The Consequences of Increasing the Number of Terms Used to Score Open-ended Concept Maps

DR. ROY B. CLARIANA & ELLEN M. TARICANI
The Pennsylvania State University

ABSTRACT

This descriptive study considers the consequence of increasing the number of terms used when scoring open-ended concept maps. Participants (n=24) read an instructional text and drew concept maps of the content, then completed a multiple-choice posttest that measured vocabulary and comprehension. The distances between terms in each participant's concept map were transformed into three proximity arrays using the 16, 26, and 36 most important terms from an expert's map. A Pathfinder Network analysis approach was used to analyze the distance proximity array data. The concept map scores derived using the 16 most important terms were all significantly related to the multiple-choice posttest scores, with Common scores based on term spatial location being most related to comprehension (r=0.57) and Terms scores based on the number of important terms included on the map being most related to vocabulary (r=0.55). However, contrary to expectation, increasing the number of terms used to score the maps did not increase the predictive ability of the map scores, probably due to students not selecting enough of the most important words to include in their maps. Recalling important terms to include in a map appears to be an important and discrete cognitive task. Cautions and implications are provided.

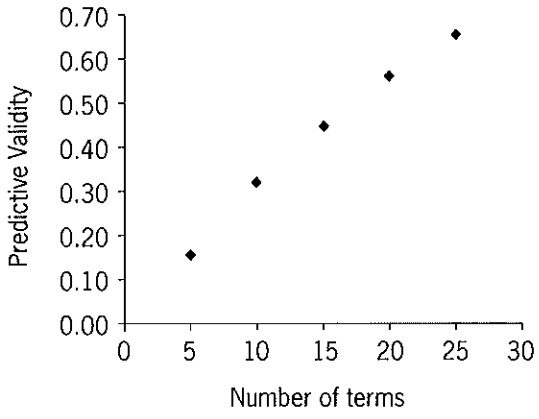
Concept maps (Novak & Gowin, 1984) are sketches or diagrams that show the relationship of a set of terms by the spatial position of the terms and by labeled lines and arrows connecting some of the terms (Jonassen, Beissner, & Yacci, 1993). Concept maps are most often scored by raters using rubrics to help quantify content, links, and the holistic visual arrangement of the concept maps (Ruiz-Primo & Shavelson, 1996), though there is growing interest in computer-based approaches for scoring concept maps (Poindexter & Clariana, 2006).

What information in concept maps can be measured? There are at least three or four different cognitive processing tasks involved when creating a concept map that may leave a residue in the map. First, if the map is 'open ended' where students may use any terms in their map, then a critical task is recalling (or possibly recognizing from a list) the most important terms/concepts to include in their map. Alternately, if a list of terms is provided and the students are told to use all of the terms (fixed or closed mapping), then recall of terms is not a factor. Note that it is easier for both people and computers to score closed maps compared to open maps. Next, students must group related terms together, often in an intuitive way, and this most likely relates to their internal network structure. Then students identify propositions by linking pairs of terms with a line (sometimes labeled) and this most likely relates to the meaning of the proposition in that context. While students work on the later stages of their map, they continually revise small components of their map making it easier to grasp, and this also seems to be an intuitive activity of making it 'feel' right and reflects both the structure of their knowledge and an internalized graphic grammar or norm of what things like this should look like.

This investigation considers an automatic approach for scoring concept maps based on *Pathfinder Networks* (PFNets) that does not require human raters (for details, see Clariana, Koul, & Salehi, 2006; Taricani & Clariana, 2006). This investigation considers the proximity of terms on the map, captured as distances measured in screen pixels. This technique uses the relative spatial location of concepts to communicate hierarchical and coordinate concept relations (Robinson, Corliss, Bush, Bera, & Tomberlin, 2003, p.26; Taricani & Clariana, 2006). This distance data may provide a direct measure of what Yin, Vanides, Ruiz-Primo, Ayala, and Shavelson (2004) call map structure complexity, which are both a local and a global aspect of structural knowledge (p.88, Goldsmith, Johnson, & Acton, 1991).

Goldsmith *et al.* (1991) have shown experimentally that increasing the number of terms used in a pair-wise rating task increases the predictive validity of the resulting *Pathfinder Networks* (*PFNets*) with domain performance measures (e.g., final course grades) in a nearly linear way (see Figure 1).

FIGURE 1. THE RELATIONSHIP BETWEEN THE NUMBER OF TERMS INCLUDED IN PATHFINDER NETWORK ANALYSIS AND THE PREDICTIVE ABILITY OF THE RESULTING PFNETS (GOLDSMITH ET AL., 1991)



A relationship between *Pathfinder Network* derived concept map scores and pair-wise rating task scores have been reported by Clariana and Marker (2007) and by Clariana and Wallace (2009). So, *will increasing the number of terms in an open-ended concept map increase the predictive ability of concept map scores*? This premise is consistent with the traditional test heuristic that more test items are usually better than fewer items.

The ultimate goal of this line of research is to develop a software tool that can be used to create and then automatically score concept maps. When using human raters, it seems reasonable to limit the number of terms used in the maps. Previous investigators have suggested that 25 to 30 terms are the optimal number of terms. This may be due to cognitive load of the raters (and the students) but also the increasing complexity of the spatial arrangement on the page (or screen). Since cognitive load is not an issue for computer-based scoring, this present investigation considers whether including more terms when scoring a concept map improves the predictive ability of the scores. The resulting concept map scores are compared to traditional multiple-choice posttests that measure vocabulary and comprehension of the lesson content, as a measure of the criterion-related validity of the alternate concept map scores. Specifically, a set of concept maps will be scored automatically using (a) 16 terms most critical for the content, (b) 26 terms including the 16 most critical terms plus 10 slightly less critical terms, and (c) 36 terms includes these 26 plus 10 even less critical terms. It is anticipated that, consistent with traditional tests, using more terms in scoring will add salient information, not just error, and so will result in the 'best' scores.

METHOD

The concept maps and the multiple-choice posttest scores used in this present investigation were appropriated from a dissertation by Taricani (2002). Note that this present investigation does not reexamine the original research questions, but rather asks new questions. Also, because the original study used 26 terms, it was necessary to reanalyze the existing paper-based maps to include 36 terms (i.e., the original 26 + 10 more terms).

Participants

In Taricani's (2002) original investigation, undergraduate students were randomly assigned to one of five treatment conditions. Only the learner-generated concept map without feedback treatment was used in this present investigation. The participants (n=24) were freshmen students at a large northeastern university recruited as volunteers from both science and non-science courses. Participation was voluntary and participants were rewarded with either extra course credit or pizza and ice cream for their participation (but not for performance).

Materials

The treatment consisted of completing a generic tutorial on how to create a concept map, a 1,900-word lesson text passage on the human heart developed by Dwyer (1972) called "The Human Heart: Parts of the Heart, Circulation of Blood, and Cycle of Blood Pressure", creating an open-ended concept map of the lesson content, and then taking two 20-item multiple-choice posttests.

The multiple-choice criterion posttest developed and validated by Dwyer (1972) provided 20 questions that dealt with lesson vocabulary and 20 questions that dealt with lesson comprehension. The vocabulary test (lesson terminology) was designed to

measure declarative knowledge of facts, terms, and definitions. The comprehension test was designed to measure a more thorough understanding of the processes of the human heart, with a specific focus on the functions of different parts of the heart. The KR-20 reliability for the posttest was 0.83.

Procedure

Participants completed a brief survey regarding demographic information, their previous experience with concept mapping, and which biology courses they have completed. Next they completed the 2-page generic lesson on how to draw a concept map. Then participants read the 3-page instructional text on the human heart and were asked to draw a concept map of that information on a blank piece of paper *while* reading. Participants *could use any terms and any number of terms in their concept maps.* After completing their maps, the lesson materials and concept maps were collected and then all participants completed the multiple-choice posttests.

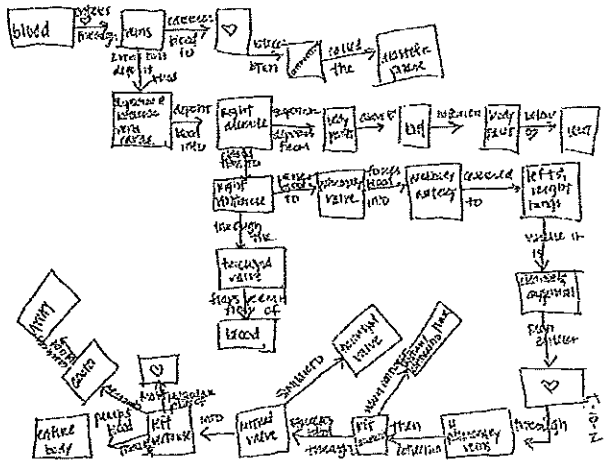
Collecting Raw Concept Map Data

A content expert was provided with a list of all of the terms used by the participants in their concept maps, arranged in order of frequency of occurrence. While considering this list and the instructional text, the expert was asked to draw a concept map with about 35 terms. The expert was not required to use terms from the list, though the list was a priming tool intended to influence the expert's choice of terms. The expert used 36 terms in his map, and this map is used as the referent map for comparison to the participants' maps.

Next the expert was asked to prioritize this list of 36 terms, and then determine the 26 most important terms on the list by removing terms of lesser importance, and finally reducing the list even more to include only the 16 most important terms. The 16 most important terms included: aorta, auricle (left), auricle (right), body, inferior vena cava, lungs, membranes, mitral valve, myocardium, oxygenate, pulmonary artery, pulmonary veins, superior vena cava, tricuspid, ventricle (left), and ventricle (right). The 26 most important terms included these 16 plus aortic valve, cleansed, contract, diastolic, endocardium, epicardium, pericardium, pulmonary valve, relax, and systolic. And finally, the 36 most important terms included these 26 plus arteries, blood, chamber, flaps, heart, muscular, pressure, pump, septum, and vein.

ALA-Mapper software (Clariana, 2002) was used to convert the spatial location of the 36 expert terms in the expert's and in each concept map into a 630-element proximity array containing all of the pair-wise distances between the 36 terms (see Figure 2). To accomplish this, each paper-based concept map was recreated in *ALA -Mapper*, a manual process that can introduce error, though care was taken to maintain the original spatial proportional relationships between terms in the map. When a student did not include one of the 36 terms, all of that term's data elements were coded as "blank" to serve as a flag that the term is missing. The 325-element proximity arrays (26 terms) were established from the 630-element array by deleting row and column data associated with the 10 less important terms from the original 36 terms. Similarly, the 120-element arrays (16 terms) were established from the 630-element array by deleting row and column data associated with the 20 less important terms from the original 36 terms.

FIGURE 2. A STUDENT'S CONCEPT MAP REPRESENTED AS A 120-ELEMENT PROXIMITY ARRAY (I.E., USING 16 TERMS).



	aorta	auricle (left)	auricle (right)	body	inferior vena cava	lungs	membranes	mitral valve	myocardium	oxygenate	pulmonary artery	pulmonary veins	superior vena cava	tricuspid	ventricle (left)	ventricle (right)
aorta	–															
auricle (left)	0.59	–														
auricle (right)	0.65	0.68	–													
body	0.93	0.59	0.60	–												
inferior vena cava	0.75	0.57	0.83	0.69	–											
lungs	0.32	0.69	0.58	0.31	0.41	–										
membranes	0.59	0.21	0.43	0.54	0.60	0.01	–									
mitral valve	0.74	0.84	0.72	0.75	0.68	0.55	0.36	–								
myocardium	0.78	0.38	0.53	0.73	0.69	0.14	0.81	0.53	–							
oxygenate	0.33	0.72	0.56	0.32	0.40	0.94	0.00	0.57	0.14	–						
pulmonary artery	0.45	0.77	0.71	0.43	0.55	0.86	0.15	0.66	0.27	0.84	–					
pulmonary veins	0.45	0.86	0.61	0.46	0.47	0.81	0.09	0.70	0.24	0.85	0.82	–				
superior vena cava	0.73	0.55	0.83	0.66	0.97	0.40	0.60	0.66	0.68	0.39	0.54	0.46	–			
tricuspid	0.72	0.76	0.89	0.69	0.81	0.58	0.42	0.83	0.56	0.58	0.71	0.66	0.79	–		
ventricle (left)	0.89	0.69	0.68	0.90	0.73	0.41	0.49	0.85	0.67	0.43	0.53	0.55	0.70	0.78	–	
ventricle (right)	0.70	0.73	0.93	0.66	0.82	0.58	0.43	0.79	0.55	0.57	0.72	0.65	0.81	0.96	0.75	–

Proximity data (distances) are dissimilarity data, where smaller values indicate stronger relationship. To handle missing terms in the proximity data (the "blanks"), proximity data were scaled and then converted into similarity data by dividing each value in the array by the maximum actual value in that array to create a range of values from 1 to 0, and then this scaled proximity dissimilarity data were inverted into similarity data by subtracting every value in the array from 1. Finally, all "blanks" were converted to 0's, which means that missing terms have no relationship to the other terms on the concept map. Thus proximity data values ranged from 0 to 1 rounded to 2 decimal places (i.e., 0.22), with 0 indicating no association and 1 indicating maximum association.

Converting Raw Concept Map Data into Scores

Knowledge Network and Orientation Tool (*KNOT*, 1998) software was then used to transform all of the participants' and also the expert's proximity arrays (i.e., based on 36, 26 and 16 terms) into *PFNets* using the *KNOT* parameters of Minkowski's r set to infinity, q set equal to n-1, the proposition arrays were defined as "Similarity", the maximum range value was set as 1, and the minimum range value was set as 0.1 (Goldsmith & Davenport, 1990). Note that this last parameter, minimum, is specific to this data set and is necessary in order to exclude the missing terms from the *KNOT* analysis. In this case, all missing terms obtained 0's in the proximity array and so setting the minimum parameter to 0.1 informs *KNOT* to exclude missing terms when forming *PFNets*.

Finally, *KNOT* was used to compare all of the participants' *PFNets* to the expert's. The two most commonly reported similarity measures are *Common* and *Configural Similarity*. *Common* is the sum of the links shared by two *PFNets* (the intersection of two *PFNets*). *Configural Similarity*, which is also called neighborhood similarity, is obtained by averaging the average of the intersection divided by the union for every term in the two *PFNets*. Usually one of the *PFNets* is an expert referent and the other is the participant's *PFNet*, thus *Common* and *Configural Similarity* show the relationship between the participant's concept map and the expert's concept map.

Reliability of the Concept Map Scores

KNOT analysis takes a large amount of raw proximity data and reduces it to the most salient information in that data set. We do not know of a way to measure Cronbach alpha for term data because it is a single value for each student. However, for *Common* (*Map-Cmn*) data it is possible as follows. Note that the expert's *PFNet* based on 36 terms obtained 44 links, the one based on 26 terms obtained 27 links, and the one based on 16 terms obtained 17 links. Thus, for the reliability analysis based on 36 terms, a participant's *PFNet* common (*Map-Cmn*) score can range from 0 (no links in common with the expert) to 44 (all links in common with the expert). In a sense, *Map-Cmn* for 36 terms is like a 44-item test, where each item contributes to the total *Map-Cmn* score. Cronbach alpha for the concept map *Common* were α_{36}=0.90, compared to α_{26}=0.84, and α_{16}=0.68.

RESULTS AND DISCUSSION

The variables used in this investigation include the 20-item vocabulary (*MC-Vocab*) and 20-item comprehension (*MC-Comp*) multiple choice posttests, the number of important terms (*Map-Trm*) actually used by each participant in their concept map, the number of *Common* links (*Map-Cmn*) shared by the participant's and the expert's *PFNets*, and the *Configural Similarity* (*Map-Sim*) of the participant's and the expert's *PFNets*. Descriptive statistics for each included maximum possible score, observed range of scores, mean, and standard deviation (see Table 1).

TABLE 1. MEANS AND STANDARD DEVIATIONS FOR THE MULTIPLE-CHOICE POSTTEST AND THE VARIOUS CONCEPT MAP SCORES

MC Posttest	Maximum	Range	Mean	s.d.
Vocabulary (*MC-Vocab*)	20	5 to 19	10.8	4.1
Comprehension (*MC-Comp*)	20	4 to 19	9.3	4.2
For 16 terms				
Terms (*Map-Trm*)	16	8 to 16	13.3	2.6
Common (*Map-Cmn*)	17	2 to 13	7.3	3.2
Similarity (*Map-Sim*)	0 to 1.00	0.08 to 0.65	0.33	0.16
For 26 terms				
Terms (*Map-Trm*)	26	9 to 26	20.0	4.8
Common (*Map-Cmn*)	27	4 to 17	9.8	4.4
Similarity (*Map-Sim*)	0 to 1.00	0.10 to 0.49	0.27	0.12
For 36 terms				
Terms (*Map-Trm*)	36	12 to 36	27.5	6.1
Common (*Map-Cmn*)	44	4 to 21	11.1	4.8
Similarity (Map-Sim)	0 to 1.00	0.06 to 0.34	0.18	0.08

Comparing Concept Map scores to Posttest Scores

The concept map scores based on 16 terms, 26 terms, and 36 terms are compared to each other and to the multiple choice vocabulary and comprehension test scores using Pearson's correlation (see Table 2). It was anticipated that as the number of terms used to score a concept map increases, the predictive ability of the scores would also increase (Goldsmith *et al.*, 1991). However, the correlations are opposite of what was expected. The predictive ability of the concept map scores decreased as the number of terms increased. Note that the Goldsmith *et al.* approach was NOT open-ended; the students in that case manipulated a pre-determined list of terms. Perhaps concept maps created from a pre-determined list of terms would obtain scores more like those observed in the Goldsmith et al. data. Note also that the *Map-Cmn* scores tended to predict the multiple-choice test scores a little better than the *Map-Sim* scores, a finding that has been observed in other studies. Since *Map-Sim* scores are designed to 'punish' student errors (and guessing) while *Map-Cmn* scores do not, this means that students' map errors are not as important for predictive ability. Specifically, concept map errors do not necessarily predict multiple-choice test errors.

TABLE 2. POSTTEST AND CONCEPT MAP SCORE CORRELATION MATRIX

	MC-Posttest		For 16 terms			For 26 terms			For 36 terms		
	MC-Vocab	*MC-Comp*	*Map-Trm*	*Map-Cmn*	*Map-Sim*	*Map-Trm*	*Map-Cmn*	*Map-Sim*	*Map-Trm*	*Map-Cmn*	*Map-Sim*
For 16 terms											
Terms *(Map-Trm)*	0.55	0.51	1								
Common *(Map-Cmn)*	0.49	0.57	0.81	1							
Similarity *(Map-Sim)*	0.45	0.55	0.73	0.99	1						
For 26 terms											
Terms *(Map-Trm)*	0.37	0.41	0.89	0.68	0.60	1					
Common *(Map-Cmn)*	0.37	0.49	0.78	0.89	0.87	0.81	1				
Similarity *(Map-Sim)*	0.35	0.48	0.70	0.89	0.89	0.69	0.98	1			
For 36 terms											
Terms *(Map-Trm)*	0.31	0.34	0.83	0.58	0.49	0.97	0.74	0.61	1		
Common *(Map-Cmn)*	0.29	0.28	0.60	0.67	0.67	0.73	0.86	0.83	0.71	1	
Similarity *(Map-Sim)*	0.28	0.24	0.51	0.63	0.64	0.62	0.81	0.80	0.60	0.99	1

Values greater than 0.38 are significant at the $p < .05$

In addition, step-wise multiple regression analyses were conducted for the *Map-Trm*, *Map-Cmn*, and *Map-Sim* values at each terms level (for 16, 26, and 36) compared to the multiple choice *MC- Vocab* and *MC-Comp* scores. For concept map scores based on 16 terms, for *MC-Vocab*, only *Map-Trm* entered the equation (r=.548) and for *MC-Comp*, only *Map-Cmn* entered the equation (r=.574). For concept map scores based on 26 terms, for *MC-Vocab*, none entered the equation and for *MC-Comp*, only *Map-Cmn* entered the equation (r=.495). For concept map scores based on 36 terms, for *MC-Vocab*, none entered the equation and for *MC-Comp* none entered the equation. These multiple-regression findings suggest that recalling terms to include in your concept map relates to your vocabulary test score, while the geometric positions of terms on the map relate to your comprehension. Though concept map errors do not relate to multiple choice test errors, here concept map 'corrects' do relate to multiple choice test 'corrects'.

DISCUSSION

Our goal for this line of research is to develop a software tool that can be used to create and then automatically score concept maps. This investigation considers the consequences of including more terms in the mathematical analysis process. The concept map scores derived using the 16 most important terms were all significantly related to the multiple-choice posttest scores, with *Common* scores (i.e., *Map-Cmn*) based on term spatial location being most related to comprehension multiple-choice test scores (r=0.57) and Terms scores (i.e., *Map-Trms*) based on the number of important terms included on the map being most related to vocabulary multiple-choice test scores (r=0.55).

However, contrary to expectation, increasing the number of terms used to score the concept maps decreased the predictive ability of both map scores that are derived from the geometric location of terms on the map (*Common and Configural Similarity*) and also those measures based simply on the number of terms included in the map. Probably these students did not select enough of the most important words to include in their maps, possibly because of convention (maps feel right with about 25 to 30 terms) or perceived complexity of their maps. Students were more likely to include the 16 most critical terms in their maps but as the number of terms used to score the maps increase (i.e., 26 and then 36), students included some and excluded other of the less important terms. Apparently, a few of the most important terms contribute the most to the predictive ability of the concept maps created by these non experts.

One logical and obvious finding of this investigation that wasn't obvious at the start is that map scores based on links and term location are bound by the number of important terms included during open-ended concept map creation. A participant cannot link to a term that is not present on the map. For this reason, the kind of information captured in open-ended concept map is fundamentally different than that in fixed/closed concept maps.

Many investigators refer to open-ended concept mapping, where participants may use any concepts and linking terms in their maps, as the gold standard for capturing students' knowledge structures (McClure, Sonak, & Suen, 1999; Ruiz-Primo, Schultz, Li, & Shavelson, 1999; Yin et al., p.24). As noted above, the number of important terms included in an open ended concept map is a confounding variable in automatically scoring these maps because the number of terms is profoundly related to the number of links that can be formed. Looking back at Clariana et al. (2006), that study also found that the number of important terms included in the concept maps was strongly related to the human-rater concept map scores (r=0.75). Is there a compromise approach to maintain the open concept mapping gold standard yet escape the terminology recall penalty?

In the headings as signals research base, eliciting lists of important words from participants after they have read a text passage is commonly used as a measure of knowledge of text structure (Clariana & Marker, 2007). Pragmatically, a simple concept mapping approach could consist of software that first asks students to list all the important terms that they would like to include in their concept maps, and on a second screen, it would provide students with the a preset list of important terms to use to actually create their maps. This approach would allow the software to capture both term recall and then term spatial relationships.

What is the value of a tool that can automatically score concept maps? Concept map scores can complement traditional assessment formats (e.g., multiple-choice, constructed response, essay, lab practical) in both the classroom and also in large-scale assessment, such as statewide tests of students' science content knowledge. Such a tool could be used at the classroom-level for both instruction and assessment by providing teachers with another way to gauge student understanding of important concepts and to identify student misconceptions. Further refinement of the tool and the approach is warranted.

Direct Reprint Requests to:
Dr. Roy Clariana
Penn State University
30 E. Swedesford Road
Malvern, PA 19468
Email: RClariana@psu.edu

REFERENCES

Clariana, R. B. (2002). *ALA-Mapper, version 1.0*. Available for download online: http://www.personal.psu.edu/rbc4

Clariana, R.B., Koul, R., & Salehi, R. (2006). The criterion-related validity of a computer-based approach for scoring concept maps. *International Journal of Instructional Media, 33*(3), 317-325.

Clariana, R. B., & Marker, A. W. (2007). Generating topic headings during reading of screen-based text facilitates learning of structural knowledge and impairs learning of lower-level knowledge. *Journal of Educational Computing Research, 37*(2),173-191.

Clariana, R. B., & Wallace, P. E. (2009). A comparison of pair-wise, list-wise, and clustering approaches for eliciting structural knowledge. *International Journal of Instructional Media, 36*(3), 287-300.

Dwyer, F. M. (1972). A guide for improving visualized instruction. State College, PA: Learning Services.

Goldsmith, T.E., & Davenport, D.M. (1990). Assessing structural Similarity of graphs. In Schvaneveldt (ed.), *Pathfinder associative networks: studies in knowledge organization*, 75-87. 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.

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

KNOT (1998). Knowledge Network and Orientation Tool, version 4.3. Available online from http://interlinkinc.net/.

McClure, J., Sonak, B., & Suen, H. (1999). Concept map assessment of classroom learning: reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36*, 475-492.

Novak, J.D., & Gowin, D.B. (1984). Learning how to learn. Cambridge, UK: Cambridge University Press.

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*(2),177-184.

Robinson, D.H., Corliss, S.B., Bush, A.M., Bera, S.J., & Tomberlin, T. (2003). Optimal presentation of graphic organizers and text: A case for large bites? *Educational Technology Research and Development, 51*(4), 25-41.

Ruiz-Primo, M. A., & Shavelson, R.J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*, 569-600.

Ruiz-Primo, M. A., Schultz, S., Li, M., & Shavelson, R. J. (1999). *On the cognitive validity of interpretations of scores from alternative concept mapping techniques.* (CSE Tech. Rep. No. 503). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).