J. EDUCATIONAL COMPUTING RESEARCH, Vol. 37(3) 211-227, 2007

A COMPUTER-BASED APPROACH FOR DERIVING AND MEASURING INDIVIDUAL AND TEAM KNOWLEDGE STRUCTURE FROM ESSAY QUESTIONS

ROY B. CLARIANA
The Pennsylvania State University

PATRICIA WALLACE
The College of New Jersey

ABSTRACT

This proof-of-concept investigation describes a computer-based approach for deriving the knowledge structure of individuals and of groups from their written essays, and considers the convergent criterion-related validity of the computer-based scores relative to human rater essay scores and multiple-choice test scores. After completing a classroom-based course, undergraduate participants in a sophomore-level management course completed a 100-item multiple-choice final examination and then answered an extended-response essay question comparing four management theories. The essays were quantified with ALA-Reader software applying both sentence-wise and linear lexical aggregate approaches, and then analyzed with Pathfinder KNOT software. The linear aggregate approach was a better measure of essay content structure relative to the sentence-wise approach, with significant Spearman correlations of 0.60 and 0.45 with the human rater essay scores. The group network representations of low and high performing students were reasonable and straightforward to interpret, with the high group being more similar to the expert, and the low and high groups more similar to each other than to the expert. Suggestions for further research are provided.

Following Anderson (1984), we view structure as the essence of knowledge (p. 5), and human neural networks are the biologically plausible source of knowledge structure (Elman, 1993; Rumelhart & McClelland, 1986). In our connectionist

doi: 10.2190/EC.37.3.a
http://baywood.com

view, knowledge structure is the antecedent underpinning of knowledge and the precursor of meaning (Goldsmith & Johnson, 1990; Goldsmith, Johnson, & Acton, 1991). Specifically, an individual's knowledge structure is an expression of their language association neural network. Unique in each individual, this network is a sort of lexicon (i.e., albeit a multi-dimensional mental representation) that is at the most fundamental level of memory (pre-meaningful). We hold that measurers of knowledge structure are primarily measures of this association network (Deese, 1965) but are not necessarily measures of semantic meaning (e.g., word association approaches), though some do also measure meaning (e.g., concept maps).

The notion of structure of knowledge has become so main stream that a report by the National Research Council (2001) stressed the importance of classroom assessment that measures it. Knowledge structure has been measured in a variety of ways including card-sorting tasks, network diagrams, concept maps, and *Pathfinder* networks derived from word association data. Jonassen, Beissner, and Yacci (1993) provide details for 22 different ways to measure knowledge structure.

Compare-contrast type essay questions have also been used to assess relational understanding that is part of knowledge structure (Gonzalvo, Canas, & Bajo, 1994). Goldsmith, Johnson, and Acton (1991) state "Essay questions, which ask students to discuss the relationships between concepts, are perhaps the most conventional way of assessing the configural aspect of knowledge" (p. 88). It is rather critical to keep in mind that an essay contains different kinds of information, and that the scoring approach determines what is actually measured (Nitko, 1996), and most if not all essay scoring approaches do not intentionally measure knowledge structure.

But whether it is intentionally measured or not, essays contain a reflection of an individual's knowledge structure. For example, essay questions require students to recall information about a topic, select the most relevant aspects in the light of the particular question prompt, and then to organize the material into a logical and coherent response (Taber, 2003). Recalling, organizing, and integrating ideas while writing requires far more structuring of response than that required by objective test items (Jonassen & Wang, 1992; Valenti, Neri, & Cucchiarelli, 2003) and this structure ought to reflect the individual's internal knowledge structure. Thus, essays contain many aspects of knowledge structure, but as Goldsmith, Johnson, and Acton (1991) point out, "there is no simple and objective method to derive the structural relations from the written answers" (p. 88).

This current investigation purports to provide a simple and objective method for deriving and measuring individuals (and teams) knowledge structure from essays. In addition, two alternate scoring methods are considered, a within-sentence aggregation approach that focuses on key concepts within sentences (Koul, Clariana, & Salehi, 2005) and a new linear aggregation approach that considers key concepts both within and across sentences.

CONVERTING ESSAYS INTO CONCEPT MAPS

Lomask, Baron, Greig, and Harrison (1992) devised a method for scoring student essays based on essay content structure. As part of the Connecticut statewide assessment of science content knowledge, trained teacher raters converted student essays into concept maps and then hand scored the concept maps as a measure of science content knowledge and comprehension (Shavelson, Lang, & Lewin, 1994). For example, students were asked to "describe the possible forms of energies and types of materials involved in growing a plant and explain fully how they are related." One student responded in part:

> Plants need solar energy, water, and minerals. The chlorophyll is used to give it color. And I also think that the plant grows toward the suns direct (Lomask et al., p. 22).

This text sample is shown as a concept map in Figure 1. Plants are the central and most connected concept node in the map representation of this text. Note that the terms in the map differ from the student's written text response because several more or less equivalent terms (synonyms and metonyms) can stand for the same concept (Shavelson, Lang, & Lewin, 1994, p. 7).

The resulting concept maps were scored by Lomask's team relative to an expert map based on (a) the *number of concept terms* used by the student, (b) the *number of correct links* between concepts, and (c) the *expected number of links* among all of the concepts mentioned in an essay. These concept map scoring features in Lomask's et al. rubric considered the extent of knowledge, declarative knowledge, and one aspect of knowledge structure.

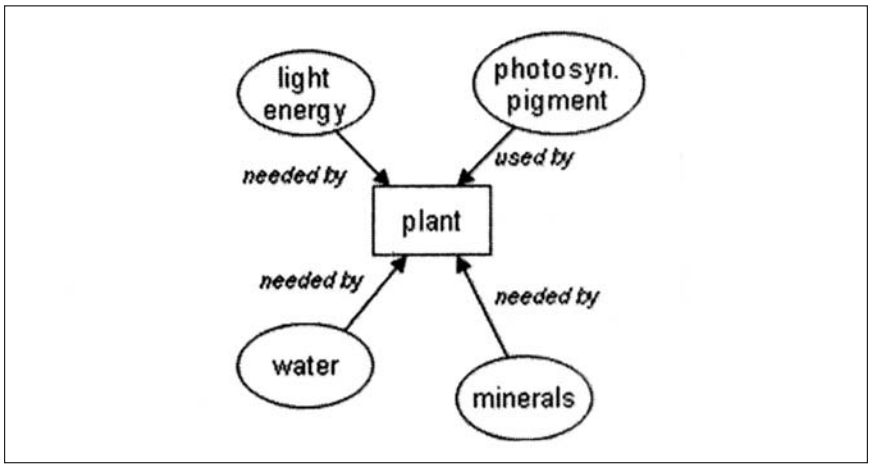

Figure 1. Concept map derived from a portion of a student's essay (in Lomask et al., 1992).

SOFTWARE FOR CONVERTING ESSAYS INTO NETWORK GRAPHS

Along these lines, Clariana (2004) developed a software tool called *ALA-Reader* (ALA, analysis of lexical aggregates) for converting short essays into concept map-like *Pathfinder* network representations (*PFNet*) by aggregating pre-selected key concepts from the essays at the *sentence* level (Shavelson, 1972). For example, these two sentences taken from a participant's essay in the present investigation (with the key words underlined) were analyzed by *ALA-Reader* and then represented as a *PFNet* (see Figure 2).

> Humanists believed that job satisfaction was related to productivity. They found that if employees were given more freedom and power, then they produced more.

ALA-Reader aggregated the individual sentence networks into a *Pathfinder* proximity data file. Then Knowledge Network Organizing Tool software (*KNOT,* Schvaneveldt, 1990) was used to convert the proximity data into a single network *PFNet* representation of the essay (see Figure 2). *Productivity* is the most connected concept term with five links (degrees) and so is the central term.

PFNet representations of essays from *ALA-Reader* can be compared to the *PFNet* representations of a referent *PFNet* (i.e., an expert essay) to produce a measure of structural similarity between the student and expert essays. Using this approach, a student's essay can be reduced to a number that represents the amount of similarity between the student's essay and the expert referent. Similarity is defined as the intersection of the links in common between the student and expert *PFNets* divided by the union of all links in those two *PFNets*.

Using this *ALA-Reader* sentence aggregation approach (e.g., the co-occurrence of key concepts within each sentence of the essay), Koul, Clariana, and Salehi (2005) considered the criterion-related validity of *ALA-Reader* essay scores. In pairs, teachers ($N = 22$) enrolled in a graduate-level course researched a science topic online and created a concept map of the topic. Later, individually they wrote a short essay from their concept map. The concept maps and essays were scored by the computer-based tools and by human raters using rubrics. The *ALA-Reader* computer-based essay scores, aggregated at the sentence level, were an adequate measure of essay content compared to human-rater scores ($r = 0.71$).

SENTENCE LEVEL AGGREGATION VERSUS AGGREGATION ACROSS SENTENCES

The present investigation uses *ALA-Reader* to aggregate key text terms at the sentence level (within sentences), but adds another method that aggregates terms both within and across sentences (linearly). This new linear approach was developed because (a) the sentence aggregation approach obtains disconnected graphs, (b) writing is a *serial* artifact of a multi-dimensional mental representation,

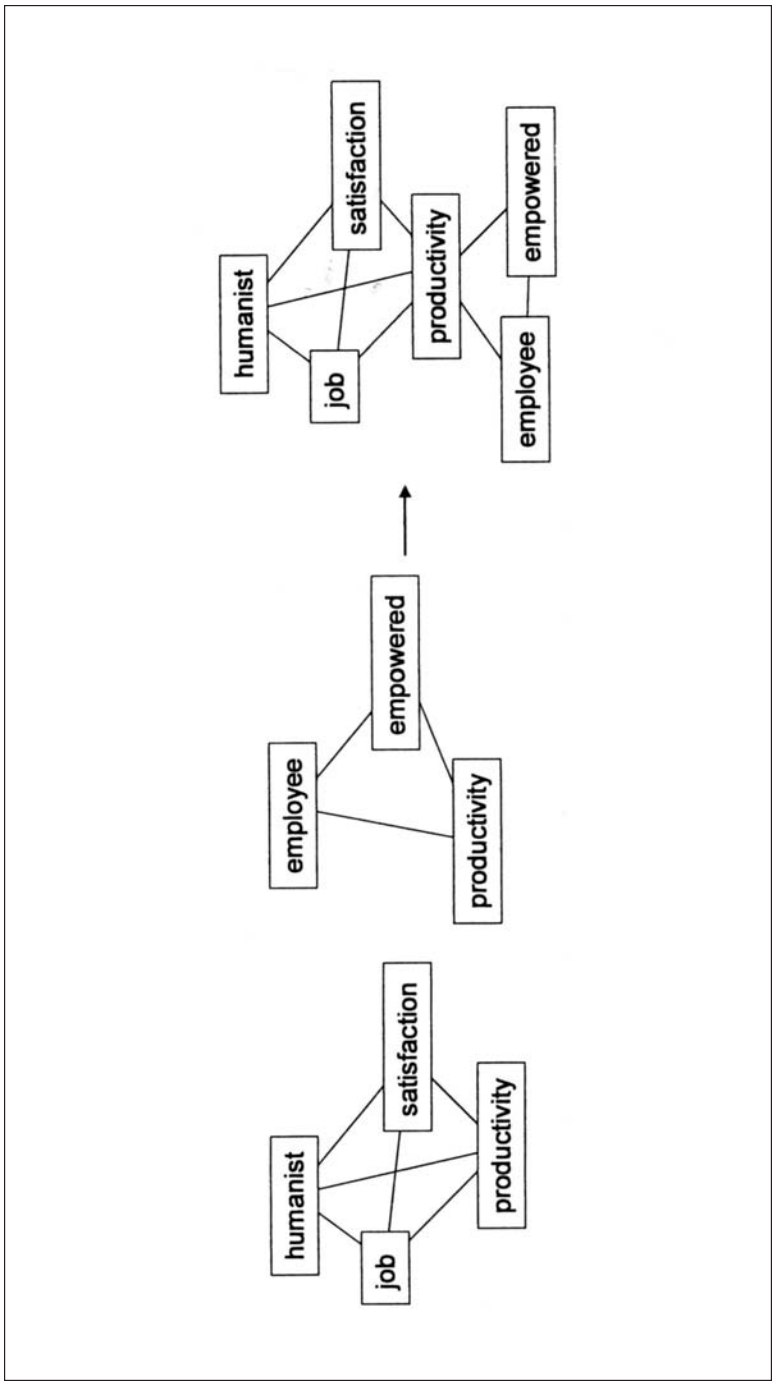

Figure 2. *ALA-Reader* sentence-wise *PFNet* representations of the first sentence (left panel), the second sentence (middle panel), and the combined aggregate of the first and second sentences (right panel).

and (c) because of the body of research that shows the influence of serially contiguous associations on cognitive processing, such as free association (Deese, 1965), associative models of memory (Anderson, 1983; Shavelson, 1974), the effects of text signals (Brooks, Dansereau, Spurlin, & Holley, 1983), and Bayesian essay scoring (Rudner & Liang, 2002). On the other hand, Landauer, Foltz, and Laham (1998) argue that word order is not very important for indexing an essay (Landauer, Laham, Rehder, & Schreiner, 1997), although their approach, Latent Semantic Analysis (LSA), must depend on the serial contiguity of words at least partially since the linear order of words in the training texts inevitably and profoundly influences the structure of the final LSA semantic space.

Therefore, a linear aggregation approach was developed and added to the *ALA-Reader* software for this investigation. During linear analysis, initially, the developing network is fairly linear (see left panel of Figure 3), but as key terms are repeated in the essay, such as *productivity*, the representation begins to fold back on itself creating a network structure. This network structure can be represented as a force-directed graph (see right panel of Figure 3). As with the sentence-wise approach (compare to the right panel of Figure 2), *productivity* is the central and most connected concept term in this text sample; however the linear approach reduces the text to only six links while the sentence-wise approach has nine, the same six links plus three other links. In addition, the linear approach will always produce a network structure that is a connected graph, while the sentence-wise approach is likely to produce multiple unconnected clusters (a disconnected graph), especially if the text is not coherent or uses pronouns or other substitute words extensively (the problem of linguistic anaphoric reference).

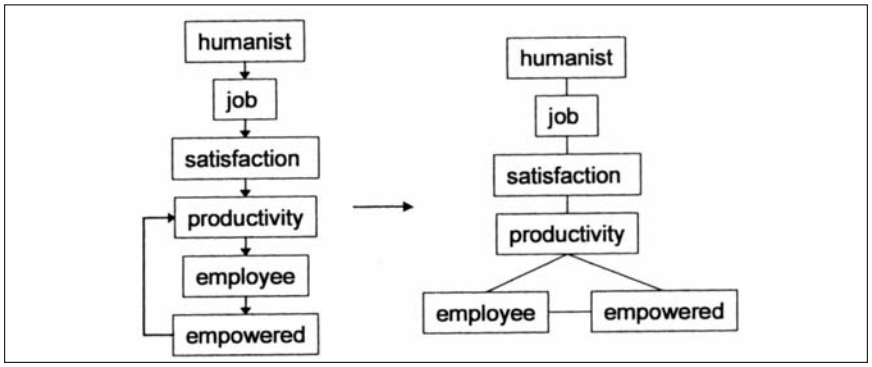

Figure 3. Using the same text that was used to derive the *PFNet* in Figure 2, the *ALA-Reader* linear representation of the first and second sentences (left panel) and the force-directed graph of this same representation (right panel).

Besides *ALA-Reader*, there are a number of computer-based tools for scoring essays (Valenti et al., 2003). Of these, *ALA-Reader* is most similar to Conceptual Rater software (called *C-Rater*) in that both approaches: (1) evaluate whether a response contains information related to specific domain concepts (writing style is not assessed), (2) do not require a large collection of graded answers for training, (3) can score short or long essays, and (4) use a correct-answer referent for grading purposes. This is important because most of the other computer-based tools are costly to setup, involving the development of a large dataset of example text and marked responses (at least 500 essays) either for training (neural network, Bayesian, and natural language processing approaches) or calibration (setting weight values through correlation analyses). Further, the training or calibration for these other computer-based essay scoring tools only apply to a specific essay prompt and content sub-domain, and so substantial cost is incurred for each new writing prompt. *ALA-Reader* is neither prompt nor domain specific, and so does not have a high setup cost. It can score essays given only a list of important terms and an expert referent essay.

USING ALA-READER TO REPRESENT TEAM KNOWLEDGE

Perhaps more radically, *KNOT* software has a data averaging feature that can be used to combine the proximity files from several students' essays into a group *or team* representation of that content. A group representation can provide a clearer view of the team's knowledge structure by averaging out the idiosyncrasies in individual student responses. Using this feature, it is possible to compare the knowledge structure of a group during pre- and post-instruction to determine growth and change of the group; and also to compare subgroups within a group, for example, low versus high performing students.

SUMMARY

This current investigation reports the results of the ongoing development of *ALA-Reader* software by adding a new linear aggregation approach that knits terms together both within and across sentences. The linear scores will be compared to within sentence aggregation scores as well as to human rater scores and multiple-choice test scores. In addition, the use of this tool for measuring team (group) knowledge structure will be considered. This formative information will contribute to the next round of development of the software tools.

METHOD

Participants

One section of a sophomore-level course that must be taken by all Business majors, Management 201, consisting of 29 students was selected as a sample of

convenience for this investigation. There were 12 females and 17 males. The smaller number of female students participating in the study is directly related to the disproportionate number of female students versus male students that are enrolled in the Management major.

Instructional Setting

Management 201, Managing in the 21st Century, considers the historical roots of management theory and the basic functions of management including the essential management skills for planning, organizing, leading, and controlling a department or an entire organization The course, which is required of all School of Business students, utilizes experiential learning activities such as cases, team work, and exercises in order to realize the role of the manager in a changing workplace. The course met twice a week for a semester.

During the regularly scheduled examination week at the end of the semester, all students completed the customary final examination for the course (worth 25% of their final course grade) and also answered an extended-response essay question for extra credit. The final examination, which was designed to evaluate the student's ability to apply concepts learned in the course, consisted of a 100-item multiple-choice test that was comprehensive in scope and was delivered by computer using *ExamView* testing software. The final examination score mean was 78.0 (sd = 8.3).

The essay question was typed by the participants using Microsoft Word software and the finished essay was printed in the computer lab and submitted to the course instructor. Participants were given a maximum time of 30 minutes for the essay. The extended-response essay question stated, "Describe and contrast in an essay of 300 words or less the following four Management theories: Classical/scientific management, Humanistic/Human Resources, Contingency, and Total Quality Management."

Essays Scored by ALA-Reader and by Raters

The 29 essays contained 7,917 total words, of which 1,289 were unique. The longest essay was 490 total words, the shortest was 97 words, and the average essay length was 273 words. The 20 most frequent stop words (i.e., *a, and, the*; Luhn, 1959) accounted for about 33% of all words, and the top five stop words in order are *the* (5.2% of all words), *and* (3.3%), *of* (3.3%), *a* (3.1%), and *to* (3.2%).

ALA-Reader software must be provided with a list of key terms (maximum limit of 30) with synonyms and metonyms. In order to optimize the value of the information across the entire set of 29 essays, a frequency list containing all of the content terms used in the participants' essays was prepared. A common important term, such as *employee*, had a high frequency count, occurring 90 times in the 29 essays. *ALA-Reader* analysis requires that functionally equivalent terms,

such as *employee*, *worker*, and *subordinate*, be grouped together as the same concept list word. When the synonyms and metonyms for *employee* are combined, the number of occurrences of this concept in the 29 essays increases to 150 times, or about 5 occurrences per essay. The frequency list of terms was used to create a list of the 30 most important conceptual terms. The 30 important concept terms arranged in order of frequency of occurrence are: Management (manager, boss, employer), Employee (workers, subordinates), Company (corporation, organization), Work (job, task), Classical (scientific), Environment (situation, place), Productivity (performance, output), Contingency, Efficiency (well, best), Empowerment (freedom, satisfied), Needs (encourage, appreciate), Humanistic, Concerns (stress, problem), Leadership, Plan (process, implementation), Human, Feelings (happy, positive, intrinsic), Customer (consumer, everyone), Relationship, Motivation, Individual (unique), Goal, Success, Machine (mass production), TQM, Quality (reputation), Pay (salary), Serve (service), Measured, and Focus (emphasis).

Next the list of 30 terms was stemmed. Stemming refers to the process of carefully removing suffixes to obtain word roots that have similar meanings. For example, "*educ"* is the common stem for educate, education, educates, educating, educational, and educated (Rudner & Liang, 2002). Stemming may be a crucial step in essay analysis, though the potential value of stemming is not clear. Probably, stemming is better in some cases and not in others, depending on both the specific terms involved and the essay domain content, for example whether it is important to distinguish between educated and educational or not. In previous *ALA-Reader* studies, stemming was easier and direct for the biology content used, where specialized technical vocabulary comprised most of the important terms in an essay. For example, Yager (1983) estimated that introductory college biology courses have more new vocabulary, about 3,000 unique terms, than even foreign language courses. The management essays in this investigation contained relatively fewer specialized words, and substantially more synonyms than the previous study involving biology content.

All of the essays were also scored by a pair of human raters using a generic content rubric. The essay scoring rubric considered whether the content was clear, relevant, accurate, and concise. The four essay sub-sections (classical, humanistic, contingency, and TQM) were scored separately and each section was worth 2.5 points, thus the maximum essay score was 10 points. To simplify analysis comparisons, the *ALA-Score PFNet* common subtest raw scores were linearly transformed to a 2.5 point scale.

The Cronbach alpha reliability for the essay scores based on four sources, the two raters and the two *ALA-Score* scores (sentence and linear) are: classical subtest alpha = 0.70, humanistic subtest alpha = 0.61, contingency subtest alpha = 0.65, TQM subtest alpha = 0.69, and total test alpha = 0.81.

RESULTS

Data analysis consists of the subtest means and standard deviations, evaluation of the criterion-related validity of the computer-based essay scores relative to the rater scores and the multiple-choice tests, and finally description of the group network structure for low and high performing students. Means and standard deviations for raters' and for the *ALA-Reader* essay scores are shown in Table 1. Rater 2 tended to award relatively higher essay scores, while Rater 1 was stricter. The *ALA-Reader* mean scores tended to fall between the two human rater mean scores, with the *ALA-Reader* sentence-wise scores greater than the linear-aggregate scores.

Spearman rank correlation for the combined essay test score (i.e., sum) for Rater 1 and Rater 2 was $r_s = 0.71$ (see Table 2). Rater 1 total essay scores were most related to the 100 item multiple-choice test scores, $r_s = 0.53$. Rater 2's and *ALA-Reader* linear-aggregate essay scores were also significantly related to the multiple-choice test scores, although *ALA-Reader* sentence-wise essay scores were not. Shavelson, Ruiz-Primo, and Wiley (2000) reported correlations between multiple-choice tests and measures of knowledge structure (i.e., concept maps) in the range of 0.45 to 0.64, and proposed that both measurement methods tap aspects of declarative knowledge that are interrelated; concept maps reflect the structure of knowledge while multiple-choice tests reflect the extent of knowledge. Thus the *ALA-Reader* measures (i.e., knowledge structure) should not necessarily correlate highly with the multiple-choice test scores (extent of knowledge). The *ALA-Reader* linear total essay scores correlated significantly with

Table 1. Essay Score Means and Standard Deviations

	Essay subtest				
	Classical*	Humanist*	Contiguity*	TQM*	Sum**
Rater 1	1.14 (0.75)	0.98 (0.54)	0.66 (0.57)	0.64 (0.63)	3.41 (2.01)
Rater 2	1.84 (0.71)	1.91 (0.52)	1.47 (0.79)	1.53 (0.80)	6.76 (2.34)
ALA-Reader (sentence)	1.77 (0.50)	1.50 (0.53)	1.09 (0.48)	1.58 (0.84)	5.94 (1.47)
ALA-Reader (linear)	1.47 (0.44)	1.21 (0.51)	1.14 (0.63)	0.90 (0.41)	4.73 (1.39)

*2.5 points maximum; **10 points maximum.

Table 2. Spearman Correlations of Essay Scores and the Multiple-Choice Test Score

	MC test	R1	R2	ALA (S)	ALA (L)
MC test	1				
Rater 1	0.53**	1			
Rater 2	0.43*	0.71**	1		
ALA-Reader (sentence)	0.17	0.47*	0.29	1	
ALA-Reader (linear)	0.39*	0.60**	0.45*	0.73**	1

*Sig. at .05. **Sig. at .01.

Rater 1 and Rater 2 total essay scores (r = 0.60 and 0.45). The *ALA-Reader* sentence-wise total essay scores correlated less well with Rater 1 and Rater 2 total essay scores.

Group Comparisons

The students were grouped into high (n = 14) and low (n = 15) groups based on median split of their human rater essay scores, and then the individual essay proximity files within each group were averaged to obtain a *PFNet* for each group. Then the similarity of the low and high groups' *PFNets* and the expert's *PFNet* were computed (see Table 3).

Similarity is defined as the intersection of the links in common between two *PFNets* divided by the union of all unique links in the two. Similar to a percent test score that ranges from 0 to 100%, similarity ranges from zero, no similarity, to one, perfect similarity. The high group's essay network structure was more similar to the expert than was the low group's (i.e., 0.31 compared to 0.19), and the high and low groups were most like each other (i.e., 0.41).

Both groups' network structure (see Figure 4) showed *employee* and *management* as the central and most connected concepts. The terms *company*, *environment*, *individual*, *productivity*, and *work* were also important concepts for both low and high performing groups. The four super-ordinate management theory categories from the essay prompt (classical, humanistic, contingency, and TQM) were all associated with *management*. Action terms such as leadership and success were also associated with *management*, while emotion-related terms such as *feelings*, *concerns*, and *motivation* were associated with *employee*. Both groups associated *productivity* and *pay*, although the high group had a richer representation of productivity that included customers and efficiency. The main differences between the *PFNets* displayed in Figure 4 are that the high-performers had relatively more terms that were central (e.g., with more than 1 link) and the association structure was more extensive and complex.

Table 3. Similarity of the Expert's, and the Low and High Groups' Essay *PFNets*

	Exp.	High	Low
Expert	1		
High group (n = 14)	0.31	1	
Low group (n = 15)	0.19	0.41	1

In general, an expert's extensive knowledge base is organized into elaborate, richly connected and integrated structures; whereas novices tend to posses less domain knowledge and a less coherent organization of it (personal communication from Reviewer B; Chi, Glaser, & Farr, 1988; Chi, Glaser, & Rees, 1982), and this qualitative difference has been shown even for "stronger novices" compared to "weaker novices" (Zajchowski & Martin, 1993). The *PFNets* in Figure 4 clearly demonstrate this idea; the high-performers knowledge structure as represented by the group *PFNet* (see the right panel of Figure 4) is visually more integrated and elaborated than that of the low-performers (see the left panel of Figure 4).

DISCUSSION

This investigation described a computer-based approach using *ALA-Reader* software to aggregate key terms at either the sentence level or else linearly across sentences. The data from *ALA-Reader* were analyzed by Pathfinder *KNOT* software to derive the knowledge structure of individuals and of groups from their written essays. Based on the linear aggregation approach compared to human rater scores (r = 0.60 and 0.45) and to performance on a multiple-choice test (r = 0.39), the convergent criterion-related validity of the computer-generated essay scores was supported to some degree. Further, the representations of group knowledge structure seemed reasonable and consistent with expectations (e.g., the high group is more similar to the expert than the low group). This supports the premise that this approach using *PFNet* representations of *ALA-Reader* text aggregate data provides a method for comparing the knowledge structure of sub-groups in a classroom to each other and to some standard. Thus the approach may also be useful for measuring team knowledge structure.

There are limitations and caution notes regarding the findings of this investigation. Writing prompts are a critical consideration in essay writing and scoring. Following Gonzalvo, Canas, and Bajo's (1994) use of compare-contrast type essay questions to assess relational understanding that is part of knowledge structure, in this investigation, students were asked to "Describe and

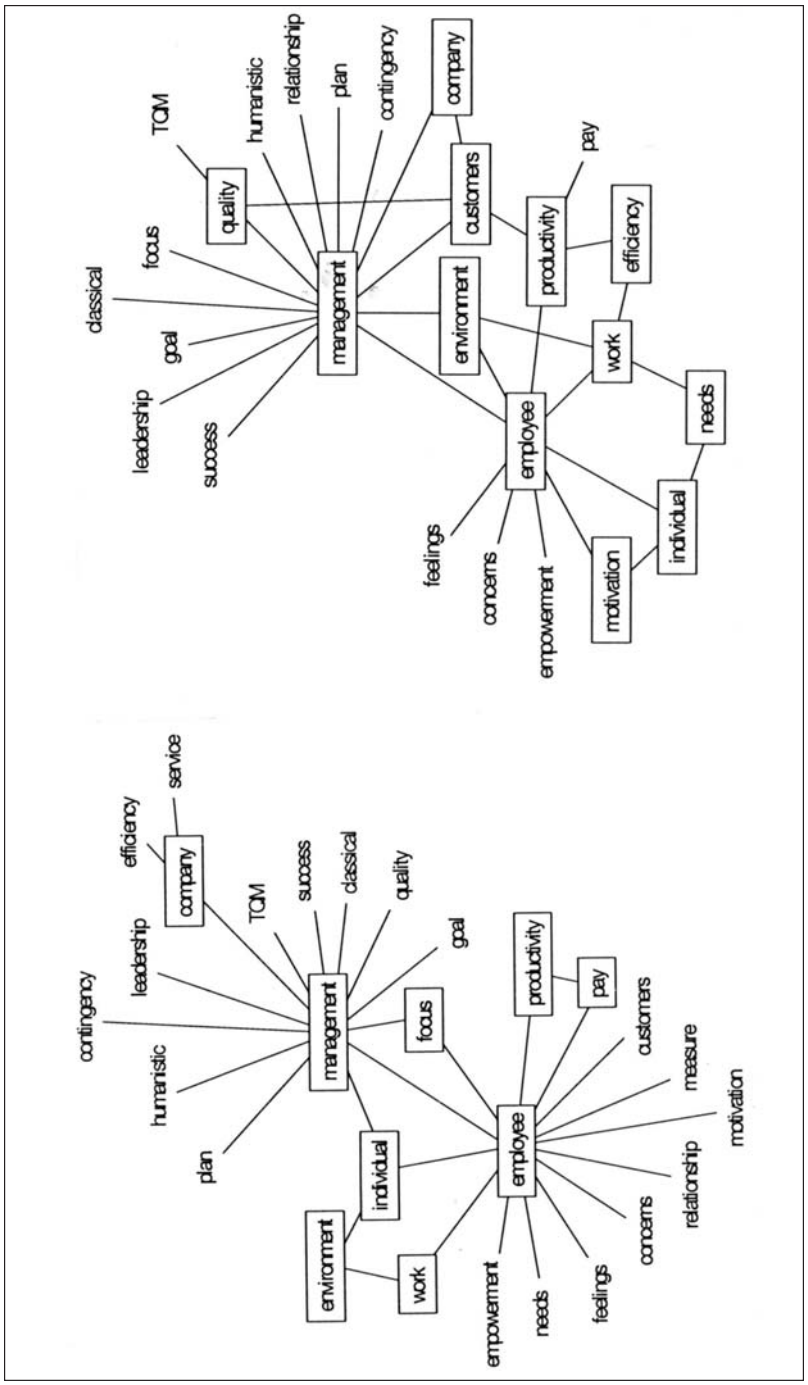

Figure 4. *PFNets* of the low (left panel; $n = 15$) and high (right panel; $n = 14$) groups.

contrast . . . the four management theories." However, all students "described" but very few of the students followed the "contrast" aspect of the prompt and so the essays are less likely to show integration or elaboration. Disregarding the "compare-and-contrast" part of a writing prompt may be quite common (Biggs, 1992). Such "knowledge telling" results in essays that are little more than un-integrated lists of terms, and so is less likely to reveal the students' knowledge structure.

Another limitation of this investigation is that human raters' essay scores are not necessarily an appropriate measure of knowledge structure (Shavelson, Ruiz-Primo, & Wiley, 2000). What do raters actually evaluate when they judge student essays holistically (Barritt, Stock, & Clark, 1986)? In this case, the raters were not intentionally looking for the student's knowledge structure; they were looking for a meaningful and complete response to the prompt. However, we hold that among the many other things present in an essay, essays reflect the structure of the student's association network (knowledge structure). Note that the *ALA-Reader* measure of knowledge structure does significantly correlate with two raters' holistic essay scores (e.g., $r = 0.60$ and 0.45).

Another limitation of this investigation involves the centrality of the expert essay referent for scoring purposes. The students' scores depend entirely on the quality of the expert essay referent, since the score is the number of links in common between the student essay *PFNet* and the expert essay referent *PFNet*. Though the essay prompt is well constrained, it seems likely that different experts could write different essays, and so the resulting student scores would vary depending on which expert referent essay was used for the analysis. Further research is needed in order to specify how to derive the expert essay referent.

For the express purpose of essay marking, this approach is probably less powerful than other approaches, such as Latent Semantic Analysis. Although further research and development are needed, since *ALA- Reader* is free shareware, then this approach has promise right now as a low cost measure of knowledge structure.

Considering the next steps, instruction intends to increase not only the number of what Novak (2003) refers to as "valid notions" but also to decrease the number of "invalid notions." The approach described in this investigation can be used for instructional purposes such as highlighting correct, incorrect, and missing propositions in students' essays. For example, the participants' network representations contain correct links that agree with the referent network representations, and these "correct" propositions could be pointed out, such as with check-plus marks beside sentences, green highlighting, or underlying. Sentences without such highlights could be reconsidered by the student. Also, the links contained in the expert essay that were not included in the student's essay could be pointed out to the student in various ways, such as with hints and clues, so that the student could add this missing content to

their essay. Future development of the free *ALA-Reader* software will include these features.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University.

Anderson, R. C. (1984). Some reflections on the acquisition of knowledge. *Educational Researcher, 13*(10), 5-10.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication, 37*, 315-327.

Biggs, J. (1992). A qualitative approach to grading students. *Higher Education Research and Development Society of Australasia News, 14*(3), 3-6

Brooks, L. W., Dansereau, D. F., Spurlin, J. E., & Holley, C. D. (1983). Effects of headings on text processing. *Journal of Educational Psychology, 75*, 292-302.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. S. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 1-75). Hillsdale, NJ: Erlbaum.

Clariana, R. B. (2004). *ALA-Reader.* Downloaded May 3, 2006, from http://www.pesronal.psu.edu/rbc

Deese, J. (1965). *The structure of associations in language and thought.* Baltimore, MD: The Johns Hopkins Press.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Goldsmith, T. E., & Johnson, P. J. (1990). A structural assessment of classroom learning. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 241-253). Norwood, NJ: Ablex Publishing Corporation.

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

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

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Knowledge structure: Techniques for representing, conveying, and acquiring knowledge structure*. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Jonassen, D. H., & Wang, S. (1992). *Acquiring structural knowledge from semantically structured hypertext*. Proceedings of Selected Research and Development Presentations at the 14th Annual Convention of the Association for Educational Communications and Technology. (ERIC Document Reproduction Service No. ED 348 000).

Koul, R., Clariana, R. B., & Salehi, R. (2005). Comparing several human and computer based methods for scoring concept maps and essays. *Journal of Educational Computing Research, 32*(3), 261-273.

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

Landauer, T. K., Laham, D., Rehder, R., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mahwah, NJ: Erlbaum.

Lomask, M., Baron, J. B., Greig, J., & Harrison, C. (1992, March) *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science*. Paper presented at the annual meeting of the National Association of Research in Science Teaching, in Cambridge, MA.

Luhn, H. P. (1959). Keyword-in-context index for technical literature (KWIC index). (Technical report RC-127, IBM Corporation, Advanced Systems Development Division). Reprinted in C. K. Schultz (Ed.). (1968). *H. P. Luhn: Pioneer of information science: Selected works* (pp. 227-235). New York: Spartan Books.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Nitko, A. J. (1996). *Educational assessment of students*. Englewood Cliffs, NJ: Merrill.

Novak, J. D. (2003). The promise of new ideas and new technology for improving teaching and learning. *Cell Biology Education, 2*, 122-132.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *Journal of Technology, Learning, and Assessment, 1* (2). Retrieved May 5, 2005, from http://www.jtla.org

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing* (Vols. I and II). Cambridge, MA: MIT Press.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex Publishing Corporation.

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology, 63*, 225-234.

Shavelson, R. J. (1974). Some methods for examining content structure and cognitive structure in instruction. *Educational Psychologist, 11*, 110-122.

Shavelson, R. J., Lang, H., & Lewin, B. (1994). *On concept maps as potential "authentic" assessments in science* (CSE Tech. Rep. No. 388). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing (CREST).

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (2000). *Windows into the mind*. Based on an invited address, Facolta' di Ingegneria dell'Universita' degli Studi di Ancona, June 27, 2000. Retrieved May 5, 2005, from http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/Windows%20into%20the%20Mind_8_4_03_Final.doc

Taber, K. S. (2003). *Examining structure and context—Questioning the nature and purpose of summative assessment*. Seminar presentation to Cambridge International Examinations, University of Cambridge Local Examinations Syndicate, July 2003. Retrieved May 04, 2005 from http://www.leeds.ac.uk/educol/documents/00003134.htm

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, *2*, 319-330. Retrieved May 14, 2005 from http://jite.org/documents/Vol2/v2p319-330-30.pdf

Yager, R. E. (1983). The importance of terminology in teaching K-12 science. *Journal of Research in Science Teaching, 20*(6), 577-588.

Zajchowski, R., & Martin, J. (1993). Differences in the problem solving of stronger and weaker novices in physics: Knowledge, strategies, or knowledge structure? *Journal of Research in Science Teaching, 30*(5), 459-470.

Direct reprint requests to:

Dr. Roy Clariana
30 E. Swedesford Road
Malvern, PA 19355
e-mail: RClariana@psu.edu