How Latent Semantic Analysis (LSA) Represents Essay Semantic Content: Technical Issues and Analyses

M. E. Schreiner, Bob Rehder, Thomas K. Landauer, and **Darrell Laham**
Department of Psychology & Institute of Cognitive Science
University of Colorado, Boulder
Boulder, CO 80309-0345
{missy, rehder, landauer, dlaham}@psych.colorado.edu

LSA is a method for extracting the meaning of words and passages based on statistical analysis of large text corpora and their representation as vectors in a high-dimension "semantic space" (Landauer & Dumais, in press). The representation is accomplished with Singular Value Decomposition followed by optimal dimension reduction. Recent research (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, in press; Rehder, Schreiner, Wolfe, Laham, Landauer and Kintsch, in press) has shown that the LSA vector of an individual's essay on a technical domain is a veridical summary representation of the individual's knowledge in that domain. Ninety-four undergraduates at the University of Colorado were asked to write essays of approximately 250 words on the anatomy, function and purpose of the human heart. The essays were given to two professional readers at Educational Testing Service, Inc., who assigned a quality score from 1 to 5 reflecting how much the student knew about the subject. The students were also given a 40 point short answer test on the same topic.

LSA was trained on a set of 27 articles relevant to the heart taken from *Grolier's Academic American Encyclopedia*. A vector was computed for each student essay by averaging the vectors of the words contained in it. We evaluated the two components of an essay's LSA vector, its length and direction, as potential measures of domain knowledge. The vector length and the cosine between the vector and a short expert text (a section on the heart from a high school biology textbook) were each correlated with the short-answer test scores and the scores assigned by the ETS graders. The results are presented in Table 1.

	Correlation with	
	Average ETS Score	Short Answer Test Score
Vector Length:	.62	.68
Cosine between Vector and Expert Text:	.65	.65

Table 1: Heart essay results.

Note that the vector length and cosine measures are correlated with one another ($r=.48$, $p<.0001$), so to some extent are explaining the same variance in the external scores. However, a multiple regression using the vector length and cosine as predictors revealed that both are highly significant above and beyond each other (partial correlations of .51 and .58, respectively), and they do not interact. Thus, vector length and cosine are both unique, independent predictors of domain knowledge.

The cosine between the essay vector and expert text reflects the direction of an essay's vector in the high-dimensional LSA space, and is interpreted as the *quality* of the semantic content of the essay relative to the expert text. Vector length, on the other hand, may be interpreted as the *quantity* of general heart knowledge in the essay relative to the general heart knowledge embedded in the encyclopedia articles.

To what extent does the fact that essay vector length predicts domain knowledge merely reflect the number of words in the essay? In the Wolfe, et al. (in press) study, essay word count was not significantly correlated with either the short-answer test scores or the scores assigned by human graders ($r=.02$ and $=.01$, respectively). When non-content words such as "the" and "although" were removed from the essays, essay word count was still only weakly correlated with the external scores ($r=.25$, $p<.05$ for the short-answer test, and $r=.16$, $p<.15$ for the essay grades). 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. We conclude that the combination of the weighting of words and dimension reduction performed by LSA are of crucial importance for representing the knowledge contained in an essay.

References

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.*

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

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