Running head: LATENT SEMANTIC ANALYSIS AND KNOWLEDGE ASSESMENT

# Using Latent Semantic Analysis to assess knowledge: Some technical considerations

Bob Rehder, M. E. Schreiner, Michael B. W. Wolfe, Darrell Laham,

Thomas K Landauer, and Walter Kintsch

University of Colorado, Boulder

Abstract

In a previous paper (Wolfe, Schreiner, Rehder, Laham, Foltz, Landauer, & Kintsch, this issue) we have shown how Latent Semantic Analysis (LSA) can be used to assess student knowledge - how essays can be graded by LSA and how LSA can match students with appropriate instructional texts. We did this by comparing an essay written by a student with one or more target instructional texts in terms of the cosine between the vector representation of the student's essay and the instructional text in question. This simple method was effective for the purpose, but questions remain about how LSA achieves it’s results and how they might be improved. Here we address four such questions: (a) what role use of technical vocabulary per se plays, (b) how long should the student essays be, (c) whether the cosine is optimal measure of semantic relatedness, and (d) how to deal with the directionality of knowledge in the high-dimensional space.

# Using Latent Semantic Analysis to assess knowledge: Some technical considerations

## The Role of Technical Terms

The semantic relatedness between a student's essay on a certain topic and an instructional text in that domain proved to be a reliable measure of student knowledge and a valuable predictor of how much the student could learn from the text. What features of the essays were responsible for these properties? Specifically, what role does the technical vocabulary play? Does LSA owe its success merely to the students' use of technical vocabulary? If a student had generated an unstructured 'bag of technical words' that didn't really reflect high level understanding (e.g., a well-developed situation model), would LSA have done equally well?

To investigate this question we re-analyzed data from Wolfe et al (see for necessary details) as follows:

(1) A list of technical heart/circulatory terms was developed using a lose criterion (e.g., "left" and "right" were counted as non-technical terms; "pump", "body", "purple" and "red" were counted as technical terms, as were more sophisticated terms, such as "superior", "bulbo" and "spiral").

(2) The participants' original (Original) pre- and post-learning essays was separated into technical terms (Technical) and other words (Non-Technical). 47.9% of the words on which LSA bases it's analysis which the students wrote were classified as technical.

(3) Two new LSA vectors for each student's essay were computed, one based on the technical words only and one based on the non-technical words only.

(4) Cosines between the Technical and Non-technical essay vectors and Text C[1] (in addition to the original, whole-essay cosines obtained by Wolfe et al.) were computed.

(5) The pattern of correlations between the three types of cosine measures (i.e. Original, Technical and Non-Technical) and the students' performance on the pre-questionnaire test were compared. (Recall that the pre-questionnaire test was a 40 item short answer test, taken before writing an essay and reading the instructional text.)

The correlations between the pre-questionnaire scores and the three types of cosine measures for all 106 participants in the Wolfe, et al study (94 undergraduates and 12 medical students) are given in Table 1; all correlations are significant $p < .0001$ level.

| | Pre-Questionnaire | Original | Non-Technical |
|---|---|---|---|
| Original | .71 | | |
| Non-Technical | .59 | .83 | |
| Technical | .69 | .94 | .63 |

Table 1. Correlations between pre-questionnaire scores and the three cosine measures.

It is true, as one might have suspected, that cosines computed from essay vectors that contain only a list of the technical terms the students used correlated about as highly with the pre-questionnaire scores as cosines

computed from the intact essays. Thus, the technical vocabulary a student uses makes a difference for LSA, and, indeed, an appropriately weighted count of technical word use in an essay can serve as an effective measure. What is surprising, however, about the correlations in Table 1 is that cosines computed from the *non-technical* words in the essays yielded almost equally good predictions as the cosines computed from the technical vocabulary! The non-technical words students use in describing the functioning of the heart contain a great deal of information about their knowledge of the heart - indeed almost as much as their technical vocabulary, or the essays as a whole.

We conclude that nothing is to be gained by separating essays into technical and non-technical terms (which is neither easy nor straightforward).

What if we had students simply generate a list of technical, heart-related terms, instead of writing an essay, and then used the vector representation of that list in the LSA space as our estimate of heart knowledge? Table 1 suggests that such a procedure might be effective, although one cannot simply equate such a list with the technical terms we extracted from the students' essays. It is not clear that the same or even a similar list would be generated by the two procedures, given the very different task demands involved (e.g., our list contains repetitions). Perhaps the only way that a student can generate an accurate set of words is to compose a good essay. This conjecture require further research.

## Essay Word Count

In the Wolfe et al experiment, participants were instructed to write an essay of approximately 250 words. Nevertheless, there was a fair amount of variability in essay word count. The mean length of the complete pre-essays was 261.2 words (s.d. = 15.02), with word counts ranging from 209 words to 306 words[2]. The correlation between essay word count and pre-questionnaire

score was non-significant (r = .12). This result may be due to the fact that the length of participants' essays were constrained (to about 250 words). In studies where essay length is less constrained, essay word count is strongly related to knowledge. For example, Laham and Landauer (1996) found that the length of essays written during a class period as part of a psychology test predicted the grade that the essay received. Page (1994) has also found essay word count to be a strong predictor of domain knowledge.

How long should an essay be in order to get an accurate estimate of how much a participant knows? In order to answer this question, we looked at the effectiveness of the LSA cosine measure in predicting pre-questionnaire scores as a function of essay word count. From each participant's essay we created 19 sub-essays: the first sub-essay consisted of the first 10 words, the second sub-essay consisted of the first 20 words, and so on up to 200 words (up to the minimum essay word count). We then calculated cosines between each sub-essay and the standard instructional text, Text C. Next we calculated the correlation between the set of essays of a given length (e.g., first all 106 10 word essays, then all 106 20 word essays, etc.) and each participant's pre-questionnaire score. The proportion of the variance accounted for by each of the 19 essay / pre-questionnaire correlations is given in Figure 1.

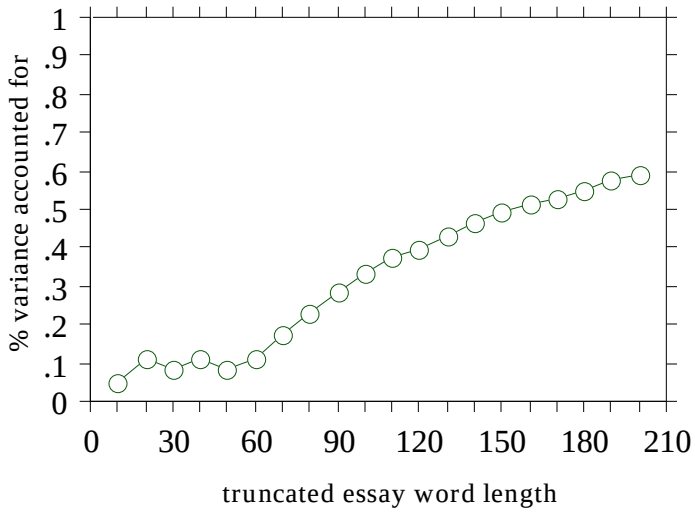



Figure 1. The proportion of variance accounted for (r^2) when predicting pre-questionnaire scores from the cosines of students' essays and Text C as a function of truncated essay word count.

The first 60 words of essays are non-predictive of knowledge level, at least under the conditions of the present study, where students were instructed to write 250-word essays. Between 70 and 200 words the cosine of the essay becomes increasingly predictive of the participants' knowledge level, but with decreasing marginal returns. Thus, the accuracy gained in measuring domain knowledge with essays considerably longer than 200 words may be negligible. Given the practical difficulties in gathering essays from students, 200 word essays appear to be a reasonable compromise.

## Alternative LSA-Derived Measures of Knowledge

In previous LSA work, the cosine between two document vectors has been the primary measure of the similarity between two documents. However, there are other possible LSA-derived measures that could be used, such as the dot product or Euclidean distance between the two vectors, or the

length of an individual vector. In this section, we explore the usefulness of these alternatives as measures of knowledge in a domain.

A vector can be thought of as a position within an n-dimensional space. The value of a vector is represented as a series of coefficients, each coefficient representing a value (or distance) along a particular dimension in the n-dimensional space. Thus, if X and Y are vectors in an n-dimensional space, then X and Y are written as:

$$X = (x_1, x_2, \ldots, x_n) \quad \text{and} \quad Y = (y_1, y_2, \ldots, y_n)$$

The inner product, or dot product, of vectors X and Y is defined as:

$$X \bullet Y = x_1y_1 + x_2y_2 + \ldots + x_ny_n$$

It is important to note that $X \bullet Y$ is a scalar, not a vector. The length of a vector X is defined as:

$$||X|| = (X \bullet X)^{1/2} = (\sum_{i=1..n} x_i^2)^{1/2} \qquad (1)$$

If we use the symbol $\theta$ to denote the angle between vectors X and Y, then the cosine of $\theta$ is defined as:

$$\cos\theta = \frac{X \bullet Y}{||X|| * ||Y||}$$

The value of $\theta$ can range between 0 and $\pi$; $\cos\theta$ can range between -1 and 1.

As it is used in LSA, the cosine between two document vectors roughly refers to the similarity of the document vectors while factoring out the length of the vectors. As we apply it to the Wolfe et al. (this issue) data, the cosine between the vector of an individual's pre-essay about the heart (E) and the vector of a standard instructional heart text is considered as a measure of pre-knowledge about the heart. We use the vector for Text C (see footnote 1) as the standard instructional text, so we refer to this cosine measure as cos EC. How well cos EC serves as a measure of knowledge may be assessed by correlating it with independent knowledge measures. The dot product

between an essay's vector and the vector for Text C, (E•C), the Euclidean distance between the essay's vector and Text C's vector, dist EC, and the length of the essay's vector itself, (||E||), are also considered as candidate knowledge measures.

For these analyses we use the data of the 94 undergraduates from Wolfe et al. (this issue). Cos EC, E•C, dist EC, and ||E|| are each individually highly significant predictors of pre-knowledge as measured by the pre-questionnaire and the pre-essay grades. These correlations are presented in Table 2 (the dim-method variables will discussed in the following section). The largest correlations are found using the dot product, E•C, rather than with cos EC, a surprising result in light of the superiority of cosines as a similarity measure found in the LSA and LSI (latent semantic indexing) literature (Harman, 1986).

| | Pre-questionnaire | Pre-essay |
|---|---|---|
| cos EC | 0.68 | 0.62 |
| E•C | 0.76 | 0.73 |
| dist EC | -0.72 | -0.69 |
| ||E|| | 0.65 | 0.65 |
| dim-method 1 | 0.67 | 0.62 |
| dim-method 2 | 0.70 | 0.63 |
| dim-method 3 | 0.83 | 0.72 |

Table 2: Correlations of pre knowledge assessment scores and LSA measures. All p values <.0001, n=94.

Because cos EC, E•C, dist EC, and ||E||, are all correlated with one another, an obvious next step is to perform a multiple regression using all four variables as predictors. Before proceeding however, it is instructive to

consider the mathematical relations between these variables. For example, the dot product is given by

$$E \bullet C = (\cos EC)(||E||)(||C||)$$

where $(||C||)$ is the length of the vector of Text C. Because $(||C||)$ is constant, a new variable, $E \bullet C' = (\cos EC)(||E||)$, will have the same correlations with the pre-knowledge measures as does $E \bullet C$. In other words, for purposes of predicting pre-knowledge scores, $E \bullet C$ may be viewed as a function of cos EC and $||E||$. In particular, $E \bullet C$ may be interpreted as the interaction term between cos C and $||E||$.

In the Appendix we prove that predicting pre-knowledge measures with dist $EC^2$, a monotonic transformation of Euclidean distance, dist EC, is equivalent to predicting pre-knowledge from a linear combination of $(\cos EC)(||E||)$ and $||E||^2$. The fact that $E \bullet C$ is a function of $(\cos EC)(||E||)$, and that dist EC is a function of $(\cos EC)(||E||)$ and $||E||^2$, suggests a multiple regression equation in which $(\cos EC)(||E||)$ and $||E||^2$ are used as predictors of pre-knowledge scores along with cos EC and $||E||$. The results of such a multiple regression predicting pre-questionnaire scores are shown in Table 3[3]. Cos EC and $||E||$ are highly significant predictors above and beyond all other variables, including each other. The interaction term (i.e., a scaled version of the dot product) was not. In other words, although the dot product was the most successful *individual* LSA measure for predicting pre-knowledge, it provides no additional predictive value above and beyond cos EC and $||E||$ together. The failure of the interaction term to reach significance also reveals that cos EC and $||E||$ are independent contributors to representing domain knowledge. Finally, the two terms for Euclidean distance, $(\cos EC)(||E||)$ and $||E||^2$, provide no additional predictive value above and beyond cos EC and $||E||$ together. Thus, we conclude that cos EC

and ||E|| exhaust the representation of knowledge embedded in the four LSA variables we originally considered in Table 2.

| Predictor | Partial Correlation | Standardized Beta Weight | F(1,93) |
|---|---|---|---|
| cos EC | .53 | 0.46 | 35.4 *** |
| ||E|| | .51 | 0.43 | 32.1 *** |
| (cos EC) (||E||) | .08 | -0.09 | <1 |
| $||E||^2$ | .03 | -0.03 | <1 |

Table 3: Results of multiple regression where pre-questionnaire scores are predicted from cos EC, ||E||, (cos EC) (||E||), and $||E||^2$. Multiple R^2 = 0.61. *** = p<.0001.

The fact that cos EC is a highly significant predictor of pre-knowledge is expected, as cos EC reflects the direction of an essay's vector in the high-dimensional LSA space, and the vector's direction is interpreted as the representation of the quality of the semantic content of the essay. In order to interpret the finding that essay vector length, ||E||, is a highly significant predictor of pre-knowledge above and beyond cos EC, we must consider what LSA spaces are, and how LSA vectors are computed in an LSA space. An LSA space is intended to represent a multi-dimensional semantic space. As the degree of association with one or more semantic dimensions increases, the length of an essay's vector increases.

A number of factors influence how the semantic associations represented by an essay's vector are determined. First, the LSA space we used was constructed from encyclopedia articles only about the heart. As a result, words that are "off-topic" (not about the heart) do not appear in the LSA

matrix, and hence cannot affect an essay's vector representation, including its length. Second, words that appear rarely in the heart encyclopedia articles (such as technical words) are more heavily weighted by LSA than words that appear less frequently, under the assumption that rare words are more likely to distinguish documents from one another semantically. Heavily weighted words increase the degree of semantic association relative to less-heavily weighted words, resulting in longer vectors. Third, before being converted to their LSA vector representations, essays were first submitted to a "stop list" that removed their non-content words such as "the" and "of" (reducing the size of the essays by an average of 56%, mean=116, s.d.=11.15). Fourth, broad essays that are associated with a number of semantic dimensions will have a longer vector length that essays that are written narrowly. To summarize, an essay's vector length is (a) a strong positive function of the number of rare (often technical) heart words, (b) a moderate positive function of the number of common heart words, (c) a function of the breadth of heart knowledge expressed in the essay, and (d) unrelated to the number of non-content and off-topic words.

Thus, the length of an essay's vector reflects an individual's general knowledge about the heart (or, at least, the general knowledge about the heart embedded in the encyclopedia articles). In contrast, the cosine measure reflects the more narrow knowledge embedded in the standard instructional text, Text C. Therefore, knowledge measured both broadly and narrowly is important for predicting the performance on the pre-questionnaire employed in this study.

The importance of LSA vector lengths as a representation of domain knowledge needs to be supported by replication in other domains. We can cite one additional study from our laboratory. The Laham and Landauer (1996)

study described earlier used an LSA vector representation of essays in the domain of psychology to predict the grades that the essays received, and found vector length to be a highly significant predictor of the grades of the essays above and beyond a cosine measure, and to be more strongly correlated with grade than essay word count. Thus, the importance of vector length in representing knowledge may have some generality.

Essay Word Count versus Essay Vector Length . It is instructive to consider the importance of essay word count in determining the length of an essay's vector. If all essays had the same proportion of rare heart words, common heart words, and non-content and off-topic words, then essay word count would be a strong predictor of essay vector length (because essay word count would predict the number of rare heart words, the number of common heart words, etc.). In fact, the correlation between essay word count and essay vector length is not significant (r=.14). Furthermore, whereas essay vector length is a strong predictor of pre-knowledge (even above and beyond the cosine measure), in a previous section we showed that essay word count was not significantly related to pre-knowledge. For these reasons, we conclude that as essay word count increases that the relative proportion of heart words (rare and common) and non-content and off-topic words changes such that there are fewer heart words. Thus, those subjects that wrote longer essays did not necessarily possess more heart knowledge, and our use of LSA was able to detect that fact.

However, the lack of relationship between essay word count and essay vector length may be due to the fact that participants were constrained to write an essay of 250 words. With less-constrained essays, Laham and Landauer (1996) found that essay word count and vector length were highly correlated (r=.96).

## The Goldilocks Principle, and The Problem of Directionality in High-Dimensional Space

In Wolfe et al. (this issue) we investigated the hypothesis that learning is optimal when a text is neither too easy nor too hard relative to the learner's background knowledge, a hypothesis we have come to call “the Goldilocks Principle”. The cosine between an essay the student wrote before receiving instruction and the instructional text served as an estimate of how difficult that text would be for that student: if the cosine was very high, the text was too easy, if it was very low, the text was too hard, if it was intermediate the text was just right. The cosine, however, measures relatedness as an unsigned angle in a high-dimensional space. The essays of two individuals may have the same cosine with an instructional text, but the essay of the first individual may be dissimilar to the text because the individual knows very little about the topic (relative to the text), whereas the essay of the second individual may be dissimilar to the text because the individual knows very much about the topic (relative to the text). Take the hypothetical case of elementary school students and cardiovascular surgeons, all of whom write 250 word essays about the functioning of the human heart. When compared to an undergraduate level text, these two groups of essays might well have similar cosines. Both are equally dissimilar to the undergraduate text, but for very different reasons: the surgeons' essays are dissimilar because they know significantly more than is contained in the undergraduate essay, whereas the elementary school students' essays are dissimilar because they know significantly less.

We believe that this problem, which we refer to as the directionality problem, did not materially affect our analysis of the undergraduate data in

Wolfe et al (this issue) because all the undergraduates who participated in the study knew little about the heart relative to the four instructional texts. As long as all potential learners have considerably *less* knowledge than all the potentially to-be-read text, it is clear how the LSA cosine similarity measure can serve as a proxy for a knowledge measure: A higher cosine reflects more knowledge and a lower cosine reflects less knowledge.

The directionality problem can be most clearly illustrated with the medical school students who participated in the Wolfe et al study. The fact that the medical students were relatively high-knowledge and the undergraduates were relatively low-knowledge, raises the possibility that the directionality problem might come into play. That is, these two groups might have the same cosines (similarity) with a text, even though one group is *above* the text and the other *below* it. Figure 2 presents the distribution of cosines with Text A for the undergraduates and for the medical students. The cosines of the 2 groups do not differ significantly ($t_{(104)} = 1.76$, $p > .05$).

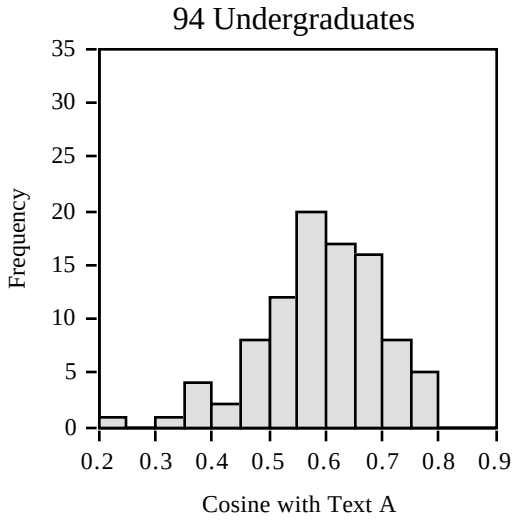



Figure 2a. Distribution of cosines with Text A for the 94 undergraduates.

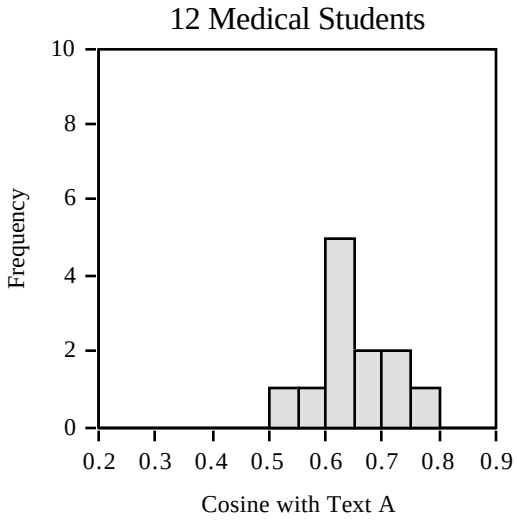



Figure 2b. Distribution of cosines with Text A for the 12 medical students.

However, on the basis of the distributions of pre-questionnaire scores presented in Figure 3 it is apparent that the medical students know far more about the heart than the undergraduate students ($t_{(104)} = 9.77$, $p < .0001$). Clearly, cosines with an instructional text are not always a sufficient means by which to discriminate between high- and low-knowledge individuals.

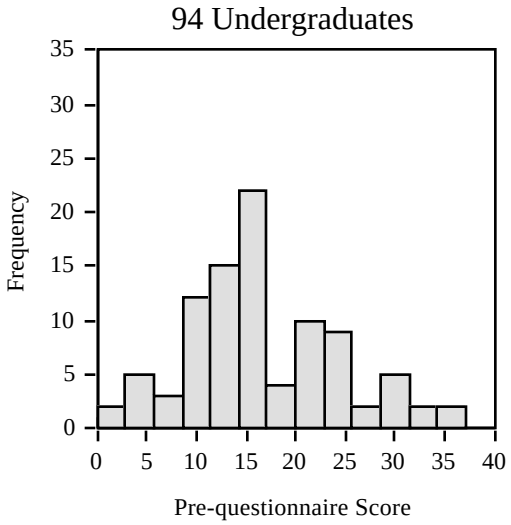



Figure 3a. Distribution of pre-questionnaire scores for the 94 undergraduates

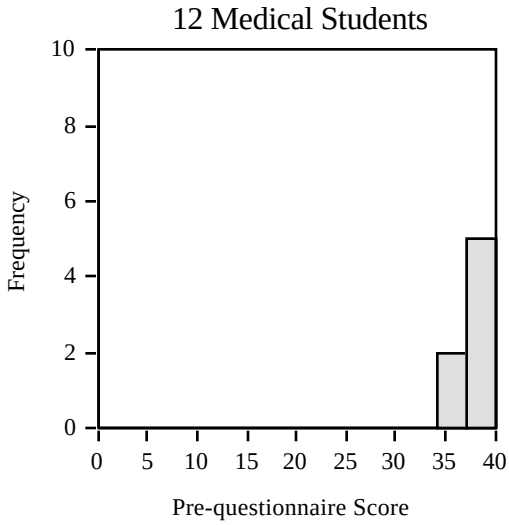



Figure 3b. Distribution of pre-questionnaire scores for the 12 medical students.

The directionality problem can arise when essays are compared with *one* instructional text. In fact, the difference between low- and high-knowledge individuals becomes apparent when additional similarities between texts and essays are considered. For example, the cosines between the essays of the undergraduates and the medical students reveal that they are quite dissimilar. Comparing the undergraduates and medical students with less-advanced or more-advanced instructional texts also reveals the differences between the groups.

Landauer & Dumais (in press) have described how an LSA space may be viewed as representing correlational information originally generated

from a high-dimensionality Euclidean semantic space. Our solution to the directionality problem employs multidimensional scaling (MDS) to re-compute a pertinent portion - a local subspace - of the original representation from similarity measures produced by LSA. Because the MDS procedure can consider the distances among several different texts and essays simultaneously, it can distinguish between low- and high-knowledge individuals. Indeed, in our scaling results (reported in detail below) essays of the undergraduates and the medical students tended to cluster in different regions of the Euclidean spaces that our MDS procedures produced.

In order to apply the zone of learnability (the Goldilocks principle), we also need a scale that reflects the amount of knowledge associated with the pre-essays written by our subjects and with the instructional texts. Three alternative methods were used to construct the MDS Euclidean subspace and the one-dimensional knowledge scale.

Method 1. Method 1 consists of two steps. In the first step, the cosines between all pairs of instructional texts were computed (see Table 3 in Wolfe, et al., this issue), and then a one-dimension scaling solution was produced. This one-dimension solution was directly interpreted as a single-dimension knowledge scale. The low- versus high-knowledge ends of the scale were assigned by relying on external knowledge about the relative difficulty of the instructional texts. In Wolfe, et al. the four instructional texts were carefully chosen according to difficulty, with Text A the easiest and Text D the hardest. As expected, the texts were arranged on the single dimension in the order A, B, C, and D, and the ends of the scale on which Text A and D appeared were interpreted as the low- and high-knowledge ends of the scale, respectively.

In the second step, for each pre-essay, the cosines between the pre-essay and the four instructional texts were computed, and the position of the pre-

essay on the dimension that was most consistent with those cosines was determined. This position becomes the measure of knowledge for a subject. Figure 4 presents the distribution of dimension scores for both the 94 undergraduates and the 12 medical students, as well as the position of the four texts on the dimension. The difference between the undergraduates and medical students is significant ($t_{(104)} = 3.69$, $p < .001$). Thus, Method 1 successfully discriminates between the two groups, and hence provides a potential solution to the directionality problem. Further evidence that the single dimension produced by Method 1 corresponds to a knowledge scale is presented in Table 2: The correlation between the single dimension and the undergraduates' pre knowledge assessment scores is highly significant.

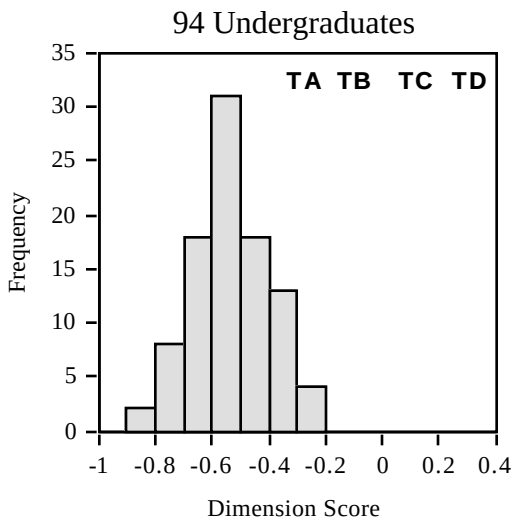



Figure 4a. Distribution of dimension scores for the 94 undergraduates computed by Method 1.

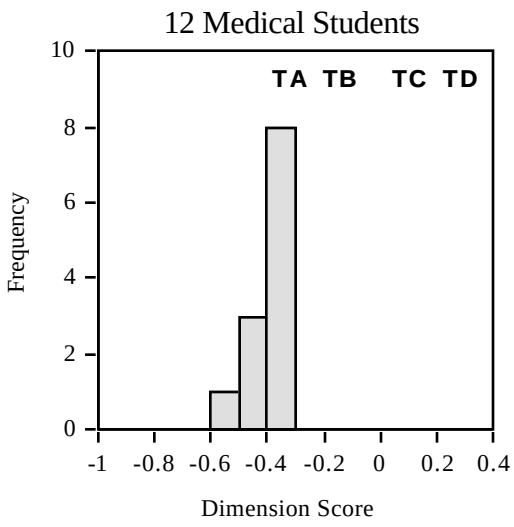



Figure 4b. Distribution of dimension scores for the 12 medical students computed by Method 1.

Method 2. The second method also generates a one-dimension MDS solution. However, whereas Method 1 only utilized cosines between the four texts (in step 1), and between the pre-essays and the four texts (in step 2),

Method 2 utilizes all available cosines: Between the texts, between all the essays (pre and post), and between the texts and all the essays. With all this information being used jointly in the subspace scaling step it is possible that a more accurate Euclidean space can be constructed. Once again, the MDS procedure arranged the four instructional texts on the resulting dimension in the expected order A, B, C, D. As in Method 1, we assumed that Text A was at the low knowledge end of the scale and Text D at the high knowledge end. The distributions of the resulting pre-knowledge dimension scores for the undergraduates and medical students, as well as the position of the four instructional texts, are shown in Figure 5. The difference between the groups is highly significant ($t_{(104)} = 4.78$, $p < .0001$). As for Method 1, the correlation between the single dimension and the undergraduates' pre knowledge assessment scores is highly significant (Table 2).

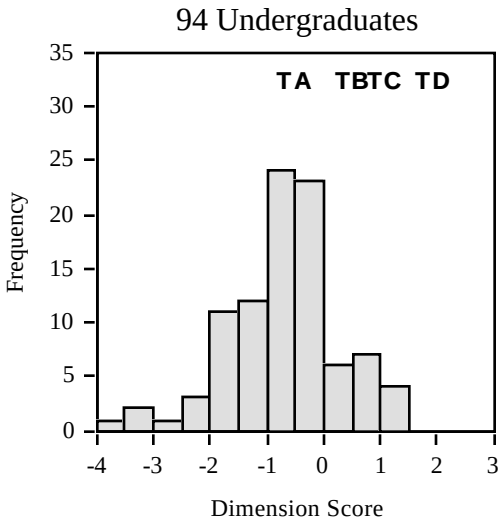



Figure 5a. Distribution of dimension scores for the 94 undergraduates computed by Method 2.

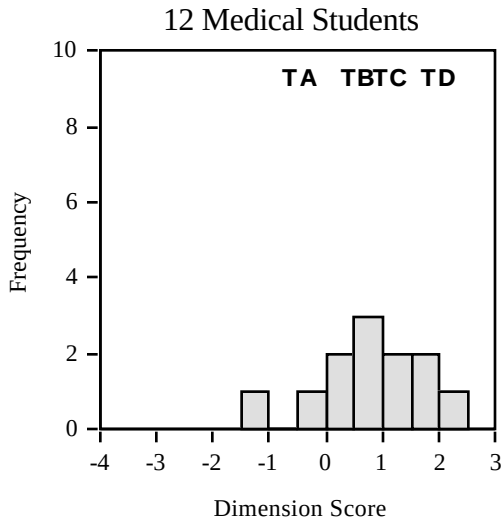



Figure 5b. Distribution of dimension scores for the 12 medical students computed by Method 2.

Note that Methods 1 and 2 are essentially an application of the technique of unfolding, pioneered by Coombs (1964).

Method 3. Method 2's one-dimension MDS solution is the single dimension that best represents differences between all the textual units that went into the MDS procedure (namely, the instructional texts and the essays). It is notable that this single dimension corresponds so closely to pre knowledge assessment scores. There are undoubtedly many ways in which the textual units differ from one another, and yet it appears that knowledge level is largest systematic difference between them. Nevertheless, these other differences may still be contributing nontrivial noise to Method 2's single dimension. In Method 3, we attempt to find yet a more accurate representation of knowledge. Once again all the available cosines were submitted to a MDS procedure, but rather than dictating the dimensionality of the solution *a priori* as we did in Methods 1 and 2, we used standard procedures (Carrol & Arabie, in press) to find a 10-dimensional solution that captured all the non-random differences represented by the cosines.

In order to construct a knowledge scale from this 10-dimension Euclidean space, Method 3 uses the pretest scores of each subject as an independent pre-knowledge measure to determine the single dimension through the 10 dimensional space that was most closely related to overall heart knowledge. The following regression equation,

$$\text{Pretest} = \beta_0 + \beta_1 \cdot \text{Dim}_1 + \beta_2 \cdot \text{Dim}_2 + \ldots + \beta_{10} \cdot \text{Dim}_{10}$$

produced beta weights $B_1$, $B_2$, ..., $B_{10}$, which were used to define a single dimension Dim,

$$\text{Dim} = B_1 \cdot \text{Dim}_1 + B_2 \cdot \text{Dim}_2 + \ldots + B_{10} \cdot \text{Dim}_{10}$$

The position of each individual's pre-essay in the 10-dimensional space was reduced to a single Dim score, which was interpreted as a measure of the

individual's knowledge of the heart. Figure 6 presents the distribution of dimension scores of the 94 experimental participants and the 12 medical students. This dimensional score does an excellent job of distinguishing low-knowledge individuals from high-knowledge individuals, as the difference between the two groups is highly significant ($t$ (104) = 6.35, $p$ < .0001). Figure 6 also presents the position of the four instructional texts. The position of Text A confirms the implication of Figure 2: Text A lies at a knowledge level intermediate between the knowledge level of the experimental participants and the knowledge level of the medical students. Finally, the dimension score correlates highly with the undergraduates' pre knowledge assessment scores (Table 2).

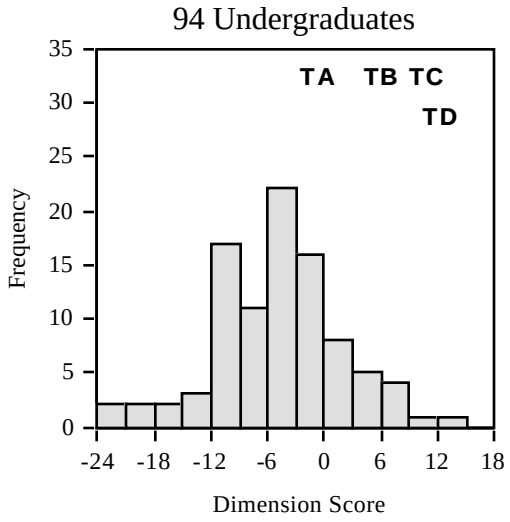



Figure 6a. Distribution of dimension scores for the 94 undergraduates computed by Method 3.

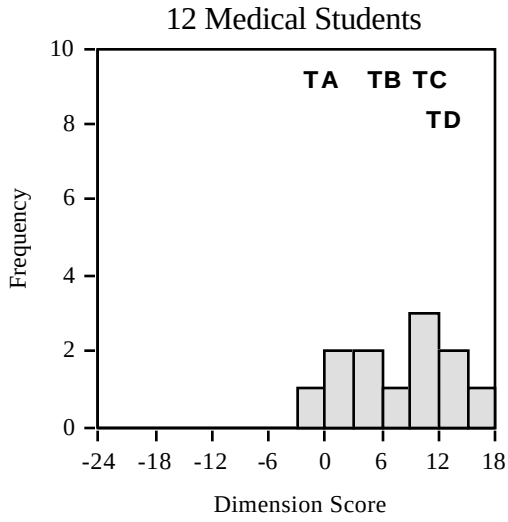



Figure 6b. Distribution of dimension scores for the 12 medical students computed by Method 3.

Summary. The directionality problem can be solved by combining LSA similarity measures and multidimensional scaling, using one of the methods

described here. According to the ability to discriminate between the undergraduates and medical students from the Wolfe, et al. (this issue) study, and according to correlations between the constructed knowledge scale and empirically measured pre knowledge assessment scores (Table 2), Method 3 is the most effective and Method 1 is the least effective. It must be noted, however, that the results of Method 3, unlike those of Methods 1 and 2, involves a step of empirical fitting of the derived scale parameters (the beta weights) to the predicted test data, and thus necessarily capitalize to some unknown extent on chance. Thus, the superiority of Method 3 is probably somewhat exaggerated.

Also, in practice Method 3 may be difficult to apply because it requires the availability of the external knowledge measure. An important finding is that Method 2 appears to discriminate between low- and high-knowledge subjects nearly as well as Method 3 without such an external knowledge measure. Research on additional datasets will be required to determine whether the effectiveness of Method 2 generalizes to other domains.

The methods we have presented all rely on multidimensional scaling to produce a Euclidean spatial representation of the knowledge level of our subjects and texts. Note that analyses similar to those we have presented could have been carried out in the original SVD space. However the dimensionality of the Euclidean spaces we have computed are vastly reduced compared to the original SVD space, and such recomputed sub-spaces may have advantages. There is a sense in which many of the dimensions of a high dimensional representation may be serving to separate hierarchical or other non-scalar product relationships, yielding an approximation rather than a true representation of the relationship of some elements to each other. This raises the possibility that linear relations among a subset of the objects are

distorted in the global SVD solution. The iterative step of recomputing the space on only the desired objects, starting with global values as initial distance estimates, may improve the local representation. Whether this is actually the general case for applications of LSA, and such issues as whether the iteration converges, etc., needs to be investigated.

## Conclusion

The technical considerations we have explored in this report in general support the way LSA has been used in Wolfe et al. (this issue). What we have found can be summarized by the following points:

(1) Student's domain knowledge can be assessed effectively by letting them write essays on a central topic in that domain. Nothing is to be gained by separating out technical terms from the non-technical vocabulary in these essays, because the two contribute about equally to the prediction of knowledge. It should be emphasized that these essays should be in the students own words - answers to questions by the teacher may not do, for instance, because such answers tend to repeat the question and hence reflect the teacher's choice of words, rather than the student's.

(2) In the present study, a 200 word essay was sufficient for the purpose of knowledge assessment. The minimum length required was at least 60 words. This observation is encouraging for practical applications as it suggests that relatively small samples of student writing need be analyzed, a particular advantage if the material must be transcribed in machine-readable form. However, these findings are possibly specific to our methodology which required the students to write essays of a fixed length.

(3) The cosine between the vector representation of a student's essay and the vector representation of an instructional text in the LSA space is probably the best single measure of their semantic relatedness. The length of the essay vector was also a significant predictor of knowledge in the present study, but further research with essays of varying length will be needed to determine the generality of this finding, and to determine the optimal manner for combining these measures.

(4) The cosine between a student's essay and an instructional text only measures their semantic relatedness, or similarity. It does not tell us whether the essay is below the text or above the text in terms of amount of domain knowledge. However, in the Wolfe et al. study and, we suspect, in most instances where our methods might be employed to match students and text, all students are well below the text in their level of sophistication. When this is the case, no directionality problem arises.

(5) If a directionality problem exists, as was the case with the medical students in the Wolfe et al. study, multidimensional scaling methods can be used to construct a representation of the readers' essays and the instructional texts. One of the three methods described above can be applied to determine their relative positions on a domain knowledge scale.

(6) Another open question that has not been addressed here concerns the role of semantic spaces based on different training corpora in LSA. A special “heart” space, obtained from a scaling only of the articles in an encyclopedia related to this topic was used in all analyses reported here and in Wolfe et al. The general encyclopedia space described in Landauer and Dumais (in press), for example, would have been less

suitable for this purpose because all heart-related items are squashed together in the larger space. What space to use when, as well as the relationship between alternative spaces, needs to be investigated further.

In conclusion we can say that the rather simple approach we took in the Wolfe et al. study was justified by the analyses presented here. As we have pointed out, there remain several open questions that can only be answered by further research, but at this point, the procedures followed by Wolfe et al. appear to be adequate.

References

Carrol, J.D. & Arabie, P. (in press). Multidimensional scaling. In M. H. Birnbaum (Eds). Handbook of Perception and Cognition, Volume 3: Measurement, Judgment and Decision Making. New York: Academic Press

Coombs, C.H. (1964). A theory of data. New York: Wiley.

Harman, D. (1986). An experimental study of the factors important in document ranking. In Rabbit, F. (Ed.). Association for Computing Machine's Ninth Conference on Research and Development in Information Retrieval. New York: Association for Computing Machines.

Judd, C.M., & McClelland, G.H. (1989). Data analysis: A model-comparison approach. San Diego, CA: Harcourt Brace Javanovich.

Laham, D., & Landauer, T. K. (1996). Unpublished manuscript.

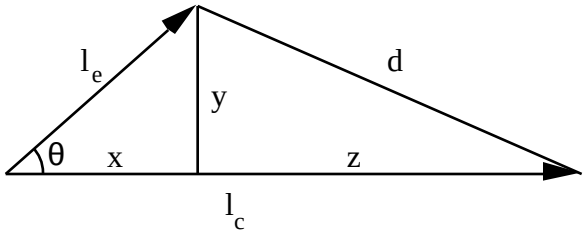
Landauer, T. K., & Dumais, S. T. (in press). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62, 127-142.

Wolfe, M. B. W., Schreiner, M.E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W. & Landauer, T. K (this issue). Learning from text: Matching readers and texts by Latent Semantic Analysis.

## Appendix

The following diagram depicts two vectors, the vector of an essay (the short one), and the vector of the "standard" text (Text C, the long one).




The goal is to express the Euclidean distance between the vectors as a function of the vector lengths and the cosine.

Definitions:

| | | |
|---|---|---|
| $l_e$ | = | length of essay vector (denoted $\|\|E\|\|$ in the main text). |
| $l_c$ | = | length of vector of standard text (denoted $\|\|C\|\|$ in the main text). |
| $\cos\theta$ | = | cosine between essay and Text C (denoted cos EC in the main text). |
| d | = | Euclidean distance between vectors (denoted dist EC in the main text). |

We know $l_e$ and $\cos\theta$. Because $l_e^2 = x^2 + y^2$ and $x = l_e \cos\theta$,

$$y^2 = l_e^2 - x^2$$

$$y^2 = l_e^2 - l_e^2 \cos^2\theta \qquad (1)$$

Because $d^2 = y^2 + z^2$, and $z = l_c - x$,

$$d^2 = y^2 + (l_c - x)^2$$

$$d^2 = y^2 + (l_c - l_e \cos\theta)^2$$

$$d^2 = y^2 + l_c^2 - 2l_c l_e \cos\theta + l_e^2 \cos^2\theta$$

Substituting in the expression derived for $y^2$ in Eq (1),

$$d^2 = (l_e^2 - l_e^2 \cos^2 \theta) + l_c^2 - 2l_c l_e \cos \theta + l_e^2 \cos^2 \theta$$

$$d^2 = l_e^2 + l_c^2 - 2l_c l_e \cos \theta \qquad (2)$$

This shows that $d^2$ (a transformed version of Euclidean distance, d) reduces to a linear combination of products of $l_e$ and cos $\theta$ (remember that $l_c$ is a constant). Suppose we were to regress pretest scores onto $d^2$. The regression equation would be,

$$\text{Pretest} = \beta_0 + \beta_1 \quad d^2$$

Substituting in the expression derived for $d^2$ in Eq (2), and rearranging,

$$\text{Pretest} = \beta_0 + \beta_1 \quad (l_e^2 + l_c^2 - 2l_c l_e \cos \theta)$$

$$\text{Pretest} = (\beta_0 + \beta_1 l_c^2) + \beta_1 \quad l_e^2 + (-2\beta_1 l_c) l_e \cos \theta$$

Thus, predicting Pretest from $d^2$ is equivalent to the following regression equation,

$$\text{Pretest} = \beta_0 + \beta_1 \quad l_e^2 + \beta_2 \quad l_e \cos \theta$$

with $\beta_1$ and $\beta_2$ constrained such that $\beta_2 = -2\beta_1 l_c$.

## Author Note

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

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 Bob Rehder, Department of Psychology (Cognitive), Campus Box 345, University of Colorado, Boulder, CO, 80309. Electronic mail may be sent via internet to rehder@psych.colorado.edu.

Table 1

Correlations between pre-questionnaire scores and the three cosine measures.

| | Pre-Questionnaire | Original | Non-Technical |
|---|---|---|---|
| Original | .71 | | |
| Non-Technical | .59 | .83 | |
| Technical | .69 | .94 | .63 |

Table 2

Correlations of pre knowledge assessment scores and LSA measures. All p values <.0001, n=94.

| | Pre-questionnaire | Pre-essay |
|---|---|---|
| cos EC | 0.68 | 0.62 |
| E•C | 0.76 | 0.73 |
| dist EC | -0.72 | -0.69 |
| \|\|E\|\| | 0.65 | 0.65 |
| dim-method 1 | 0.67 | 0.62 |
| dim-method 2 | 0.70 | 0.63 |
| dim-method 3 | 0.83 | 0.72 |

Table 3

Results of multiple regression where pre-questionairre scores are predicted from cos EC, $||E||$, (cos EC) ($||E||$), and $||E||^2$. *** = p<.0001.

| Predictor Variable | Partial Correlation | Standardized Beta Weight | F(1,93) |
|---|---|---|---|
| cos EC | .53 | 0.46 | 35.4 *** |
| $||E||$ | .51 | 0.43 | 32.1 *** |
| (cos EC) ($||E||$) | .08 | -0.09 | <1 |
| $||E||^2$ | .03 | -0.03 | <1 |

## Figure Caption

Figure 1. The proportion of variance accounted for (r^2) when predicting pre-questionnaire scores from the cosines of students' essays and Text C as a function of truncated essay length.

Figure 2a. Distribution of cosines with Text A for the 94 undergraduates.

Figure 2b. Distribution of cosines with Text A for the 12 medical students.

Figure 3a. Distribution of pre-questionnaire scores for the 94 undergraduates

Figure 3b. Distribution of pre-questionnaire scores for the 12 medical students.

Figure 4a. Distribution of dimension scores for the 94 undergraduates computed by Method 1.

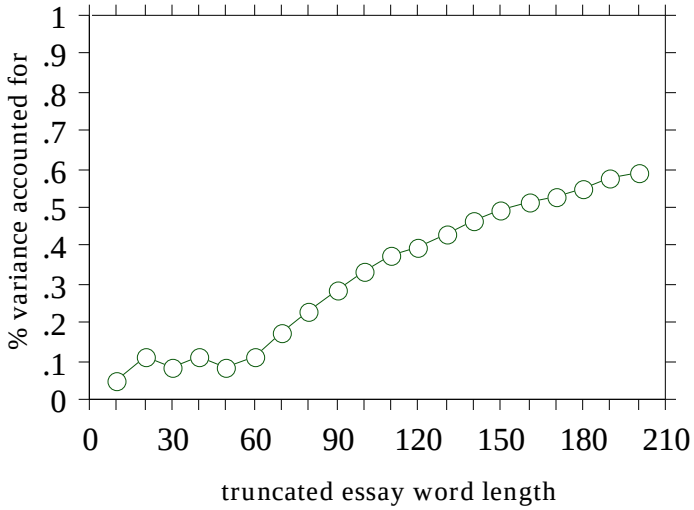
Figure 4b. Distribution of dimension scores for the 12 medical students computed by Method 1.

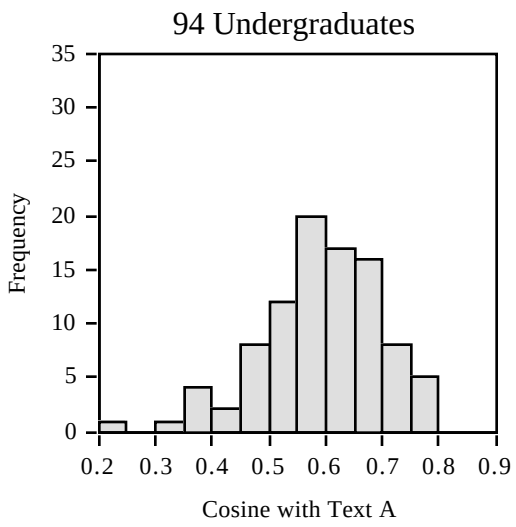
Figure 5a. Distribution of dimension scores for the 94 undergraduates computed by Method 2.

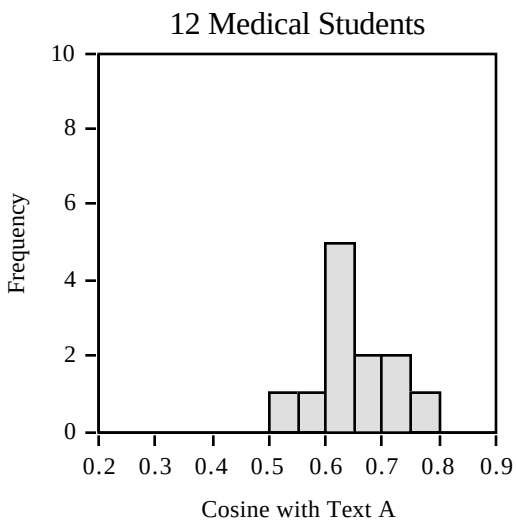
Figure 5b. Distribution of dimension scores for the 12 medical students computed by Method 2.

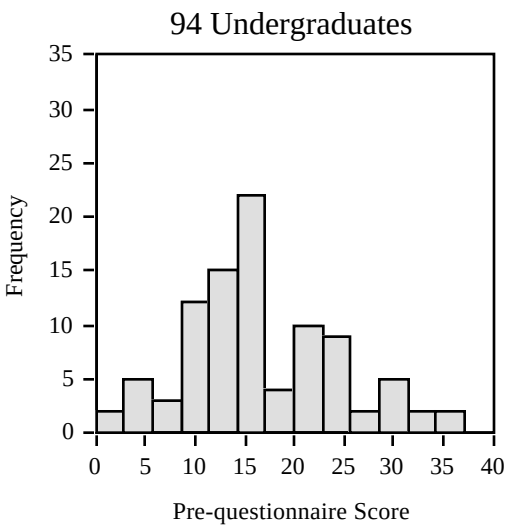
Figure 6a. Distribution of dimension scores for the 94 undergraduates computed by Method 3.

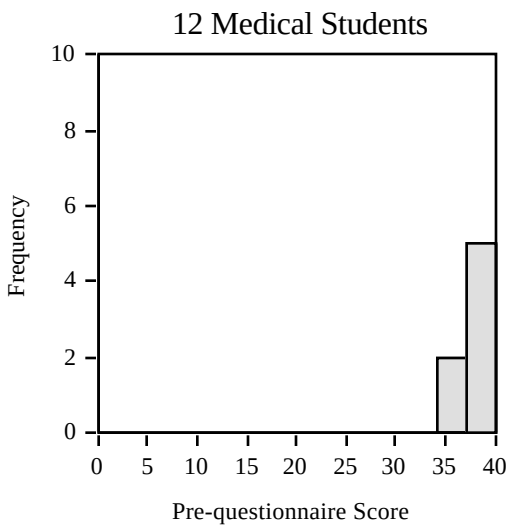
Figure 6b. Distribution of dimension scores for the 12 medical students computed by Method 3.

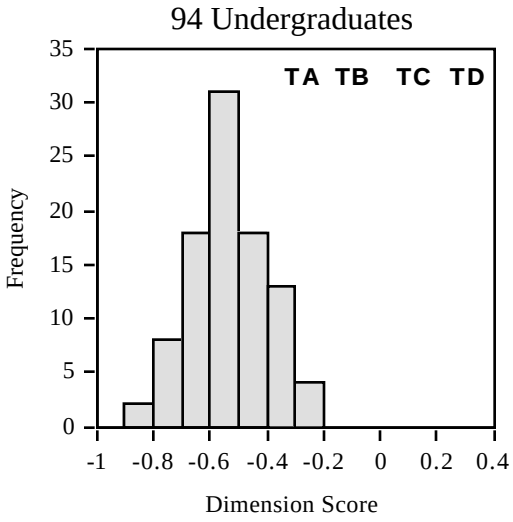

% variance accounted for
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
0
30
60
90
120
150
180
210
truncated essay word length

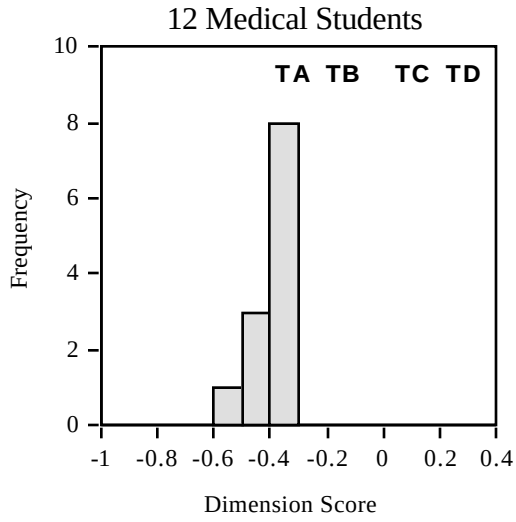

94 Undergraduates
Frequency
35
30
25
20
15
10
5
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Cosine with Text A


12 Medical Students
Frequency
10
8
6
4
2
0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Cosine with Text A

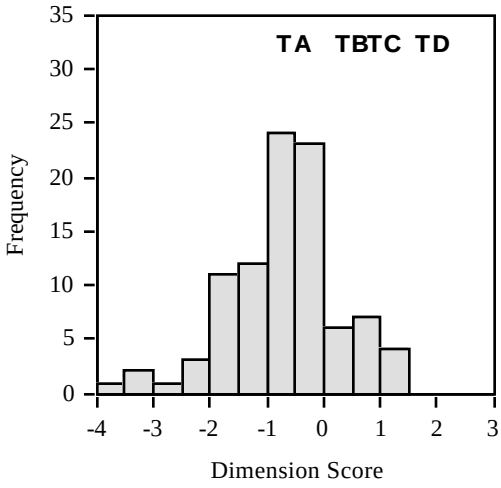

94 Undergraduates
Frequency
0
5
10
15
20
25
30
35
0
5
10
15
20
25
30
35
40
Pre-questionnaire Score


12 Medical Students
Frequency
0
2
4
6
8
10
0
5
10
15
20
25
30
35
40
Pre-questionnaire Score

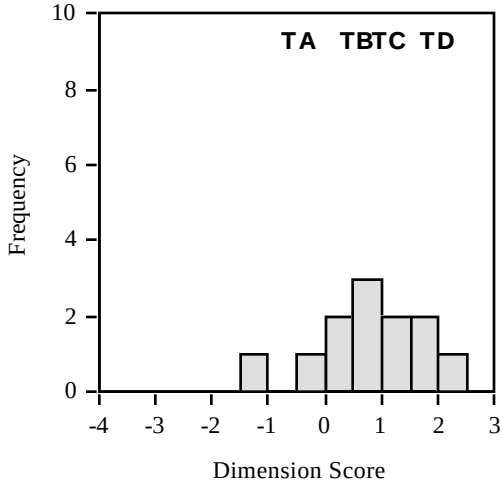

94 Undergraduates
TA TB TC TD
Frequency
35
30
25
20
15
10
5
0
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4
Dimension Score

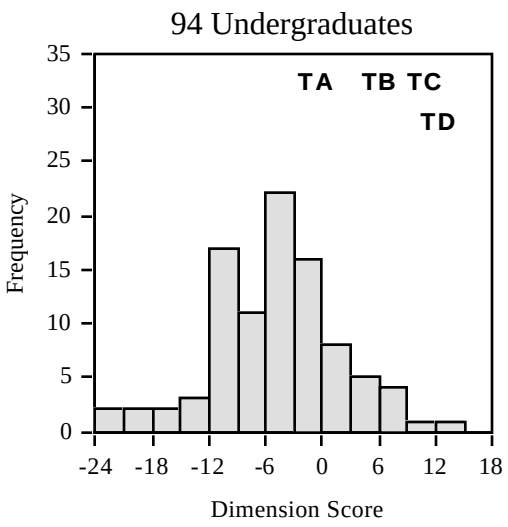

12 Medical Students
TA TB TC TD
Frequency
10
8
6
4
2
0
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4
Dimension Score

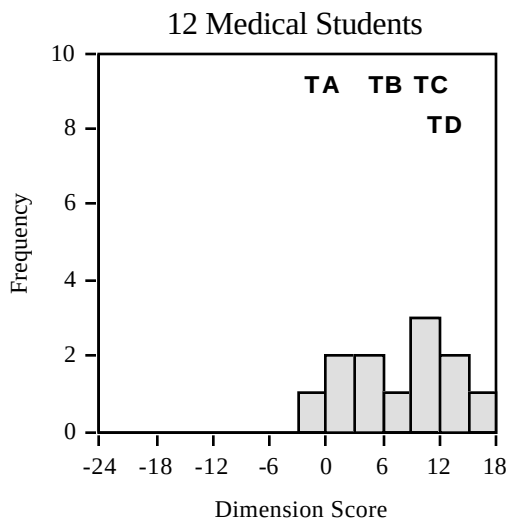
94 Undergraduates




12 Medical Students




94 Undergraduates
TA TB TC
TD
Frequency
35
30
25
20
15
10
5
0
-24 -18 -12 -6 0 6 12 18
Dimension Score


12 Medical Students
TA TB TC
TD
Frequency
10
8
6
4
2
0
-24 -18 -12 -6 0 6 12 18
Dimension Score

## Footnotes

[1] Throughout this article we use the cosine between a student's essay and a standard instructional as a LSA measure of the student's knowledge in the domain, as the instructional text is known to have correct domain knowledge. We use Text C as the standard, because the correlations between essay-to-instructional text cosines and the students' pre-questionnaire scores was greatest for Text C.

[2] It is important to note that, unlike in other sections of this paper, these are the raw lengths (in number of tokens) of the essays prior to any manipulation (i.e., separation into techical versus normal essays, LSA stop listing, SVD and vector calculation, etc.) In particular, this measure is not the same as the LSA vector length used later as one of our grading measures.

[3] In this multiple regression, cos EC and $||E||$ were first put into mean deviated form, that is, were subtracted from their mean, and then (cos EC) $(||E||)$ and $||E||^2$ were computed. As a result, the tests of the parameters for cos EC and $||E||$ correspond to a test of their "main effect", that is, their effects without the higher-order terms. See Judd and McClelland (1989), pp. 255-264. Note that a multiple regression where the dependent variable was pre-essay grades rather than pre-questionnaire scores produced the same qualitative results. A multiple regressions where $E \bullet C$ and dist EC were used as predictors rather than (cos EC) $(||E||)$ and $||E||^2$ also revealed that $E \bullet C$ and dist EC were not significant predictors above and beyond cos EC and $||E||$.