Education Tech Research Dev
DOI 10.1007/s11423-009-9115-z

DEVELOPMENT ARTICLE

Deriving and measuring group knowledge structure from essays: The effects of anaphoric reference

Roy B. Clariana · Patricia E. Wallace · Veronica M. Godshalk

Abstract Essays are an important measure of complex learning, but pronouns can confound an author's intended meaning for both readers and text analysis software. This descriptive investigation considers the effect of pronouns on a computer-based text analysis approach, *ALA-Reader*, which uses students' essays as the data source for deriving individual and group knowledge representations. Participants in an undergraduate business course ($n = 45$) completed an essay as part of the course final examination. The investigators edited the essays to replace the most common pronouns (their, it, and they) with the appropriate referent. The original unedited and the edited essays were processed with *ALA-Reader* using two different approaches, sentence and linear aggregate. These data were then analyzed using a Pathfinder network approach. The average group network similarity values comparing the original to the edited essays were large (i.e., about 90% overlap) but the linear aggregate approach obtained larger values than the sentence aggregate approach. The linear aggregate approach also provided a better measure of individual essay scores (e.g., $r = 0.74$ with composite rater scores). This data provides some support that the *ALA-*

This is an expanded version of the paper: Clariana, R. B., Wallace, P. E., & Godshalk, V. M. (2008). Deriving and measuring group knowledge structure via computer-based analysis of essay questions: the effects of controlling anaphoric reference. In Kinshuk, D. G. Sampson, J. M. Spector, P. Isaías, & D. Ifenthaler (Eds.), *Proceedings of the IADIS international conference on cognition and exploratory learning in the digital age* (88–95). Freiburg, Germany: International Association for Development of the Information Society.

R. B. Clariana (✉)
Pennsylvania State University, Penn State Great Valley, 30 East Swedesford Road, Malvern, PA 19355, USA
e-mail: RClariana@psu.edu

P. E. Wallace
The College of New Jersey, Ewing, NJ, USA
e-mail: PWallace@tcnj.edu

V. M. Godshalk
University of South Carolina, Beaufort, USA
e-mail: godshalk@uscb.edu

Published online: 21 March 2009

Reader linear approach is adequate for capturing group knowledge structure representations from essays. Further development of the *ALA-Reader* approach is warranted.

Keywords Mental models · Measuring knowledge structure · Pathfinder networks · Essays

Understanding and measuring the progress of learning in complex domains is an important issue for instructional designers, instructors, and researchers. An increasingly common way to measure such knowledge in the lab is by comparing an individual's or even a group's mental model (Craik 1943; Johnson-Laird et al. 1998) to some referent mental model such as from a more advanced peer or an expert (Seel 1999). A number of recent technology-based approaches for measuring aspects of individual and group mental models are under development (Johnson et al. 2006), such as Analysis of Constructed Shared Mental Model (ACSMM), Surface, Matching, and Deep Structure (SMD), and Model Inspection Trace of Concepts and Relations (MITOCAR). These approaches use various methods for eliciting knowledge structure such as concept maps and essays.

Compare-contrast type essay questions have been used to assess relational understanding that is part of knowledge structure (Gonzalvo et al. 1994). Goldsmith et al. (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; the scoring approach determines what is actually measured. Most if not all essay scoring approaches, human or computer-based, do not intentionally measure knowledge structure. Whether intentionally measured or not, essays contain at least a reflection of an individual's knowledge structure, a snapshot of their mental model.

This investigation considers a method called Analysis of Lexical Aggregates (e.g., *ALA-Reader* software) that uses students' essays as the data source for deriving individual and group knowledge representations (Clariana & Wallace 2007). *ALA-Reader* essay-derived knowledge structure is not a complete measure of an individual's mental model but likely represents a critical aspect of it; nor is it a direct measure of essay content though it likely captures a facet of the essay's content to a greater or lesser extent.

ALA-reader research and development

The lexical aggregate approach used in this investigation is based on a concept map scoring approach (*ALA-Mapper)* described by Taricani and Clariana (2003, 2006) but applied to text passages. The *ALA-Reader* text analysis method is described in more detail below and in Clariana and Wallace (2007) but in brief, *ALA-Reader* aggregates from the essay preselected key terms including synonyms and metonyms at either the sentence level (Shavelson 1974) or linearly across sentences (Clariana &Wallace 2007) and saves this information into a link array file for further analysis.

Clariana and Koul (2004) used *ALA-Reader* software to score students' essays on the structure and function of the heart and circulatory system relative to an expert's essay. At that time the software could only analyze using the sentence aggregate approach. For benchmark comparison, the essays were also scored by 11 pairs of human raters and these

11 scores were averaged together into a benchmark composite essay score. Compared to the composite score, the *ALA-Reader* scores ranked 5th out of 12, with an $r = 0.69$ (i.e., the 12 scores' ranged from $r = 0.11$ to 0.86).

Koul et al. (2005) also used the *ALA-Reader* sentence aggregate approach to score students' essays on the structure and function of the heart and circulatory system. Working in pairs, participants researched this topic online and created concept maps using *Inspiration* software. Later, using their concept map, participants individually wrote a short essay. The concept maps and essays were scored by *ALA-Mapper* and *ALA-Reader* relative to an expert's map and essay, by another software tool called Latent Semantic Analysis (*LSA*), and by 11 pairs of human raters using two different rubrics. As in the previous study, the rater scores were averaged together into a composite essay score. Compared to the raters' composite score, the *ALA-Reader* scores ranked 5th out of 13, with an $r = 0.71$ (the 13 scores ranged from $r = 0.08$ to 0.88) and *LSA* scores were 9th out of 13, with an $r = 0.62$ indicating that *ALA-Reader* performed more like the composite of the human raters than did *LSA*.

Clariana and Wallace (2007) used *ALA-Reader* to score essays relative to an expert referent and to establish and compare group average knowledge representations from those essays. As part of their final course examination, undergraduate business majors were asked to write a 300-word compare-and-contrast essay on four management theories that were covered during the course (this is a relevant and high stakes essay). The essays were scored by *ALA-Reader* using both a sentence and a linear aggregate approach. To provide a benchmark, the essays were also separately scored by two human raters who's Spearman *rho* inter-rater reliability was $\rho = 0.71$. The linear aggregate approach obtained larger correlations with the two human raters ($\rho_{rater1} = 0.60$ and $\rho_{rater2} = 0.45$) than did the sentence aggregate approach ($\rho_{rater1} = 0.47$ and $\rho_{rater2} = 0.29$). In addition, group average network representations of low and high performing students were reasonable and straightforward to interpret, the high group was more like the expert, and the low and high groups were more similar to each other than to the expert.

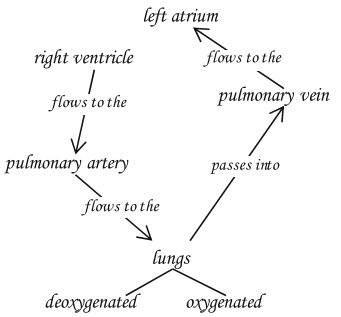
These three studies show a moderate correlation between human rater essay scores and *ALA-Reader* scores, although the correlations were considerably better in the first two studies. On further reflection, the key terms in the essays in the first two studies used mostly technical biology vocabulary that did not lend itself to pronoun referents (almost no pronouns were used) while essays in the third study used more general vocabulary that included a number of synonyms for key terms, such as manager, supervisor, and boss for the key term 'management', and that included a high frequency of pronouns. In addition, the first two studies used the sentence aggregate approach and obtained an adequate measure of essay performance, while in the third study, the linear aggregate approach provided a satisfactory measure of essay performance but the sentence aggregate approach did not. This present investigation considers these issues as part of the ongoing development of this tool. First, in order to increase essay comparison coherence, students in the present investigation were given a list of 29 key terms in the essay prompt so that these terms would more likely be included in the essays. Second, the sentence and linear aggregation approaches are both used to see if one is better than the other for group comparisons and for individual comparisons. Finally, the effects of pronouns were examined by manually editing high frequency pronouns in the essays to their referents before analysis and then comparing the original and edited versions of the essays. The next section describes the *ALA-Reader* approaches and also how group average representations were established.

Automatic essay analysis by ALA-reader software

The *ALA-Reader* essay analysis approach was adopted directly from the *ALA-Mapper* concept map analysis approach (Taricani & Clariana 2003). How does *ALA-Mapper* analyze concept maps and what is the relationship between the *ALA-Mapper* and *ALA-Reader* approaches? Concept maps are sparse representations of propositions. A proposition consists of two nodes connected by a labeled link, and is more or less a noun–verb–noun combination. The nodes are the terms or concepts that are being considered; links are lines that connect nodes; and link labels are phrases such as 'has a', 'is a', and 'leads to'. Note that link labels may not be critical for concept map analysis. Harper et al. (2004) reported that the correlation between just counting link lines (i.e., node–node) compared to counting valid propositions (i.e., node–label–node) in the same set of maps was $r = 0.97$, suggesting that link labels add little additional information over just counting links. *ALA-Mapper* converts node–node information into a mathematical form we refer to as a link array (see Fig. 1) that can be further analyzed by various approaches for example by multi-dimensional scaling, cluster analysis, and in this case, by Pathfinder network scaling.

The *ALA-Reader* sentence aggregate approach was developed to analyze at the sentence level because sentences are an important unit of text organization. Sentences contain one or more propositions; the sentence aggregation approach seeks to capture the important node–node associations represented by propositions in sentences. To analyze sentences in text, first *ALA-Reader* disregards all of the words except for preselected key terms and then replaces the synonyms and metonyms of key terms with the appropriate key term. Then the key terms that co-occur in the same sentence are entered into a proximity array, the lower triangle of an n-by-n array containing n(n-1)/2 elements (e.g., the seven terms in Fig. 1 above requires a link array of 21 elements; and in this current investigation 29 key terms require 406 elements). A '1' entered in the proximity array indicates that two key terms co-occurred in the same sentence and a '0' indicates that those two key terms did not occur in the same sentence. The software continues to aggregate sentences into the proximity array until all the text is processed (see Fig. 2), the final link array for a text passage contains only 1s and 0s, links are aggregated, not added, across sentences.

The *ALA-Reader* linear aggregate approach is similar to the sentence aggregate approach, except that sentences do not matter, links between adjacent key terms both within and across sentences are entered in the link array as these occur during a linear pass through the text. To analyze text, as with the sentence aggregate approach, first *ALA-Reader* disregards all of the words except for preselected key terms, and the synonyms and



Link Array

	a	b	c	d	e	f	g
a left atrium	-						
b lungs	0	-					
c oxygenate	0	1	-				
d pulmonary artery	0	1	0	-			
e pulmonary vein	1	1	0	0	-		
f deoxygenate	0	1	0	0	0	-	
g right ventricle	0	0	0	1	0	0	-

Fig. 1 Example concept map and its link array (from Taricani & Clariana 2003, 2006)

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

	humanist	job satisfaction	productivity	employees	empowered
humanist	--				
job satisfaction	1	--			
productivity	1	1	--		
employees	0	0	0	--	
empowered	0	0	0	0	--

sentence 1

→

	humanist	job satisfaction	productivity	employees	empowered
humanist	--				
job satisfaction	1	--			
productivity	1	1	--		
employees	0	0	1	--	
empowered	0	0	1	1	--

sentence 1 and sentence 2

Fig. 2 Example text passage and its sentence-level aggregation

metonyms of key terms are replaced with the appropriate key terms. Then the software begins to search through the text from the beginning to the end sequentially. The software adds a '1' in the proximity link array to indicate a link between a pair of consecutive terms, each succeeding term is linked to the next key term found. The software continues to aggregate linearly into the array until all of the text is processed.

The linear aggregate of the text example from Fig. 2 is displayed in Fig. 3. Note that proximity link arrays can be visually displayed as force-directed graphs (e.g., a Pathfinder networks or PFNETs). These graphs are node–node representations, and in this case represent the key word propositions in the text as a graph. In Fig. 3, the key term 'productivity' has the most links (e.g., three links) and so is the central or most important concept in this example text.

In Fig. 3, there are five '1s' in the linear aggregate link array compared to six '1s' in the sentence aggregate link array shown previously in Fig. 2; the same text passage produced a slightly different link array when analyzed by the sentence approach compared to the linear approach. Typically in key-term rich sentences, which are defined as three or more key terms per sentence, the sentence aggregate approach includes more links relative to the linear aggregate approach. Key-term rich sentences are indicative of expert responses. However, the sentence aggregate approach tends to relatively under-specify associations when key terms in sentences are sparse, for instance with poor or novice writers, and also

	humanist	job satisfaction	productivity	employees	empowered
humanist	--				
job satisfaction	1	--			
productivity	0	1	--		
employees	0	0	1	--	
empowered	0	0	1	1	--

linear aggregate

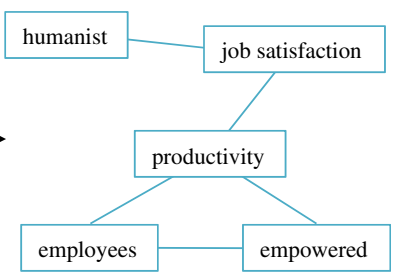


Fig. 3 The linear aggregate link array for the example text from Fig. 2 and its force-directed graph

when key terms are invisible to the software as anaphoric references (i.e., pronouns). Thus the link arrays for the linear and the sentence aggregate approaches should be quite different for good and poor essays and for essays with many pronouns.

Although the sentence aggregate approach may under- or over-specify keyword pair-wise associations relative to the linear approach, this may not be an issue for the purposes of generating group knowledge representations because the Pathfinder network scaling approach (Cooke 1992; Johnson et al. 1994) is a data reduction approach that emphasizes the main pair-wise associations in proximity data that is "a fuller representation of the salient semantic structures than minimal spanning trees, but also a more accurate representation of local structures than multidimensional scaling techniques" (Chen 1999, p. 408). Knowledge Network Organizing Tool software (KNOT 1998) for Pathfinder network analysis can be used to average together multiple link array proximity data files in order to establish a single group average data file of those essays and its PFNET representation. When multiple proximity arrays are averaged into one array, associations in arrays that are idiosyncratic, spurious, or just plain errors (both human and software errors) occur less frequently than do the frequencies of apposite associations, and these low frequency associations typically drop out of the averaged group array (although common misconceptions will tend to be included in the average group array). Thus both under- and over-specification errors are less of a problem for *ALA-Reader* analysis when the intent is to obtain averaged group representations; although it is still potentially a problem for representing an individual student's knowledge structure.

When using *ALA-Reader* and KNOT for generating an individual student's essay score relative to an expert referent, of the two common types of PFNET comparison measures (see KNOT release notes for a full description of the measures), network links in *common* measure have been shown to be a better predictor of rater essay scores than has the network *similarity* measure (Taricani & Clariana 2006). This may be because errors and spurious associations count in the similarity measure but do not count in the common measure. Specifically, only associations that match the expert are counted towards the common measure and so spurious and incorrect associations are disregarded. Thus over-specification of a student's PFNET by *ALA-Reader* is less problematic when the common measure is used for essay scoring and this may explain why the common measure has been shown to be superior to the similarity measure for ranking essays, although the similarity measure is generally better than the common measure for other purposes.

Purpose

This investigation considers the effects of pronouns on the quality of *ALA-Reader* measures of knowledge structure. Pronouns in text present a substantive problem for text processing software because pronouns carry meaning in the text but often must be disregarded because most software is not able to connect the pronouns with their referents. This investigation uses human readers to edit the most common pronouns (e.g., their, it, and they) in students' essays to the appropriate referent. By comparing the PFNETs obtained from the original essays to those of the edited essays, it is possible to consider the relative effects of pronouns on the quality of text processing. As noted above, the density of key terms per sentence has a strong effect on the resulting PFNETs, novice or poor essays are more likely to be under specified and thus negatively impacted even more by the presence of many pronouns; while good essays with a higher density of keywords may overcome the likely negative effects of pronouns. To consider this possibility that the presence of

pronouns influence the analysis of poor essays more than good essays, the first analysis in this investigation consists of comparisons of the average group representations of the 15 bottom performing students' original and edited essays to those of the 15 top performing students. As a measure of consistency across time and groups, the average group representations of the top and bottom performing students in this investigation are compared to the top and bottom performing students from the previous study (Clariana & Wallace 2007). Finally, individual students' original and edited essays relative to an expert's essay are analyzed by *ALA-Reader* with KNOT analysis, and those common measures are compared by correlation to human rater composite essay scores to consider the effects of pronouns on the *ALA-Reader* individual essay scoring method.

Method

Participants were undergraduate students enrolled in two sections of a required business course in an Eastern university in the USA. This is the same course and the same instructor who was involved in the earlier study (Clariana & Wallace 2007); this study was held a year after that study. There were 49 total students enrolled, but data from four students could not be used leaving a final sample of 45 participants. As before, the students completed an essay as part of the course final examination. The essay prompt stated,

> Describe and contrast in an essay of 300 words or less the following management theories: Classical/Scientific Management, Humanistic/Human Resources, Contingency, and Total Quality Management. Please use the terms below in your essay: administrative principles, benchmarking, bureaucratic organizations, contingency, continuous improvement, customers, customer focus, efficiency, employee, empowerment, feelings, Hawthorne studies, human relations, humanistic, leadership, management (i.e., bosses), Management by Objectives, motivate, needs, organization (i.e., corporation), plan, product, quality, relationship, scientific management (classical), service, situation (or environment), TQM, and work (or job, task).

Note that in Clariana and Wallace (2007), essentially the same essay prompt was used except that a list of terms was not provided. In this present investigation, it was anticipated that providing a list of important terms in the writing prompt would influence the students to include these terms in their essays and this would improve the consistency of the *ALA-Reader* analysis. These 29 terms including their synonyms and metonyms were the key terms used by the *ALA-Reader* software during text analysis, 21 of the terms were the same as those used for analysis in the previous investigation and eight terms are new for this analysis.

Word frequency counts were determined using the free online program *Textalyser* (see textalyser.net). The 45 student essays contained 13,464 total words (1,844 unique words that include proper nouns) which is an average of 299 words per essay (range from 170 to 476 words, standard deviation of 70.1). The ten most common words account for 27% of all of the text (3,651 occurrences) and include in order: (1) *the*, 852 occurrences in 45 essays or 18.9 average per essay, (2) *and*, 460 occurrences, 10.2 per essay, (3) *of*, 446 occurrences, 9.9 per essay, (4) *to*, 431 occurrences, 9.6 per essay, (5) *management*, 326 occurrences, 7.2 per essay, (6) *a*, 265 occurrences, 5.9 per essay, (7) *is*, 244 occurrences, 5.4 per essay, (8) *in*, 228 occurrences, 5.1 per essay, (9) *that*, 212 occurrences, 4.7 per essay, and (10) *employees*, 187 occurrences, 4.2 per essay. The three most common pronouns are: *their* (ranked 15th with 2.7 per essay), *they* (ranked 16th with 2.6 per essay), and

it (ranked 23rd with 2.2 per essay). Students use of the terms *their* and *they* in their essays indicates that they were assuming the perspective of either a manager or of an employee, and only a few wrote as an independent observer.

Comparing group average data representations

ALA-Reader software was used to process the original unedited essays (with pronouns present) and the edited essays (i.e., the pronouns *their*, *it*, and *they* were manually replaced with the appropriate referent) using both a linear aggregate approach and a sentence aggregate approach. To identify the best and worst essays for grouping purposes, human rater essay scores for these 45 essays were used to rank the student essays; then Pathfinder KNOT software was used to average together the proximity files of the 15 top performing and the 15 bottom performing students for each of the four data sets to create four top performing groupings and four bottom performing groupings consisting of linear aggregate of original essays, linear aggregate of edited essays, sentence aggregate of original essays, and sentence aggregate of edited essays. The group average proximity raw data Pearson correlations between these eight groupings are shown above the diagonal in Table 1. In addition, the PFNET similarity values of these eight groupings, calculated as the PFNET intersection divided by PFNET union, are shown below the diagonal in Table 1.

There are strong correlations ($r > 0.90$) between the proximity raw data within the top group and within the bottom group (see the values above the diagonal in Table 1), but not between the top and bottom groups (r range from 0.61 to 0.79). This indicates that editing the pronouns to their referents had little effect on the top or the bottom groups' raw proximity data (see the four underlined values above the diagonal, all $r > 0.97$). The network similarity values comparing the original to the edited essays were also large (see values below the diagonal in Table 1), but the linear aggregate approach obtained slightly larger similarity values (top group = 0.81; bottom group = 0.83) compared to the sentence aggregate approach (top group = 0.78; bottom group = 0.71). This suggests that pronouns influence sentence aggregate PFNETs more. Even so, these are quite similar PFNETS with many links in common. For example, the top group's PFNET based on their original essays analyzed using the linear approach contains 31 links (see group 'a' in Table 1 and the left panel of Fig. 4) and the PFNET based on their edited essays contains

Table 1 Top and bottom performing groups' correlations (above) and similarities (below)

	a	b	c	d	e	f	g	h
Top group ($n = 15$)								
a. Linear, original	–	0.99	0.98	0.99	0.70	0.79	0.70	0.69
b. Linear, edited	0.81	–	0.97	0.98	0.66	0.75	0.67	0.73
c. Sentence, original	0.56	0.52	–	0.99	0.61	0.73	0.66	0.67
d. Sentence, edited	0.49	0.46	0.78	–	0.64	0.74	0.71	0.74
Bottom group ($n = 15$)								
e. Linear, original	0.41	0.42	0.43	0.40	–	0.98	0.92	0.90
f. Linear, edited	0.37	0.38	0.36	0.33	0.83	–	0.95	0.91
g. Sentence, original	0.30	0.29	0.35	0.32	0.47	0.43	–	0.99
h. Sentence, edited	0.29	0.35	0.33	0.31	0.45	0.40	0.71	–

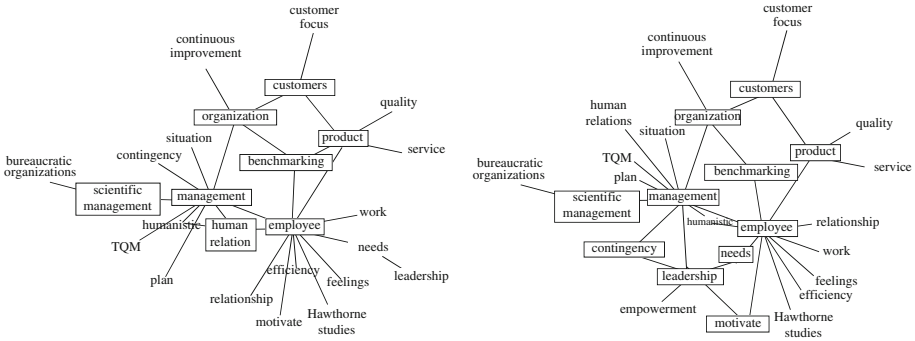


Fig. 4 PFNET group representations of the original (*left*) and the edited essays (*right*)

34 links (see group 'b' in Table 1 and the right panel of Fig. 4). These two PFNETs share 29 links in common which is an 89% overlap (i.e., 29/(31 + 34/2). The graphical representations of these two PFNETs as force-directed graphs are also visually quite similar (see Fig. 4).

It is useful to compare the PFNET visual depictions of the averaged essays of the top performing students. The two terms *management* and *employee* are exceptionally well connected and so are the central terms in both representations. Other well connected terms in both representations include *organization*, *benchmarking*, *customer*, and *product*. Three of the four super-ordinate essay prompt categories stated in the essay prompt, *scientific management*, *contingency*, and *TQM*, were all associated with *management* while the fourth category, *humanistic*, was associated with both *management* and *employee*. Emotion-related terms such as *feelings*, *needs*, *relationship*, and *motivation* were associated with *employee* in both PFNET representations as were the action words *benchmarking*, *work*, and *product*.

However, the key term *leadership* is relatively more connected in the PFNET of the edited essays (compare the left and right panels of Fig. 4) and so it assumes a more central position in that representation. In contrast, in the PFNET of the original unedited essays, *leadership* is not a central term and is only connected to the term *needs*. Be sure to note that the differences in structure between these two PFNETs relate to how students used pronouns in their essays, probably the pronoun *it* to refer to *leadership*, and not necessarily to their views of the centrality of leadership in these theories.

How related are these essays to those of the previous study? Using the 29 key terms from this investigation, the original (unedited) essays of the top ten and bottom ten students from the previous investigation and the top ten and bottom ten from this present investigation were processed by *ALA-Reader* using the linear aggregate approach to obtain four group average PFNET representations. These were then compared to each other and to the expert essay (see Table 2). The top and bottom groups in the present investigation were most alike (54% overlap) and the top and bottom groups from 2007 were also quite alike (48% overlap). The two top groups (49% overlap) were more similar to each other than the two bottom groups (41% overlap). Comparisons to the expert essay show that the two top groups, now and in 2007, were considerably more similar to the expert than the two bottom groups.

Table 2 Percent overlap of the top and bottom groups for the present investigation and for Clariana and Wallace (2007)

Similarity	A	B	C	D
A. Top group (now)	1			
B. Bottom group (now)	54%	1		
C. Top (2007)	49%	38%	1	
D. Bottom (2007)	42%	41%	48%	1
% Overlap with the expert	43%	29%	45%	36%

Comparing individual data representations

ALA-Reader software was used to process both the original and the edited essays using both a linear and a sentence aggregate approach to produce four PFNETS for each participant. Then the sentence aggregate PFNETs were compared to a sentence aggregate expert referent PFNET and the linear aggregate PFNETs were compared to a linear aggregate expert referent PFNET to obtain four scores for each original essay that consisted of the number of links in common.

For comparison proposes, two separate composite essay benchmark scores were determined based on (1) the three human essay scores (specifically, the factor score derived by the SPSS version 15.0 factor analysis regression option) and (2) the three human essay scores plus the *ALA-Reader* score (also using SPSS factor score). Pearson correlations were conducted between these two benchmark composite scores and the four sets of *ALA-Reader* essay common scores. The linear aggregate method obtained better correlations with the human raters than did the sentence aggregate approach (see Table 3). Further, replacing pronouns with their referents had little effect on linear aggregate scores but had a small positive effect on sentence aggregate scores. Overall this indicates that this sentence-level analysis approach was influenced more by the presence of pronouns relative to this linear approach. Also, adding key words to the essay writing prompt did not appear to improve the quality of the output relative to the earlier investigation, although perhaps the wrong key words were used in the essay prompt.

Conclusion

The *ALA-Reader* sentence aggregate approach obtained somewhat different PFNET representations for the original essays relative to the edited essays, while there was little

Table 3 Pearson correlations between the *ALA-Reader* essay scores and the two composite scores

Analysis approach	Composite score	
	3 Raters only	3 Raters and *ALA-reader*
Original essays		
Linear aggregate	0.59	0.74
Sentence aggregate	0.34	0.44
Edited essays		
Linear aggregate	0.57	0.71
Sentence aggregate	0.41	0.51

All significant at the $P < 0.05$

difference between the PFNETs obtained for edited and unedited essays using the linear aggregate approach. Based on the analyses of individual essay scores, pronouns in text passages had a negative effect on sentence aggregate PFNET representations. This indicates that the linear approach is less affected by pronouns and is generally superior to the sentence approach for the narrow purposes of average group knowledge representation and is also adequate for scoring individual essay content.

Students' essays are grammatically imperfect and sometimes incoherent. Our experience from editing the pronouns in these essays is that raters do not always agree on a pronoun's referent. Pronouns in text can be a substantial issue for text analysis software and subroutines for handing pronouns are expensive to develop and generally are not perfectly accurate and so such subroutines may actually add more error to the data than they correct. Although this is a small sample of essays, these results suggest that the *ALA-Reader* linear approach would not benefit from a pronoun handling subroutine. Thus the development cost of a pronoun handler is not warranted.

There are several avenues for further development of *ALA-Reader*. The greatest need is concurrent and divergent validity studies. *ALA-Reader* is not necessarily an essay scoring tool, but rather it is probably a tool to measure knowledge structure which indirectly relates to essay scores. But when used as an essay scoring method, *ALA-Reader* is likely to be more appropriate for some types of essays than others and is probably inappropriate for many types of essays. The more technical or specific the vocabulary in the essays, the better *ALA-Reader* should perform. So another area of further study is refining the type of essay genre and the specific writing prompt, for example should key words be included in the prompt, to improve the quality of the tool.

But also, the *ALA-Reader* approach can be further developed and extended. For example, the distance between terms in concept maps has been shown to be important information related to inference and comprehension (Cernusca 2007; Poindexter & Clariana 2006; Taricani & Clariana 2003, 2006); similarly the distances between key terms in a text passage may also be important information. A feature will be added to *ALA-Reader* to capture these distances between terms as a proximity array in order to consider this notion.

Which key terms to use during analysis and how many should be used is another area for further investigation because some key terms appear to be far more important than others. In this investigation, the 29 key terms selected by the course instructor to be included in the essay prompt were also used as key terms for the analysis. Further research should try different sets of key terms as well as different numbers of terms in the analysis stage and also what is the influence of synonyms and metonyms on the quality of the analysis.

As a side note, over-specification is potentially a big problem for the expert essay used as the referent to score the students' essays. To handle this, it is critical to meticulously design the expert referent as an ideal PFNET rather than just convert an expert essay into a PFNET so as to avoid the effects of unintended spurious and error associations in the expert referent essay data set. Thus another area of *ALA-Reader* development is how to establish the expert referent used to score essays.

In summary, because these findings are reasonable, more research is warranted to further develop and validate this *ALA-Reader* approach for establishing group network representations and for comparison of individual's content knowledge structure.

References

Cernusca, D. (2007). *A design-based research approach to the implementation and examination of a cognitive flexibility hypertext*. Unpublished doctoral dissertation, University of Missouri, Columbia. Retrieved November 11, 2008, from http://edt.missouri.edu/Summer2007/Dissertation/CernuscaD-071907-D8036/research.pdf.

Chen, C. (1999). Visualizing semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35*, 401–420.

Clariana, R. B., & Koul, R. (2004). A computer-based approach for translating text into concept map-like representations. In A. J. Canas, J. D. Novak, & F. M. Gonzales (Eds.), *Proceedings of the first international conference on concept mapping: Vol. 2. Concept maps: Theory, methodology, technology* (pp. 131–134). Pensacola, FL: Institute for Human and Machine Cognition. Retrieved November 11, 2008, from http://cmc.ihmc.us/papers/cmc2004-045.Pdf.

Clariana, R. B., & Wallace, P. E. (2007). A computer-based approach for deriving and measuring individual and team knowledge structure from essay questions. *Journal of Educational Computing Research, 37*, 209–225.

Cooke, N. M. (1992). Predicting judgment time from measures of psychological proximity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 640–653.

Craik, K. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.

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.

Harper, M. E., Hoeft, R. M., Evans, A. W. III, & Jentsch, F. G. (2004). Scoring concepts maps: Can a practical method of scoring concept maps be used to assess trainee's knowledge structures? *Human Factors and Ergonomics Society Annual Meeting Proceedings, 48*, 2599–2603.

Johnson, P. J., Goldsmith, T. E., & Teague, K. W. (1994). Locus of the predictive advantage in Pathfinder-based representations of classroom knowledge. *Journal of Educational Psychology, 86*, 617–626.

Johnson, T. E., O'Connor, D. L., Pirnay-Dummer, P. N., Ifenthaler, D., Spector, J. M., & Seel, N. (2006). Comparative study of mental model research methods: relationships among ACSMM, SMD, MITOCAR and DEEP methodologies. In A. J. Canas & J. D. Novak (Eds.), *Proceedings of the second international conference on concept mapping: Vol. 1. Concept maps: Theory, methodology, technology* (pp. 177–184). Pensacola, FL: Institute for Human and Machine Cognition. Retrieved November 11, 2008, from http://cmc.ihmc.us/cmc2006Papers/cmc2006-p177.pdf.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (1998). *Mental models: A gentle guide for outsiders*. Retrieved November 29, 2008, from http://www.si.umich.edu/ICOS/gentleintro.html.

KNOT. (1998). *Knowledge network organizing tool*. Retrieved June 11, 2004, from http://interlinkinc.net/.

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*, 261–273.

Poindexter, M. T., & Clariana, R. B. (2006). The influence of relational and proposition-specific processing on structural knowledge and traditional learning outcomes. *International Journal of Instructional Media, 33*, 177–184.

Seel, N. M. (1999). Educational diagnosis of mental models: Assessment problems and technology-based solutions. *Journal of Structural Learning and Intelligent Systems, 14*, 153–185.

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

Taricani, E. M., & Clariana, R. B. (2003). *Semantic map automated assessment techniques. Paper presented at the meeting of Association for Educational Communications and Technology (AECT)*, Anaheim, CA, October, 2003.

Taricani, E. M., & Clariana, R. B. (2006). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development, 54*, 61–78.

Roy B. Clariana is an Associate Professor and the Education Division Head in the School of Graduate Professional Studies at the Pennsylvania State University. He designed and developed the ALA-Reader software used in this investigation and made it available to the research community in 2004. Since then, he has conducted and published several investigations to validate and improve the usefulness of the tool as a measure of knowledge structure and as a form of essay assessment. Download this software from http://www.personal.psu.edu/rbc4/score.htm.

Patricia E. Wallace is a Full Professor in the School of Business, an AACSB accredited business school, at The College of New Jersey, in Ewing, New Jersey. Dr. Wallace has authored numerous publications in highly-competitive academic journals; in addition, she has published refereed conference proceedings and made presentations at both international and national conferences. Dr. Wallace's research focus is on the impact of Information Technology on teaching, learning, and assessment in Information Systems. Currently, she is researching trends in globalization and teamwork in Business Education while on sabbatical.

Veronica M. Godshalk is an Associate Professor of Management in the Department of Business Administration at the University of South Carolina, Beaufort. She is also the Business Administration Department Chair, and has been at USCB since August 2007. Dr. Godshalk teaches courses in organizational behavior, management and leadership, business research methods, and career management. She previously taught at the Pennsylvania State University for fourteen years, where she received the 2000 Arthur L. Glenn Award for Faculty Teaching Innovation and received tenure and Associate rank. Dr. Godshalk has also taught at Drexel University. Dr. Godshalk's research interests include issues surrounding career management and mentoring. She has published extensively with forty articles, books, book chapters and refereed conference proceedings in print. In 2000, she published a book, Career Management, with co-authors Jeff Greenhaus and Gerry Callanan and is currently working on a 4th edition. She is an active member and presenter in professional associations, such as the Academy of Management and the Society for Industrial and Organizational Psychology. Dr. Godshalk had worked in the computer industry in sales and sales management prior to entering academia, and has been a consultant for several Fortune 500 companies. She earned her Ph.D. from Drexel University, and her M.S. from the University of Pennsylvania.