INL/CON-05-00426
PREPRINT

The Validity of Human and Computerized Writing Assessment

Human Factors and Ergonomics Society 49th Annual Meeting -- 2005

Ronald L. Boring

September 2005

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may not be made before publication, this preprint should not be cited or reproduced without permission of the author. This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in this report, or represents that its use by such third party would not infringe privately owned rights. The views expressed in this paper are not necessarily those of the United States Government or the sponsoring agency.

THE VALIDITY OF HUMAN AND COMPUTERIZED WRITING ASSESSMENT

Ronald Laurids Boring, PhD
Idaho National Laboratory
Idaho Falls, Idaho 83415-3605, USA

This paper summarizes an experiment designed to assess the validity of essay grading between holistic and analytic human graders and a computerized grader based on latent semantic analysis. The validity of the grade was gauged by the extent to which the student's knowledge of the topic correlated with the grader's expert knowledge. To assess knowledge, Pathfinder networks were generated by the student essay writers, the holistic and analytic graders, and the computerized grader. It was found that the computer generated grades more closely matched the definition of valid grading than did human generated grades.

INTRODUCTION

Pedagogues have largely assumed that writing assessment maintains a high degree of validity, despite important claims to the contrary (Coffman, 1971). Assessment research has focused almost entirely on issues of reliability without regarding the equally important question of validity (Charney, 1984). While it is useful to know the consistency of graders as measured by reliability, it is perhaps more important to know that an essay grade truly reflects the student's abilities (Huot, 1990). Much can be done to reduce grader inconsistencies, but improved reliability does not necessarily signify a valid grading method.

At first, it does not seem difficult to add a measure of validity to an experiment about writing assessment. The experimenter would simply need to compare the essay grades to some external true indicator of ability or performance. This task proves easier in conception than in implementation. The task becomes daunting when one attempts to find such a true indicator. For example, a comparison of essay grades to grades on a multiple choice test can only measure validity inasmuch as the multiple choice test is a true gauge of ability. While it is possible to find or construct valid multiple choice measures, such a test may not even measure the same abilities as an essay examination (Coffman, 1971). There can be no discussion of cross-validation when two forms of assessment inherently measure different aspects of ability. Even if a clear validity benchmark for specific essay-writing abilities is determined, adoption of any indicator of validity must likely be done on a priori theoretic grounds, not on definitive methodological grounds.

A useful approach to validity determination was suggested by Madigan and Brosamer (1991). Using teaching assistants as essay graders, the researchers compared the grades assigned by the teaching assistants to the grades assigned by experienced faculty members. In this case, the measure of validity was the correlation between the novice and expert graders, which was 0.730 for long essays and 0.480 for short essays. The crucial assumption made was that expert graders are more valid in their grading practices. While this approach is an important step in the right direction toward validity research, there is little evidence to support the assumption that experienced graders are valid graders. Experience in grading may, in fact, serve only to fortify existing bad grading habits. Without additional evidence, there is insufficient proof to support the assumption that expertise in grading begets validity in grading.

STRUCTURAL KNOWLEDGE ASSESSMENT

A research technique has emerged that holds considerable promise for educational validity research. Structural knowledge assessment addresses the issue of validity by providing a means of assessing a student's cognitive representation in a particular domain (Goldsmith, Johnson, and Acton, 1991). This cognitive representation encapsulates the student's underlying knowledge in a quantifiable form. As such, it provides a necessary tool to verify that an assessment method truly measures a student's mastery of subject matter.

Structural knowledge assessment generally builds upon theories of structural knowledge (Jonassen, Beissner, and Yacci, 1993) but is specifically founded upon Pathfinder networks (Schvaneveldt, Durso, and Dearholt, 1989). Pathfinder networks are derived from proximity or similarity ratings between pairs of concepts. Given a list of concepts, each concept is paired and subsequently evaluated on a similarity scale. Raters assign a high similarity rating to two concepts that are highly similar, whereas they assign a low similarity rating to two concepts that are highly dissimilar. In the Pathfinder algorithm, only the strongest similarities are maintained in the form of links between concept nodes. The resulting Pathfinder networks are easily represented in the form of a graph, which shows all concept nodes, whereby only the highly similar nodes are connected to one another by links.

Pathfinder networks are particularly useful in determining knowledge structure differences between novices and experts. This ability to differentiate knowledge expertise lends itself well to educational assessment, where it is often desirable to know how similar a student's knowledge is to another student's or an instructor's knowledge. Wilson (1994) found that high achieving students' networks were more coherent, more structured, and more hierarchical than those of their lower achieving classmates. Similarly, Gonzalvo, Cañas, and Bajo (1994) found that experienced class instructors had more structured networks than did less experienced instructors. They also showed that the students' Pathfinder networks became more structured over the course of a semester, suggesting increasing domain expertise in the students as a result of classroom instruction. Berger and Dershimer (1993) found that students' Pathfinder networks became more structured with repeated exposure to course content and that the students' Pathfinder networks became more similar over time. McGaghie (1996) further found that the similarity of student and instructor Pathfinder networks increased over the duration of a course. Likewise, Gomez, Hadfield, and Housner (1996) discovered that the similarity between student and instructor Pathfinder networks was predictive of the students' grades in a course.

Structural knowledge assessment allows a direct comparison between a grader's semantic representation of a topic and a student's representation, as in the cited experiments. The similarity of two Pathfinder networks can be gauged using a standardized similarity coefficient developed by Goldsmith and Davenport (1990). This coefficient determines the similarity of neighborhoods of concept nodes. This computation generates a real number with a range from 0 to 1, where 0 represents completely dissimilar node neighborhoods and 1 represents identical node neighborhoods. Using the Pathfinder node neighborhood similarity coefficient, it is possible to compare the degree to which a novice's knowledge network structure matches that of an expert. Comparing this similarity with external assessment measures, it is possible to determine the extent that an assessment methodology reflects the overlap in student and grader knowledge.

It should be noted that a student's expertise in knowledge representation does not necessarily reflect that student's ability to elicit knowledge. If a student has a high network similarity to his or her grader yet does not achieve a high grade, this may reflect either a failure in the validity of the assessment method or a failure of the student to convey his or her expertise. This confound is central to the difficulties in researching validity in assessment methods.

Despite the promise of structural knowledge assessment, there yet remains no research using structural knowledge assessment in studying the validity of essay grading methods. The research presented in this paper represents the first incorporation of structural knowledge assessment in essay grading. If a student's knowledge is elicited in a finished essay, then that student's Pathfinder network should encapsulate that knowledge structure. The student's reproduction of his or her knowledge should correspond with the student's underlying knowledge structure. It can therefore be argued that in a valid assessment, the degree of similarity between the grader's and the student's Pathfinder networks should highly parallel the academic mark awarded to that student for his or her essay by the grader. Any deviation in this parallelism denotes a decrement in the validity of a given essay grading method.

EXPERIMENT

Overview

The present experiment was designed to compare two common forms of human writing assessment with an implementation of computerized essay grading. Among human graders, *holistic grading* (White, 1984) entails assessing an essay as an indivisible whole. Instead of viewing the essay in terms of distinct properties such as mechanics or ideas, the grader looks at the composite quality of the essay when grading. In contrast, the scale in *analytic grading* is a rubric of factors that are considered important in essay assessment (Diederich et al., 1961). The grader analyzes the quality of an essay in terms of the specified component factors, which are summed together to produce an overall grade for the essay. *Computerized grading* using latent semantic analysis (LSA) functions by extracting the context and usage of words in a large textual corpus (Landauer, Foltz, and Laham, 1998). The meaning of words is determined by the relationship of each word to other words in a given text, as represented in a high dimensional vector space. The composite meaning of the text is the multidimensional representation of all word contexts compared to source or ideal texts.

Participants

Fifty-nine undergraduate students who were enrolled in an Introductory Psychology course at a university participated by submitting essays and filling out similarity ratings. The essays were written outside class by students in partial fulfillment of their course semester grade. Fifteen essays were randomly selected from the student essays for assessment.

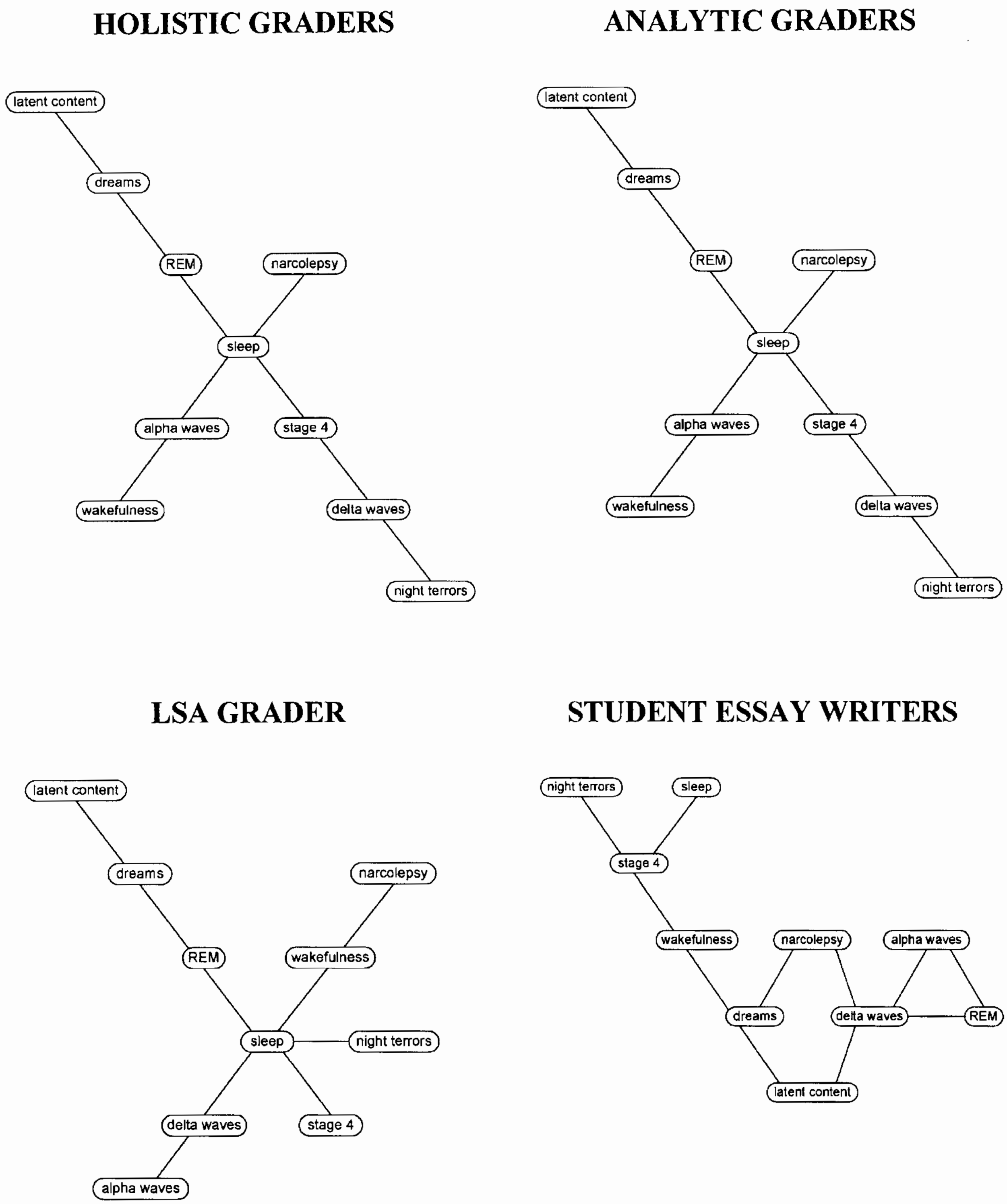
Twenty Psychology graduate teaching assistants with previous essay grading experience served as volunteer graders for the experiment. The graduate students were randomly assigned as holistic or analytic essay graders, resulting in ten holistic and ten analytic graders.

The experiment also featured a computer grader based on a unique implementation of LSA. A one-to-many document-to-document comparison was performed using a 400-factor semantic space based on the course textbook. A specially composed idealized essay served as an anchor against which the 15 student essays were graded.

Design

Separate groups of holistic and analytic graders assessed the student essays according to holistic or analytic grading methods. The computerized essay grader assessed the same set of essays. The validity of each method was determined by comparing the similarity of student and grader Pathfinder networks with the actual grades awarded.

Graders and students completed 7-point similarity ratings for a set of ten highly relevant concept terms selected by an expert panel of three instructors. Pathfinder networks were generated from the similarity ratings of the individual graders and students, *PFnet*(n-1, ∞). Similarity ratings were generated using LSA in a term-to-term comparison of the ten subject concepts in a 400-factor semantic space based on the student's textbook. These similarity ratings were transformed to a 7-point similarity scale and used to compute a Pathfinder network.



HOLISTIC GRADERS

ANALYTIC GRADERS

LSA GRADER

STUDENT ESSAY WRITERS

Figure 1. Pathfinder knowledge structure similarity networks for three types of graders and the student essay writers.

Results

The structural similarity of Pathfinder networks based on links in common was calculated using PC-Knot software. Figure 1 depicts the Pathfinder networks generated for the students and human and computer graders. Since the holistic and analytic graders were randomly sampled from a pool of graduate student teaching assistants, it was expected that the two groups would possess comparable, if not identical knowledge representations of the topic. The Pathfinder networks generated by averaged holistic and analytic graders were identical, resulting in a similarity, *C(Holistic:Analytic)*, equal to 1.000, $p < 0.001$. The similarity between the LSA-based concept distance metric and the averaged holistic and analytic graders was nonsignificant, *C(Holistic:LSA)* = *C(Analytic:LSA)* = 0.286, $p > 0.05$, suggesting that the knowledge structure elicited by the source textbook through LSA was not comparable to the knowledge structure possessed by the essay graders. The similarity between the averaged holistic and analytic graders and the averaged student essay writer was also nonsignificant, *C(Holistic:Student)* = *C(Analytic:Student)* = 0.111, $p > 0.05$. Likewise, there was no significant similarity between the knowledge representations of LSA and the averaged student essay writer, *C(LSA: Student)* = 0.176, $p > 0.05$. As expected, the expert networks generated by the graders did not generally match the novice networks generated by the student essay writers.

Another way to indicate the similarity of two networks comes by calculating the correlation between the raw similarity ratings used to generate the Pathfinder network. The correlation between the similarity ratings of the holistic and analytic graders was 0.904, $p < 0.01$. The similarity ratings generated by LSA were significantly correlated with those by holistic graders, $r = 0.379$, $p < 0.05$, and analytic graders, $r = 0.462$, $p < 0.01$. As expected, there was no significant correlation between the similarity ratings of the student essay writers and those of the holistic graders, $r = 0.071$, $p > 0.05$, the analytic graders, $r = 0.046$, $p > 0.05$, and the LSA grader, $r = 0.060$, $p > 0.05$.

As discussed earlier, validity is the degree to which the scores awarded to essays reflect the match between the essay writer's and grader's underlying semantic representation of the essay topic. In order to assess the validity of the three grading methods, the scores awarded for each essay were regressed against the similarity between student and grader Pathfinder networks. A valid essay assessment should reflect a match between student knowledge and grader knowledge. If the student has a poor knowledge representation of the topic compared to the expert knowledge representation by the grader, the essay score should be low. Conversely, if the student has an expert knowledge representation closely matching that of the grader, the essay score should be high.

As measures of grading validity, the standardized grades for each essay were regressed against the Pathfinder network similarity coefficient for holistic graders. The resulting regression line took the form, $g = 0.414C - 0.048$, where g was the grade and C was the similarity coefficient. The equation had an R^2 value equal to 0.001, $p > 0.05$. The degree of similarity between a holistic grader's Pathfinder network and the student essay writer's network was clearly not a good indicator of the grade awarded for that paper. Similarly, for the analytic graders, $g = -1.394C + 0.161$, where $R^2 = 0.012$ and $p > 0.05$, suggesting that the degree of Pathfinder network similarity was not a good indicator of the grade awarded by analytic graders. The situation was slightly different for the LSA grader, $g = -7.772C + 0.978$, where $R^2 = 0.247$ and $p = 0.06$. The relationship between the Pathfinder network similarities and the score awarded by the LSA grader was marginally significant. This finding cautiously suggests that the implemented LSA grader may be more valid at grading than human graders. LSA grades are closely coupled to the elicitation of knowledge in the essay and to the degree to which the student essay exhibits a semantic knowledge space congruent with the expert computer grader representation.

DISCUSSION

The student's Pathfinder coherence was minimally correlated to holistic and analytic grades and the LSA grades. These results suggested that the student essay writers did not generally possess a coherent or expert-like knowledge structure regarding the topic. The particular exercise of essay writing did not facilitate the formation of coherent knowledge structures in the students. It is impossible to determine the extent to which this finding about essay writing is prescriptive. Further research is necessary to reveal the general role of different writing scenarios on knowledge formation.

As noted in the results section, for the LSA grader, the similarity between the student and grader Pathfinder knowledge networks was marginally indicative of the score awarded for that essay. In contrast, network similarity had no significant effect on the score awarded to the essay by either holistic and analytic graders. This finding suggests that LSA utilizes the elicited knowledge structure as a determinant of the grade awarded to an essay, which does not appear to be the case with human essay graders. It would therefore appear that the LSA grader uses a more valid approach to writing assessment than the human graders because of its stronger reliance on structural knowledge. An allied conclusion is that in those cases where LSA deviates from human graders, the LSA grader may offer a more valid grade than that grade awarded by human graders. Given the present experiment, it is impossible to assess the veracity of this conclusion. Further research is necessary to assess the validity of LSA vs. human generated essay grades.

Caution is necessary before generalizing from the relationship between underlying knowledge representations and the actual essays. It is possible, for example, that a student's underlying knowledge structure does not directly impact the ability of that student to express him- or herself effectively or coherently in writing. A student might have a solid domain expertise but an inability to write about that knowledge. No grading method can nor should completely omit the effects of surface characteristics of an essay, because the myriad of surface characteristics serve as bridges to understanding the deep structure of the essay. An important part of the grader's assessment of a student's performance is the grader's ability to understand the student's essay. Understanding is subject to the processes of adherence to good rhetorical form (Kintsch, 1998). Without adherence to the conventions of writing and reading discourse, the student's essay would be ineffective, because it would be impossible for the grader to transcend surface shortcomings to reconstruct the student's underlying knowledge structure. Pathfinder networks consider only the deep structure of student expertise, thereby potentially overlooking other influential factors in grading such as the importance of coherent surface structure.

If a grader assigns a good grade to a knowledgeable student who is a poor essay writer, is that grade any more or less valid than if another grader assigns a poor grade to the same writer? There are good and bad qualities to the same essay, and the two graders have capitalized on these disparate qualities. The grader who awards a poor grade is reflecting the writer's inability to express his or her knowledge cogently, while the grader who awards a good grade is echoing the writer's underlying strong knowledge of the topic. Given the clearly defined grading factors of the analytic grading method, it is probable that analytic graders would strike the best balance of deep and surface features when grading an essay. On the other hand, it may be holistic essay graders, who in using their overall impression of an essay, manage the most seamless integration of deep and surface features when grading an essay. The LSA grader must use the surface features of the essay to derive the deep structure. This process is akin to the stages of discourse processing, without the benefit of previous knowledge to aid the process.

Ultimately, the ideal balance of surface features and deep structure is left undetermined by this experiment. Analytic grading methods attempt to control for both structures, holistic grading methods allow the individual grader to determine the best mixture, and the LSA grader performs a top-down analysis from the surface features to the deep structure. Further research is necessary to determine the exact interplay of surface features and deep structure and to disambiguate their effects to arrive at the most valid form of writing assessment.

ACKNOWLEDGEMENTS

This research was completed while the author was a graduate student in the Department of Psychology at New Mexico State University. The author expresses his gratitude to Professor Peter Foltz for supervising this research, and to the LSA Group at the University of Colorado at Boulder for providing free access to the software tools used in grading the essays in this research.

REFERENCES

Berger, C., & Dershimer, C. (1993). *Using Technology to Measure Change in Students' Science Learning.* Paper presented at the meeting of the National Association for Research in Science Teaching, Atlanta, GA.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.

Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 271-302). Washington, DC: American Council on Education.

Goldsmith, T.E., & Davenport, D.E. (1990). Assessing structural similarity of graphs. In R.W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 75-87). Norwood, NJ: Ablex Publishing Corporation.

Diederich, P.B., French, J.W., & Carlton, S.T. (1961). *Factors in judgements of writing ability* (Research Bulletin RB 61-15). Princeton, NJ: Educational Testing Services.

Goldsmith, T.E., Johnson, P.J., & Acton, W.H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*, 88-96.

Gomez, R.L., Hadfield, O.D., & Housner, L.D. (1996). Conceptual maps and simulated teaching episodes as indicators of competence in teaching elementary education. *Journal of Educational Psychology, 88*, 572-585.

Gonzalvo, P., Cañas, J.J., & Bajo, M.T. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology, 86*, 601-616.

Huot, B.A. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*, 201-213.

Madigan, R.J., & Brosamer, J.J. (1991). Holistic grading of written work in introductory psychology: Reliability, validity, and efficiency. *Teaching of Psychology, 18*, 91-94.

Jonassen, D.H., Beissner, K., and Yacci, M. (1993). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge.* Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

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

McGaghie, W.C. (1996). *Comparison of knowledge structures with the Pathfinder scaling algorithm.* Paper presented at the meeting of the American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. ED 401 282)

Schvaneveldt, R.W., Durso, F.T., & Dearholt, D.W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 24, pp. 249-284). New York: Academic Press.

White, E.M. (1984). Holisticisim. *College Composition and Communication, 35*, 400-409.

Wilson, J.M. (1994). Network representations of knowledge about chemical equilibrium: Variations with achievement. *Journal of Research in Science Teaching, 31*, 1133-1147.