J. EDUCATIONAL COMPUTING RESEARCH, Vol. 32(3) 227-239, 2005

COMPARING SEVERAL HUMAN AND COMPUTER-BASED METHODS FOR SCORING CONCEPT MAPS AND ESSAYS

RAVINDER KOUL
ROY B. CLARIANA
ROYA SALEHI
College of Education at Penn State University

ABSTRACT

This article reports the results of an investigation of the convergent criterion-related validity of two computer-based tools for scoring concept maps and essays as part of the ongoing formative evaluation of these tools. In pairs, participants researched a science topic online and created a concept map of the topic. Later, participants individually 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. Computer-based concept map scores were a very good measure of the qualitative aspects of the concept maps ($r = 0.84$) and were an adequate measure of the quantitative aspects ($r = 0.65$). Also, the computer-based essay scores were an adequate measure of essay content ($r = 0.71$). If computer-based approaches for scoring concept maps and essays can provide a valid, low-cost, easy to use, and easy to interpret measure of students' content knowledge, then these approaches will likely gain rapid acceptance by teachers at all levels.

There is an extensive and growing literature on the use of concept maps as a *complementary* form of classroom assessment (Markham, Mintzes, & Jones, 1994; Ruiz-Primo, 2000; Ruiz-Primo, Schultz, Li, & Shavelson, 2000; Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Shavelson, Lang, & Lewin, 1994). A recent report by the National Research Council (2001) stresses the importance of using classroom assessments that evaluates cognitive structure. Concept maps

are sketches or diagrams that show the relationships among a set of content terms by the positions of the terms and by labeled lines and arrows connecting some of the terms (National Research Council, 2001, p. 265). Concept maps and other similar graphical representations are one way, and maybe the best way, to measure students' "structural knowledge" of content (Jonassen, Beissner, & Yacci, 1993).

But this form of assessment requires consideration of two important issues, how the map is constructed and how it is scored (Ruiz-Primo et al., 2000; Stoddart, Abrams, Gasper, & Canaday, 2000; Turns, Atman, & Adams, 2000). The interpretation and scoring of concept maps involves judgments along several dimensions that represent the breadth, depth, and connectedness of the knowledge all based on only the terms, links, cross links, and levels of hierarchy in the concept map. In order to validate a particular method for scoring concept maps, there should be a strong relationship between the concept map score and the student's actual understanding of that content. Thus, the scoring approach used is especially important, since different approaches will obtain different scores for the same set of maps.

Novak and Gowin (1984) first proposed a method for scoring maps that relies solely on manually counting the number of *valid* components in a map. Their approach considers propositions, levels of hierarchy, examples, and cross links. Each component must be judged against a universe of possible valid responses, and so the rater requires extensive expert knowledge in that specific content. After all valid components have been collected from a map, the count of each component is multiplied by a weighting factor for that category, and then the weighted counts are added together to establish the final score. Lomask's method (Lomask, Baron, Greig, & Harrison, 1992) for scoring concept maps relies on the match between terms and propositions in a student's map compared to an expert's map. Here, each component is judged against the finite set of terms and propositions contained in the expert's map. Both of these methods stress the quality or importance of each individual component in a map.

In contrast, a recent scoring system proposed by Kinchin and Hay (2000) assigns a score based on the overall organization (or structure) of the map. Their approach views knowledge structure as a more holistic variable, rather than as a "sum of the individual components" type variable. Their views arose from their extensive experiences with concept maps and with the actual understood nature of biology content (Safayeni, Derbentseva, & Cañas, 2003, p. 12). Their method depends on raters identifying progressive levels of understanding by judging the overall structure of the concept map, for example whether it is linear, spoke, or net.

These different concept map scoring methods affect, in a regular way, the correlations observed between the concept map scores compared to scores obtained from other criterion test formats such as multiple-choice and essay tests. For example, concept map scoring methods that stress the importance of

key content or propositions have been found to be highly correlated with traditional multiple choice tests, apparently both tap "declarative aspects" of knowledge such as the number of key concepts, ideas, or principles (Stoddart et al., 2000). Apparently, a concept map can contain several dimensions of information, but so far, no single method of scoring concept maps captures every aspect of domain knowledge contained in a concept map.

CONCEPT MAPS AND ESSAYS

Essays are a well researched, important, versatile, somewhat reliable, and authentic but expensive form of assessment (Nitko, 1996). Essays and concept maps are considered to be highly related and complementary forms of assessment (Goldsmith & Johnson, 1990; Goldsmith, Johnson, & Acton, 1991; Gonzalvo, Cañas, & Bajo, 1994), especially if the concept map is used as a scaffold or outline for writing the essay and vice versa. Based on this expectation, one of the first large-scale uses of concept maps in assessment, the Connecticut statewide assessment, involved converting student essays into concept maps by human raters, who then hand scored the concept maps as a measure of science content knowledge and comprehension (Lomask et al., 1992).

Recently, Clariana, Koul, and Salehi (2005) examined the criterion-related validity of a computer-based software called *ALA-Mapper* (Clariana, 2003) for scoring concept maps compared to essays scored by human raters. Working in pairs, graduate students in an evaluation course were asked to construct a paper-based concept map while conducting online research on the structure and function of the heart and circulatory system. The students then used their concept maps to write a 250-word summary essay on this topic. Prototype *ALA-Mapper* software was used to compare the propositions (lines drawn between terms) and associations (distances between terms) in the students' maps to an expert's map. The *ALA-Mapper* scores were significantly correlated to human-rater essay scores (maximum $r = 0.83$). The authors concluded that relative to teachers scoring these summary essays, these computer-based concept map scores provided a relatively low-cost, easy to use, and easy to interpret measure of students' science content knowledge.

In a follow-up study (Taricani & Clariana, in press), sixty undergraduate students read a print-based instructional text on the heart and circulatory system and then created concept maps of that content. Half were given feedback in the form of a prepared concept map and the other half, no feedback. Then all completed a multiple-choice posttest with 20 terminology and 20 comprehension questions. The concept maps were scored using *ALA-Mapper* and these concept map scores were compared to the terminology and comprehension posttest scores. Concept map scores derived from proposition data were more related to terminology whereas concept map scores derived from association data were more related to comprehension. Their findings suggest that a set of concept maps

contain *quantitative* information related to valid propositions as well as *qualitative* information related to knowledge structure.

Unlike essays, concept maps can provide a visual and holistic way to describe declarative knowledge relationships, often providing a clear visual indication of student understanding that can highlight student misconceptions. A software utility called *ALA-Reader* (Clariana, 2004) can translate text summaries into a concept map-like representation of the text that can then be scored using the *ALA-Mapper* software approach. Using this software, Clariana and Koul (2004) considered the criterion-related validity of visual representations of student essays. *ALA-Reader* essay scores were compared to essay scores from 11 pairs of human raters applying a generic content rubric. The correlation between human-rater and computer-based essay scores was $r = 0.69$, and the computer scores ranked 5th out of 12. The concept-map like representations of the essays provided the students (and their instructor) with another way of thinking about their written text, especially by highlighting correct, incorrect, and missing propositions in their text.

This current investigation reports the results of ongoing investigations into the convergent criterion-related validity of these computer-based approaches for scoring concept maps and essays. Computer-based concept map scores are compared to concept map scores from human-raters who used both *qualitative* and *quantitative* rubrics. In addition, two different computer-based approaches for scoring essay content are examined (Clariana & Koul, 2004; Landauer, Foltz, & Laham, 1998). By comparing these seven approaches together (four for scoring concept maps and three for scoring essays), we will take the next small step in understanding what is measured (and perhaps, what can be measured) by these computer-based approaches. This formative information will contribute to the next round of development of the software tools.

METHOD

Participants

The participants in this investigation (a sample of convenience) were practicing teachers ($N = 22$) enrolled in two graduate courses on our campus taught by the same instructor. The participants ranged in age from 30 to 38 years, with 15 females and 7 males. Participants were briefed on the purpose of the investigation and were asked to participate, and all agreed.

Procedure

Working in pairs in the computer lab, first the participants conducted online research on the structure and function of the human heart and circulatory system. Then, using *Inspiration*™ software, each pair together constructed a concept map

to represent the structure and function of the human heart and circulatory system. *Inspiration* software allows the user to easily place and move terms closer together and father apart on the screen in order to create the most accurate representation. The students completed their maps and printed it for later use. Later outside of class, each pair wrote a 250-word essay based on their concept map. The essay served as a precursor for the in-class activities of scoring concept maps and essays conducted in during the following class meetings.

Posttest Measures

There are seven dependent measures, four concept map scores and three essay scores. Specifically, the concept maps were scored by pairs of raters using a quantitative and then a qualitative rubric. The concept maps were also scored using a computer-based method called *ALA-Mapper* proposed by Clariana, Koul, and Salehi (2005). The written essays were scored by the same pairs of human raters using a generic content rubric and then also by a computer-based method called *ALA-Reader* (Clariana, 2004) as well as a computer-based method called *Latent Semantic Analysis* (*LSA*) (Landauer, Foltz, & Laham, 1998).

Concept Map Scores and Rubrics

The quantitative rubric was adapted from the Lomask et al. (1992) rubric. Eleven pairs of human raters scored every concept map. The rubric considered size (the count of terms in a student map expressed as a proportion of the terms in an expert concept map) and strength (the count of links in a student map as a proportion of necessary, accurate connections with respect to those in an expert map). Each pair of raters used their own understanding to reach consensus on their idea of an expert representation of content and to prepare their own "expert" map. Since most concept maps are scored by real teachers, this improves the generalizability of these scores somewhat, but at the expense of validity and reliability. Cronbach alpha reliability for the human-rated concept map scores derived using the quantitative rubric was 0.86.

The qualitative rubric for scoring concept maps was based on research by Kinchin and Hay (2000). This rubric deals with three common map structure which may be interpreted as indicators of progressive levels of understanding: 1) *Spoke*, a structure in which all the related aspects of the topic are linked directly to the core concept, but are not directly linked to each other; 2) *Chain*, a linear sequence of understanding in which each concept is only linked to those immediately above and below; and 3) *Net*, a network both highly integrated and hierarchical, demonstrating a deep understanding of the topic. Following the guidelines provided by Kinchin and Hay (2000), each of the eleven pairs of human raters used the qualitative rubric to score, on a scale of 1 to 5, every concept map from the class. Cronbach alpha reliability for the human-rated

concept map scores derived using the qualitative rubric was 0.53, substantially lower than that of the quantitative rubric.

A computer-based technique called *ALA-Mapper* was also used to score the concept maps. *ALA-Mapper* is used to measure the number of valid links (propositions) and the physical proximity (associations) of terms in a student's concept map. Scores are derived by comparing the student's proposition and association data to an expert's (see Clariana, Koul, & Salehi, 2005 for complete details). To conform to the 1-5 scale range used by raters, raw proposition and association scores were linear transformed to the 1-5 scale by adding one to the raw data (to remove any zeroes), then dividing by the maximum score (converts all scores to a proportion) and finally multiplying by five (Clariana et al., 2005).

Essay Scores and Rubric

Eleven pairs of human raters scored every written essay. The essay scoring rubric considered (a) content, whether the science content is clear, relevant, accurate, and concise, (b) style, whether the summary is a fluent and succinct piece of writing and compositionally clear, functional and effective, (c) mechanics, including technical or procedural details, and the practicalities, use of grammar, punctuation and spelling, and (d) overall score from a holistic view. This investigation focused on "content"; however the other assessment dimensions were included in the rubric in order to improve the content measure. Specifically, since raters had the option of scoring style and mechanics separately, then their content score will more accurately reflect the essay's content (Nitko, 1996). Cronbach alpha reliability for the human-rated essay content scores was 0.85.

Essays Scored by Latent Semantic Analysis

The essays were also scored automatically using *Latent Semantic Analysis* (*LSA*), a computer-based essay scoring system available on the Internet (Landauer, Foltz, & Laham, 1998). *LSA* determines the similarity between text portions, its internal reliability is 1.00 (Landauer, Laham, Rehder, & Schreiner, 1997; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002).

Essays Scored by *ALA-Reader*

Clariana's (2004) computer-based approach called *ALA-Reader* was also used to score the essays. Based on key terms identified by an expert, *ALA-Reader* aggregates the terms in the student's essay into propositional data based on the proximity of the terms within the student's essay (Clariana & Koul, 2004). In effect, *ALA-Reader* converts the student's essay into a semantic map, then the software compares the student's proposition data to the proposition data derived from an expert essay. One point is awarded for each proposition that matches the expert. To conform to the 1-5 scale range used by raters, as with the *ALA-Mapper*

raw scores above, the *ALA-Reader* proposition scores were linear transformed by adding one to the raw data then dividing by the maximum score and finally multiplying by five.

RESULTS

Correlations between the various scoring approaches are presented in Table 1. The variables in this investigation include the following: concept maps scored by raters using a quantitative rubric (quantitative, $M = 3.9$, $SD = 0.69$), concept maps scored by raters using a qualitative rubric (qualitative, $M = 3.2$, $SD = 0.41$), concept maps scored by *ALA-Mapper* software based on propositions (*ALA-Mapper* propositions, raw score $M = 3.9$, $SD = 3.3$, scaled score $M = 2.2$, $SD = 1.5$), concept maps scored by *ALA-Mapper* software based on associations (*ALA-Mapper* associations, raw score $M = 1.7$, $SD = 2.3$, scaled score $M = 1.3$, $SD = 1.2$), essays scored by human raters (essay $M = 3.9$, $SD = 0.56$), essays scored by computer-based *Latent Semantic Analysis* (*LSA* $M = 2.9$, $SD = 0.8$), and essays scored by *ALA-Reader* software (*ALA-Reader*, raw score $M = 7.4$, $SD = 5.7$, scaled score $M = 1.9$, $SD = 1.3$). The correlations between concept maps scored by human raters and concept maps scored by the computer-based approaches were high ($r = 0.50$ to $r = 0.84$). Correlations between essays scored by *ALA-Reader* software and maps scored by human or computer-based approaches were moderately high ($r = 0.38$ to $r = 0.71$). Correlations between essays scored by *Latent Semantic Analysis* software and maps scored by human or computer-based approaches were low ($r = –0.07$ to $r = 0.39$).

Table 1. Correlations of All Essay and Concept Map Scores

	A	B	C	D	E	F	G
Concept Maps							
A. Lomask et al. Quantitative rubric (raters)	(0.86)						
B. Kinchin & Hay Qualitative rubric (raters)	0.69	(0.53)					
C. *ALA-Mapper* proposition (computer)	0.65	0.84	1				
D. *ALA-Mapper* association (computer)	0.50	0.53	0.71	1			
Essays							
E. Essay (raters)	0.34	0.61	0.54	0.32	(0.85)		
F. *LSA* (computer)	–0.07	0.39	0.23	0.06	0.62	1	
G. *ALA-Reader* (computer)	0.38	0.57	0.46	0.46	0.71	(0.57)	1

Note: Cronback alpha is shown in parentheses.

DISCUSSION

In this comparison of several methods of scoring concept maps and essays, both human and computer-based methods of scoring maps and essays capture some of the same information about students' science content and process knowledge. However, *within* computer-based methods of scoring essays and maps there were differences between what the different software captured. The differences in scores among the computer-based methods used in this investigation stem from the unique designs of different software.

The *Latent Semantic Analysis* (*LSA*) software compares the interrelationship of words in a particular essay with the interrelationship of words in the essays used to train the software. *LSA* measures the content and structure of the student essay to compare it to the training essays and assigns a holistic score (on a scale of 1 to 5) by placing the essay in a category with the most similar training essays. *LSA* relies on statistical analysis of a large amount of text (typically thousands to millions of words) to derive a high-dimensional semantic space that permits comparisons of the semantic similarity of words and passages. Words and passages are represented as vectors and their similarity is measured by the cosine of their contained angle in semantic space. Essay content scores are determined by comparing the vector of the essay to a finite set of scored sample essays in the dataset, the essay receives the score of its closest sample essay.

ALA-Reader software requires the user to prepare a list of important terms and their synonyms and metonyms (maximum 30 terms) which the software then uses to look for co-occurrences in each sentence of the student's essay. *ALA-Reader* then converts the co-occurrences of terms into propositions, which are then aggregated across all sentences into a proximity array. *ALA-Reader* employs *PCKNOT* software (Schvaneveldt, 1990) to convert raw proximity data into *PFNet* representations. Students' *PFNet* representations can then be compared to the *PFNet* derived from an essay written by a content expert. The student's proposition-agreement-with-the-expert scores are converted to a 1–5-point scale. *ALA-Reader* software condenses essays into concept map like representations for proposition analysis. Since *ALA-Reader* makes visual students' written essays, essay scores on *ALA-Reader* are more closely aligned with the concept map scores (see Table 1).

Both human and computer-based quantitative methods of scoring maps rely on the number of valid" links presented in the map along with the proportion of necessary, "accurate" connections compared to the expert map. However, qualitative rubrics do not rely solely on "valid" propositions and associations in a student's map. Therefore, the scores given by raters while using quantitative rubric were less varied ($r = 0.44$ to $r = 0.80$) than while using qualitative rubric ($r = 0.09$ to $r = 0.84$) (see Tables 2 and 3). The difference stems from the procedural and philosophical differences in the nature of quantitative (analytic) and qualitative (holistic) rubrics (Kinchin, 2001; Kinchin & Hay, 2000).

Table 2. Correlation of Each Qualitative and Quantitative Concept Map Scoring Source to the Combined Concept Map Score

Quantitative (Lomask et al.)			Qualitative (Kinchin & Hay)		
Scoring Source	*r*	Rank	Scoring Source	*r*	Rank
Rater pair 10	0.80	1	*ALA-Mapper* proposition	0.84	1
Rater pair 4	0.69	2	Rater pair 8	0.76	2
Rater pair 8	0.69	3	Rater pair 10	0.70	3
Rater pair 3	0.68	4	Rater pair 2	0.65	4
Rater pair 2	0.67	5	Rater pair 7	0.61	5
Rater pair 1	0.67	6	*ALA-Mapper* association	0.53	6
Rater pair 7	0.67	7	Rater pair 9	0.41	7
ALA-Mapper proposition	0.65	8	Rater pair 4	0.38	8
Rater pair 5	0.62	9	Rater pair 5	0.29	9
Rater pair 6	0.61	10	Rater pair 1	0.25	10
Rater pair 11	0.56	11	Rater pair 6	0.01	11
ALA-Mapper association	0.50	12	Rater pair 3	–0.09	12
Rater pair 9	0.44	13	Rater pair 11	*	*

*Missing data

Quantitative rubric looks for "correctness." Qualitative rubric looks for the overall mental representation of content as an indicator of the level of understanding.

In this study, *ALA-Mapper* concept map association scores and proposition scores were related (r = 0.71). Unlike the findings of Clariana et al. (2004), *ALA-Mapper* association scores were not significantly related to concept map or essay scores generated by other approaches. A major difference between the current and previous studies is the concept mapping format. In both studies, pairs worked together on one concept map, but in the earlier study, pairs used *PostIt* notes to construct their map on a large piece of newsprint while in the present study, pairs used *Inspiration* software to create their map. The activity of working with *PostIt* notes seems to invite much more spontaneous representation of content, use of language, and negotiation of meaning. Thus using *Inspiration* software likely substantially changed the collaborative process, for example, one person usually dominates the keyboard while the other sits idle. Also, unfamiliarity with the software adds an additional cognitive burden. Further, inspection of the concept maps indicate that the automatic alignment features in the *Inspiration* software tool encouraged users to alter the spatial layout of terms

Table 3. Correlation of Each Essay Scoring Source to the Combined Essay Score

Scoring Source	*r*	Rank
Rater pair 10	0.88	1
Rater pair 7	0.83	2
Rater pair 8	0.78	3
Rater pair 11	0.72	4
ALA-Reader	0.71	5
Rater pair 5	0.67	6
Rater pair 2	0.65	7
Rater pair 6	0.65	8
LSA (content)	0.62	9
Rater pair 4	0.58	10
Rater pair 3	0.42	11
Rater pair 1	0.25	12
Rater pair 9	0.08	13

to improve the "appearance" of the overall layout pattern, and this obviously substantially alters the distances (associations) between terms. The tool forces a hierarchical layout based on links rather than distances. Some students applied these spatial alignment features more than others, thus noise was added to some of the *ALA-Mapper* association scores by the software tool, though the *ALA-Mapper* proposition scores will be unaffected by this *Inspiration* software feature.

To extend, concept maps and essays are typically viewed as an individual student's knowledge representation. Here, students worked together in pairs to complete their concept map though they worked individually to complete the essay. This instructional strategy was selected based on a social constructivist view that using language creates, maintains, and reproduces meaning. By working in pairs on the research and especially to complete the concept map, it was anticipated that individuals would make their content ideas explicit and thus develop a deeper meaning of the content through this interaction. A fundamental principle of assessment validity is that there is actually something there to be measured. By using this pair strategy, we hoped to maximize what was there to be measured. However, the concept map software tool may have mitigated these benefits. Concept map software can certainly support collaboration, but

may require better designed scripts and scaffolds for it to succeed. Further research should observe the pair concept map *process in action*, perhaps contrasting a *PostIt* note approach with an *Inspiration* approach in order to describe the potentially effective aspects of concept mapping as a collaborative tool for creating shared meaning.

The issue of expertise is critical for both the human-rater and the computer-based scoring approaches used in this study. A particular limitation of this investigation is that raters were not given a single expert map to use as a referent during scoring. Rather, they depended on their own knowledge working in pairs to establish an expert referent. This approach will likely produce fairly different expert referents for each pair of judges, and possibly accounts for the low reliability of the qualitative concept map scores (alpha = 0.53) relative to the quantitative concept map scores (alpha = 0.86). Future research should either use only content experts as raters, and/or provide one expert referent map to all of the raters, or better, all judges should first negotiate together to reach consensus on a single expert referent before scoring the participant concept maps.

Sample size is another limitation of this investigation. Because of the small sample size, the results of this investigation should be used cautiously.

CONCLUSION

This investigation addressed a perceived need for scoring mechanisms which make the benefits of concept maps and essays more accessible to classroom teachers. Computer-based methods for scoring maps and essays are cost-effective, easy to use assessment tools that bring benefits to students' learning experiences while not placing unrealistic demands on the classroom teacher. For example, use of *ALA-Reader* software provide students and teachers with a visual way of thinking about written text and science content knowledge (especially by highlighting correct, incorrect, and missing propositions). The comparison of several methods of scoring maps and essays in this study suggests that the design of each tool (computer or human) captures certain aspects of students' science content knowledge. Variables such as the design of the computing software, "spontaneity" and "fluency" in the use of tools, skill level of the student and rater, and the type of scoring rubric (qualitative or quantitative) play a role in optimizing the benefits of learning with concept maps and essays.

REFERENCES

Clariana, R. B. (2003). *ALA-Mapper*. Retrieved May 3, 2003, from http://www.personal.psu.edu/rbc4

Clariana, R. B. (2004). *ALA-Reader* (beta version). Retrieved May 3, 2004, from http://www.personal.psu.edu/rbc4

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, and F. M. Gonzales (Eds.), *Concept maps: Theory, methodology, technology, vol. 2*, in the Proceedings of the First International Conference on Concept Mapping, Pamplona, Spain, Sept 14-17, pp. 131-134. Retrieved October 3, 2004, from http://cmc.ihmc.us/papers/cmc2004-045.pdf

Clariana, R. B., Koul, R., & Salehi, R. (2005). The criterion-related validity of a computer-based approach for scoring concept maps. *International Journal of Instructional Media*, *33*(3), in press.

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 structural knowledge. *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). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum and Associates.

Kinchin, I. M. (2001). If concept mapping is so helpful to learning biology, why aren't we all doing it? *International Journal of Science Education*, *23*, 1257-1269.

Kinchin, I. M., & Hay, D. B. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research*, *42*, 43-57.

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., & Schreider, 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*, 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.

Markham, K. M., Mintzes, J. J., & Jones, M. G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching*, *31*, 91-101.

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., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge, MA: Cambridge University Press.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research*, *25*, 407-425.

Ruiz-Primo, M. A. (2000). On the use of concept maps as an assessment tool in science: What we have learned so far. *Revista Electronica de Investigacion Educative, 2* (1). Retrieved May 3, 2004, from http://www.redie.ens.uabc.mx/vol2no1/contenido-ruizpri.html

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2000). Comparison of the reliability and validity of scores from two concept mapping techniques. *Journal of Research in Science Teaching*, *38*, 260-278.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment*, *7*, 99-141.

Safayeni, F., Derbentseva, N., & Cañas, A. J. (2003). *Concept maps: A theoretical note on concepts and the need for cyclic concept maps*. Retrieved October 3, 2004 from http://cmap.ihmc.us/Publications/ResearchPapers/Cyclic Concept Maps.pdf

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

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).

Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning—A report of methodology. *International Journal of Science Education*, *22*, 1221-1246.

Taricani, E. M., & Clariana, R. B. (in press). A technique for automatically scoring open-ended concept maps. *Educational Technology Research and Development*, *53*(4), in press.

Turns, J., Atman, C. J., & Adams, R. (2000). Concept maps for engineering education: A cognitively motivated tool supporting varied assessment functions. *IEEE Transactions on Education*, *43*, 164-173.

Direct reprint requests to:

Ravinder Koul
Associate Professor of Education
Penn State Great Valley
30 E. Swedesford Road
Malvern, PA 19355
e-mail: rxk141@psu.edu