Cross-Validation Study of Methods and Technologies to Assess Mental Models in a Complex Problem Solving Situation

Abstract

This paper reports a cross-validation study aimed at identifying reliable and valid assessment methods and technologies for natural language (i.e., written text) responses to complex problem-solving scenarios. In order to investigate current assessment technologies for text-based responses to problem-solving scenarios (i.e., ALA-Reader and T-MITOCAR), this study compared the two best developed technologies to an alternative methodology. Comparisons amongst the three models (benchmark, ALA-Reader, and T-MITOCAR) provided two findings: (a) the benchmark model created the most descriptive concept maps; and (b) the ALA-Reader model had a higher correlation with the benchmark model than did T-MITOCAR's. The results imply that the benchmark model is a viable alternative to the two existing technologies and is worth exploring in a larger scale study.

Keywords: assessment technology, concept map, mental models, problem solving, validation study

Cross-Validation Study of Methods and Technologies to Assess Mental Models in a Complex Problem Solving Situation

1. Introduction

This study investigated current methods and technologies that yield concept maps − structural knowledge representations consisting of concepts and relations [1–4] − as re-representations of a student’s mental models. This study is a type of cross-validation aimed at identifying the methods that work best in terms of forming the basis for dynamic formative feedback. It is assumed that using natural language (written text) responses as a basis for concept map representations of student thinking is likely to provide a reliable foundation for use in providing formative feedback and assessment [5].

There is a common belief that problem solving includes conceptualizing the problem space, which involves creating a knowledge structure that integrates ideas and concepts that a problem solver associates with the problem situation [6–9]. As a consequence, assessing problem solving should naturally take into account the constructed knowledge structure [10]; simple knowledge tests are somewhat weak measures of problem-solving ability [11].

In order to capture structural knowledge, a number of technologies have been developed, including: DEEP (Dynamic Evaluation of Enhanced Problem-solving; [4]); SMD (Surface, Matching, and Deep Structure; [12]); T-MITOCAR (Text Model Inspection Trace of Concepts and Relations; [5]); CmapTools [3]; jMap [13]; ACSMM (Analysis Constructed Shared Mental Model; [14]); KU-Mapper [15]); ALA-Mapper (Analysis of Lexical Aggregates-Mapper; [16,17]); ALA-Reader (Analysis of Lexical Aggregates-Reader; [15,17]); and KNOT (Knowledge Network Orientation Tool; [18]).

Current technologies either require learners to create an annotated concept map with rich descriptions of links and nodes (DEEP) or else they use text responses as an interim step in generating a concept map (T-MITOCAR and ALA-Reader) that can then be assessed with tools such as SMD or KNOT. All these technologies have limitations in terms of their suitability, reliability, and validity [19–21].

This paper focuses on methods that use text responses to generate a concept map that can then be assessed and explores an alternative approach that attempts to restore rich descriptions of links between nodes. Prominent methods and technologies are classified and analyzed in terms of their merits and deficiencies. Next, alternative methods and technologies to analyze student responses in the form of written text are selected. Finally, cross-validation among the selected technologies is performed, analyzed and reported. Based on the results, an alternative approach to consider in automatically constructing and assessing concept maps based on open-ended text responses to a problem situation is then described.

2. Concept maps as re-represented mental models through language inputs

2.1 Mental models as inferred entities

Mental models are cognitive artifacts resulting from perception and linguistic comprehension, representing certain aspects of external situations [22,23]. In this perspective, knowledge appears to be a configuration of holistic mental representations.

Mental model representations consist of propositional representations as structured symbols and images [22,24]. A concept map, such as the externally-represented structural component of mental models, implies that a latent structure exists in the human mind. In other words, a concept map is a first or second order representation of a primary representation – a mental model, which is an inferred entity. The primary representations (mental models) drive actions and decisions, which are external indicators of learning. In order to provide formative feedback, however, it is necessary to make inferences about the mental models that are behind decision making and problem-solving activities. Many empirical studies utilizing concept map techniques have shown that as students gain competence in a discipline, their structural comprehension becomes more coherent and expert-like [4,25,26].

2.2 Concept maps as analogues of mental models

Concept mapping is a method that elicits cognitive representations of an individual's structural knowledge involving interrelated concepts [2,27,28]. In concept maps as representations of semantic networks [7,29,30], the strength of links may be interpreted as the strength of belief in a given semantic relationship, which is reflected by link weights [31,32].

Language plays a key role in creating a concept map. Concept maps can represent pairs of related words, such as a noun (concept)-verb (relation)-noun (concept) relationship. The data used for concept mapping is generally collected from interviews or texts [2,28]. Text-based data collection is economical in terms of time and effort [33] and is based on techniques that avoid recall bias and potentially leading or misleading questions [27].

Mental models are formulated symbolically [34]; that is, symbols play a central role in representing ideas and thoughts. Meaning is constructed with cognitive effort (thinking and reasoning) often utilizing symbolic notations which help individuals to re-represent their thoughts [35]. According to Garnham [36,37], the theory of mental models provides a unified account of language processing, thinking, and reasoning. A mental model is constructed based on situational inputs and can be re-represented in text or discourse. In order to visually represent concept maps from text, technological support in terms of natural language processing and network analysis technology are required.

3. State-of-the-art concept map technologies

Concept maps are generally visually represented through network analysis using a set of techniques to portray patterns of relations among nodes [38–40]. Most of the techniques involve mathematical algorithms derived from graph theory [40–42]. In these techniques, proximity data between and among concepts is defined as “judgments of similarity, relatedness, or association between entities frequently used in the study of human cognition” [42]. Drawing on graph theory and proximity data, specific statistical methods have been used. Pathfinder technique [7,18,26] to analyze simple association networks using multi-dimensional scaling is an early statistical technique used to assess concept maps. More recently, social network analysis has been used [40,43,44].

Figure 1 illustrates relationships amongst methods and technologies in a network analysis procedure. In general, network analysis employs a three-step procedure [16,45]: (1) elicit judgments about concept relationships; (2) construct concept maps; and (3) compare the concept maps to the reference model.

3.1 Step 1: Elicit judgments about concept relationships

The first step, eliciting judgments about concept relationships, is the essential phase because it yields data that contains all the captured concepts and their relationships in a student’s response. There are two kinds of concept map approaches involving natural language processing: the ‘closed-ended concepts approach’ and the ‘open-ended concepts approach’ (see Figure 1). The closed-ended provides the student with a predefined list of concepts and links [16]. The open-ended approach allows the student to use whatever concepts and linking terms he or she desires. In Figure 1, the class of closed-ended approaches includes KU-Mapper, ALA-Mapper, and ALA-Reader; the class of open-ended approaches includes T-MITOCAR, CmapTools, and

DEEP. CmapTools and DEEP allow students to draw concept maps based on either predefined concepts or free use of any concepts. ALA-Reader accepts text input with no limitation [16,17], but that tool is classified as closed-ended concepts because the technology only retrieves a predefined list of words from the open-ended text. Many researchers consider open-ended concept mapping as the gold standard for capturing students' mental models [4,16,46,47]. This study also centers on open-ended concept mapping as a way to elicit a natural and descriptive knowledge structure.

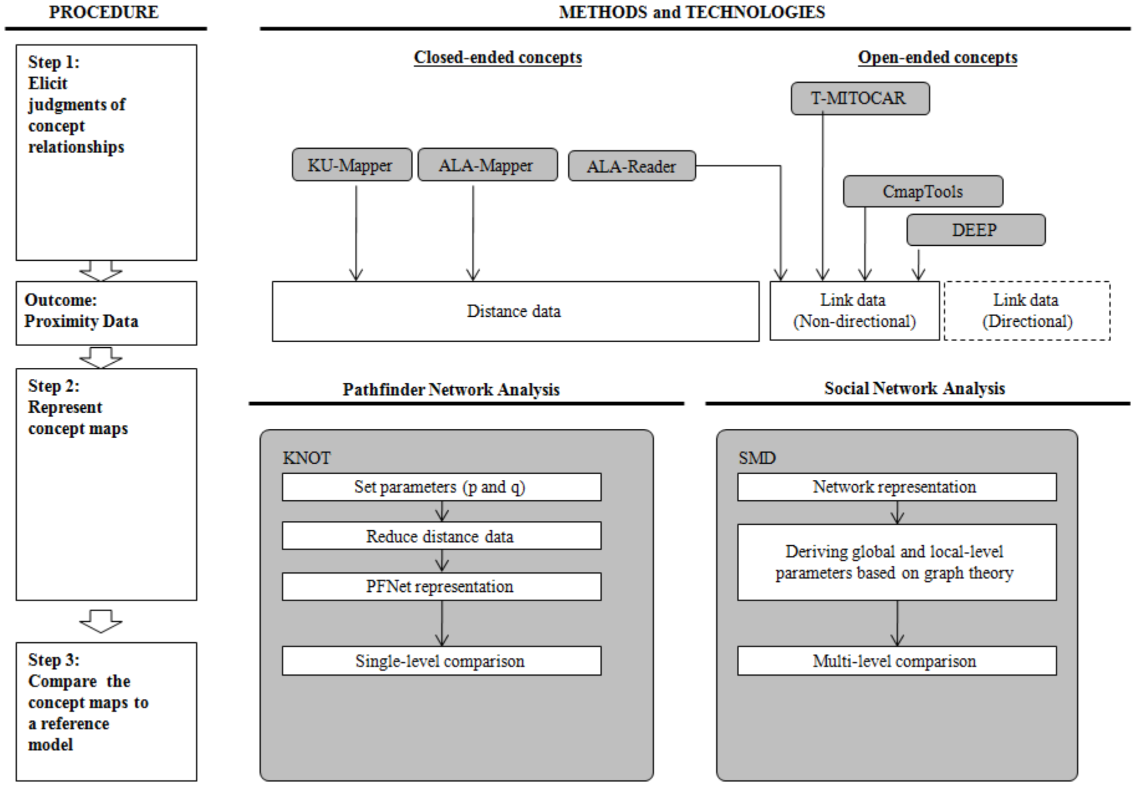

Figure 1. Procedure, methods, and technologies applied to network analysis.

The aforementioned technologies generate two types of proximity data (i.e., an *n* by *n* matrix where *n* is the number of terms−concepts), representing distance or adjacency, depending on the technologies. Distance data generated by such tools (KU-Mapper and ALA-Mapper) include all of the pair-wised distances between terms that are calculated by the location of the terms in a space (e.g., computer screen) or directly judged by students with an ordinal scale (e.g., 1 to 9) [16,17]. In the case of adjacency data, the relatedness of paired terms is represented in a matrix, the *n* by *n* matrix with ‘1’ (that is entered when two terms are connected) or ‘0’ (that is entered when two terms are not associated) [17,18,40–42]. Distance data have mathematical foundations in geometry whereas adjacency data are directly derived from graph theory [18,42]. Adjacency data is an alternative to distance data due to their benefit in describing a directional relation in which the connection between a pair of concepts has an origin and a destination [40]. In the sense that directional adjacency data can describe detailed structural information such as a causal relation between concepts, this study takes account of adjacency data that is more descriptive and complex.

In addition, Clariana and colleagues [17,48] proposed that editing pronouns to nouns the pronouns point out in text is likely to help capture relevant information about the knowledge structure better than does simply ignoring all pronouns. Although their experiment concluded that there was little difference between editing and ignoring pronouns, this study included a comparison between the two types of data when determining a better benchmark model.

In regard to the classifications discussed here, we can make a set of combinations of concept mapping approaches that are catalogued from more complex and descriptive to simple and economic. Again, all combinations listed below can be divided into two sets (i.e., noun-only and pronoun-edited) according to the data handling subroutine:

- Natural language + open-ended concepts + directional adjacency data
- Natural language + open-ended concepts + non-directional adjacency data
- Natural language + closed-ended concepts + directional adjacency data
- Natural language + closed-ended concepts + non-directional adjacency data

3.2 Step 2: Construct concept maps

The technologies process proximity data (either distance or adjacency data) and then construct concept maps. Constructing concept maps is based on two network analysis methods: Pathfinder network analysis and social network analysis. Pathfinder network analysis (PFA) is "a data reduction approach that emphasizes the main pair-wise associations in proximity data" [16] by constraining the data with *r* and *q* parameters. *r* parameter works as a function of the weights of links in the path and *q* parameter places an upper limit on the number of links in paths [18,42]. In short, pathfinder algorithms with particular *r* and *q* parameters yield the minimum number of links in an *n* by *n* array. Considering the function of PFA, it seems that PFA is a heuristic method for finding meaningful relations among many of the concepts [49]. However, it is often observed that many studies with a small set of concepts utilize PFA. Likewise, in the case that key terms are pre-defined from the referent network, the use of PFA would be less effective. Otherwise, the approach to identifying key terms and relations in a large number of concepts appears to be more pertinent to PFA. For example, Coronges and colleagues [38] conducted a more descriptive analysis (i.e., social network analysis) and then ran PFA.

Social network analysis (SNA) is useful in the study of knowledge structure although it was developed to study social relationships [38,40,50,51]. Coronges and colleagues [38] used SNA to analyze a cognitive network labeled Cognitive Associative Network (CAN). Another example is the Epistemic Network Analysis (ENA) by Rupp and colleagues [41,43] in which they created knowledge networks (concept maps) and compared the networks using measures resulting from SNA. Network measures computed by SNA include global-level measures that describe the entire network (e.g., centralization, size, density, clustering, and path length) and local-level measures, such as the centrality of each concept, which can be used to determine the salience of concepts [38,40].

Table 1

Measures for Analyzing the Organization of Cognitive Structure

SMD Measure	Operationalization	Social Network Measure[a]
Surface structure	The overall number of propositions	The number of edges (links)
Graphical structure	The complexity of a cognitive structure indicates how broad the understanding of the underlying subject matter is	Geodesics distance and diameter of a network
Connectedness	A connected cognitive structure indicates a deeper understanding of the underlying subject matter	Connectedness
Ruggedness	Non-linked vertices of a cognitive structure point to a lesser understanding of the phenomenon in question	Non-linked nodes
Average degree of vertices	As the number of incoming and outgoing edges grows, the complexity of the cognitive structure is taken as more complex	Average degree
Cyclic	A non-cyclic cognitive structure is considered less sophisticated	Cycle (closed walk) or acycle
Number of cycles	A cognitive structure with many cycles is an indicator for a close association of the vertices and edges used	Cohesive subgroups
Vertices	A simple indicator for the size of the underlying cognitive structure	The number of nodes
Vertex matching	The use of semantically correct concepts (vertices) is a general indicator of an accurate understanding of the given subject domain	Shared number of nodes
Propositional matching	The use of semantically correct propositions (vertex-edge-vertex) indicates a correct and deeper understanding of the given subject domain	Configural similarity (KNOT; Taricani & Clariana, 2006)

Note. It was modified from Ifenthaler et al., [52]. a. It referred to Wasserman and Faust [40].

As a result, when we use adjacency data (either directional or non-directional), employing SNA provides more benefits for research in that it generates diverse network parameters that are compared to a reference network or other networks (e.g., a previously elicited network). SMD, a set of analysis functions using adjacency data, is consistent with SNA. Table 1 shows that the measures used in SMD are directly adapted to the analysis techniques specified in SNA.

Table 2

Seven Similarity Measures of T-MITOCAR

Surface	Compares the number of concepts between two models (graphs).
Graphical Matching	Compares the diameters of the spanning trees of the graphical mental models as an indicator for the range of conceptual knowledge.
Structural Matching	Compares the complete structures of two graphs regardless of their contents.
Gamma	Compares the two graphs' gammas that indicate each graph's average percentage of the links that are actually present for a node.
Concept Matching	Compares the sets of concepts within a graph to determine the use of terms.
Propositional Matching	Compares only fully identical propositions between two graphs, which are used for quantifying semantic similarity between two graphs.
Balanced Semantic Matching	Uses both concepts and propositions to match the semantic potential between two model representations.

Note: It refers to Pirnay-Dummer and Ifenthaler [55].

3.3 Step 3: Compare the concept maps to the reference model

Evaluation of the derived concept maps is often done by comparison with a reference model, which is often elicited from an expert [16,38,45,53]. Comparison between concept maps is indicated by similarity measures assessed by overlaying network patterns with the concept map information [38,54]. KNOT (i.e., a PFA tool mostly employed in studies) software includes the function to gauge two similarity parameters: common and configural similarity. The common similarity is simply the total number of links shared by two concept maps. The configural similarity is calculated by dividing the total number of shared propositions by the total number of unique links in the student's and the expert's concept maps ranging from 0 (no similarity) to 1 (perfect similarity). These similarity values are used as concept map scores [16,17]. Similarity measures can be extended to include diverse network parameters. For example, T-MITOCAR provides six similarity measures as comparison results based on multiple network measures from SMD (see Table 2).

4. Method

This study has two aims: (1) to identify the methods that consistently yield the most descriptive and accurate concept maps; (2) to validate these methods and technologies using a benchmark method. The most complex condition can be expressed in this combination: natural language + open-ended + directional adjacency data + pronoun-edited. As to the network analysis method, the social network analysis (SNA) is considered as an alternative in terms of obtaining a more descriptive concept map and diverse network information. However, considering that a complex approach is only required when it offers a greater benefit over a less complex one, more complex approaches were compared to less complex approaches. In addition, this comparison process serves to validate applied technologies as well.

4.1 Participants

Participants included 20 students and seven experts. The original student data were gathered from 136 undergraduate students enrolled in a course at a university in the southern United States. The course aimed to educate students on knowledge and skills for integrating technology in teaching and learning. In the class, students made written responses to a specific complex problem. For this study, from the original group of students whose responses contained more than 350 words, a random selection of 20 students was made; the reason for restricting the selection to responses with more than 350 words is that one of the selected technologies, T-MITOCAR, requires at least 350 words for an analysis. Fifteen students were female and five were male. Of the 20, 10 were in their junior year, five were sophomores, and five were at the senior level. Seven expert responses were gathered from seven professors teaching at six major universities in the United States. It was assumed that including expert responses would enable us to investigate the technologies' ability to detect higher-level responses because expert responses are ranked at the top in an accurate and reliable concept map analysis technology.

4.2 The problem-solving task

All participants were asked for responses to a complex problem situation using natural language. The task provided a simulated situation in which students were assumed to be participating in an evaluation project, the purpose of which was to investigate an unsuccessful project that had as its goal adapting a technology (i.e., a tablet PC) for classroom teaching. In order to elicit students' knowledge in detail, the questions asked them to explicitly describe the concepts, issues, factors, and variables likely to have contributed to the result that the introduction of tablet PCs had very little effect on the instructional practices employed in the classes.

4.3 Reference modeling via a Delphi survey

This study included a reference model for the problem situation that is compared to a student's concept map in the form of a learner model so as to obtain concept map scores of each learner model. The reference model as a concept map of a targeted expert was elicited from a reference response. The reference response was created using a Delphi survey procedure [56–58]. The Delphi survey involved three iterations to develop a refined reference model that the seven experts accepted. In the first round, the participating professors created their own responses to the problem (those initial responses were used as their own inputs in the data); then, all the panel's responses were consolidated. Next, a document including all statements from the professors and a list of concepts identified from the panel's responses was again sent to the panel. The professors were asked to add their comments regarding the listed statements and concepts and rank them. After gathering the second round of surveys, the researcher created a final list of ranked statements and concepts. Based on this summary, the draft of a reference model was created. In the final round, the results of the second survey were sent to the panel and revised as necessary according to their comments. Through this procedure, a reference model containing 23 key concepts was developed.

4.4 Benchmarks

Initially, four types of benchmarks were prepared according to the combinations of noun-only vs. pronoun-edited and directional vs. non-directional. To prepare for pronoun-edited responses, all responses were reviewed. Directional relations were determined when retrieving paired concepts from text.

To distill concepts and relations from text responses, according to the author's [ ] semantic relation approach, the researcher manually distilled the semantic relations that are the underlying relations between two concepts expressed by words or phrases. The approach involves diverse types of relations of concepts beyond the typical noun-verb-noun relation form including genitives (e.g., teachers' participation), prepositional phrases attached to nouns (e.g., technology in school classrooms), or sentence (e.g., Emerging new media has always led to instructional changes.). Thereafter, all distilled concepts and relations were summarized in adjacency data. The distilled concepts and relations were cross-checked by a doctoral student. There were no significant issues regarding the data. All concepts and relations were retained without changes; potential issues would have had no significant impact on the results considering the number of concepts and relations in each concept map. The adjacency data were processed via the selected network analysis package, NetMiner (http://www.netminer.com/NetMiner/home_01.jsp), which provides the capability for creating concept maps and generating a variety of concept map information.

Finally, in order to obtain a list of similarity measures in the form of concept map scores, student models were compared to the reference model. The similarity measures were calculated using the similarity tool developed for this study. The tool was developed using C++ and validated by comparisons between randomly selected tool-generated and manually calculated sample data.

4.5 Cross-validation procedure

Along with multiple benchmarks, two natural language technologies, T-MITOCAR and ALA-Reader, were selected and validated against the benchmark model. Figure 2 illustrates the cross-validation procedure. The concept mapping approach in which each method is embedded is matched as follows:

- Benchmark 1: open-ended + directional (adjacency data)
- Benchmark 2: open-ended + non-directional (adjacency data)
- T-MITOCAR: open-ended + non-directional (distance data)
- ALA-Reader: close-ended + non-directional (adjacency data)

Regarding the benchmarks, there are two types of models: noun-only and pronoun-edited. Each model can be reclassified as directional or non-directional. In Figure 2, only the second-level classification was included. At the beginning of the analysis, the competing models of the benchmarks such as noun-only versus pronoun-edited and directional versus non-directional were compared so that a benchmark model could be determined for comparisons amongst different methods and technologies.

As Figure 2 shows, there were four comparisons made across the outcomes of each selected method through the concept mapping procedures (steps 1 through 3). The cross-validation includes two comparisons of outcomes from different sets. In the first review, the correlation and similarity were analyzed for comparisons amongst the proximity data. In the second review, visual inspections of the concept maps were conducted as a qualitative analysis. The third review involved a second correlation and similarity study among concept map parameters obtained from each method: (a) number of relations; (b) diameter; (c) gamma (cluster coefficient; see Table 2); and (d) the number of cycles (cohesive subgroups; see Table 1).

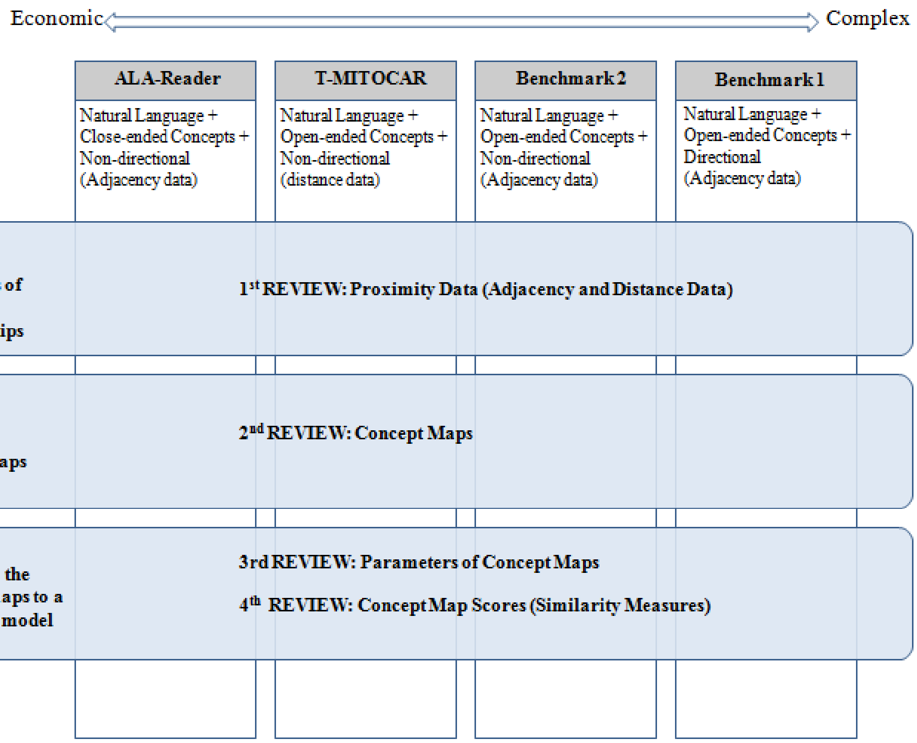
Economic
Complex
ALA-Reader
Natural Language + Close-ended Concepts + Non-directional (Adjacency data)
T-MITOCAR
Natural Language + Open-ended Concepts + Non-directional (distance data)
Benchmark 2
Natural Language + Open-ended Concepts + Non-directional (Adjacency data)
Benchmark 1
Natural Language + Open-ended Concepts + Directional (Adjacency data)
Step 1: Elicit judgments of concept relationships
1st REVIEW: Proximity Data (Adjacency and Distance Data)
Step 2: Represent concept maps
2nd REVIEW: Concept Maps
Step 3: Compare the concept maps to a reference model
3rd REVIEW: Parameters of Concept Maps
4th REVIEW: Concept Map Scores (Similarity Measures)

Figure 2. Cross-validation procedure.

The last review was a comparison across various concept map scores, which are computed similarities between the student model and the reference model (see Pirnay-Dummer & Ifenthaler, 2010): (a) surface matching; (b) graphical matching; (c) concept matching; (d) gamma matching; (e) propositional matching; and (f) balanced semantic matching.

4.6 Data Analyses

4.6.1 Data comparison

This study validated the outcomes from the competing concept map methods such as proximity data, concept map parameters, and concept map scores. The validation methods included the numerical and pattern reviews. In the following, a numerical similarity measure was applied to see how far or near the numbers of the two models are:

$$s = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)}$$

where f_1 and f_2 denote the numerical frequency of each method compared. The similarity ranges from 0 to 1, $0 \leq s \leq 1$. To review associated patterns, Pearson correlation-coefficients were calculated between the benchmark and competing methods.

4.6.2 Modification of similarity measures (concept map scores)

This study modified and adjusted the formulas that calculated the concept map scores. The modified formulas were used in obtaining concept map scores of the benchmark and ALA-Reader.

On the whole, a similarity formula assumes each part of a pair is equally significant. In the case of a concept model comparison, the reference model and student model are not equal in terms of maturity. A reference model acts as criteria and a student model is expected to progress toward the reference model. It is assumed that a reference model is likely to contain a greater number of concepts and relations and to build a larger knowledge structure than a novice model [4,59].

As for numerical similarity, a modified algorithm was applied to all concept map scores except the gamma similarity (refer to Table 2). In case f_1 is smaller than f_2, $f_1 < f_2$, the original numerical similarity formula was used so that

$$s = 1 - \frac{|f_1 - f_2|}{\max(f_1, f_2)}$$

where the frequency of a student model is f_1 and that of a reference model is f_2. Otherwise, if f_1 is not less than f_2, $f_1 \geq f_2$, the similarity value was set as ‘1’ because the student value is greater than that of the reference. That is, it indicates that the student model exceeds the reference model according to the relevant criteria.

Similarly, concerning the conceptual similarity as applied to the concept and propositional matching score, two adjustments were made. Just as a picture resembles an object rather than an object resembling the picture, a student model to some degree resembles the reference model that is more salient. In this asymmetric relationship, the features of the student model are weighted more heavily than those of the reference [60,61]. When the conceptual similarities of the benchmark was calculated by Tversky’s [62] formula,

$$\mathrm{s} = \frac{f(\mathrm{A} \cap \mathrm{B})}{f(\mathrm{A} \cap \mathrm{B}) + \alpha \cdot f(\mathrm{A} - \mathrm{B}) + \beta \cdot f(\mathrm{B} - \mathrm{A})}$$

α was weighted more heavily than β ($\alpha = 0.7$ and $\beta = 0.3$). However, in the case of the ALA-Reader, since it has a predefined set of concepts and relations, the reference mode, f (B), always includes a student model, f (A). Therefore, α was set as ‘0,’ whereas β was set as ‘1.’

5. Results

5.1 Determining a benchmark

Two comparisons amongst the methods (the noun-only versus pronoun-edited and the directional versus non-directional) for creating benchmark models were implemented so that we could decide a reliable and economical way to establish a benchmark. The first review was of the noun-only and pronoun-edited using the numerical similarity between the two approaches.

As Table 3 summarizes, very high average numerical similarities (s > 0.93) were observed between the noun-only and pronoun-edited data of concepts and relations in both the directional and non-directional models. Similar to Clariana and colleagues' [17,48] conclusion, there was such little difference that it was determined that the noun-only model was more economical.

Table 3

Numerical Similarities of Concepts and Relations between the Noun-Only and Pronoun-Edited

	Noun-Only vs. Pronoun-Edited (Directional)		Noun-Only vs. Pronoun-Edited (Non-Directional)	
	Concept	Relation	Concept	Relation
Min	0.74	0.70	0.74	0.70
Max	1.00	1.00	1.00	1.00
Average	0.97	0.93	0.97	0.93

Note. Sample size *N* = 28.

Next, the similarities between the directional and non-directional data were investigated (see Table 4). The directional and non-directional data have the same set of concepts in a given setting, either the noun-only or pronoun-edited. Only the numerical similarities of relations were calculated. The very high average numerical similarities (*s* = 0.98) demonstrated that there was little difference between the directional and non-directional data of relations in both noun-only and pronoun-edited models.

Table 4

Numerical Similarities of Relations between the Directional and Non-Directional

	Directional vs. Non-Directional (Noun-Only)	Directional vs. Non-Directional (Pronoun-Edited)
Min	0.92	0.92
Max	1.00	1.00
Average	0.98	0.98

Note. Sample size *N* = 28.

In order to certify the aforementioned findings, matrix correlations amongst the four types of benchmarks were reviewed for the selected four samples, including two randomly selected samples (students 78 and 83), the reference model, and an extreme case (student 20). As an outlier, student 20 frequently used pronouns resulting in a relatively greater difference between noun-only and pronoun-edited analyses.

Table 5

Correlation Matrix among the Four Types of Data of the Selected Samples

	Reference Model			
	ND	NN	PD	PN
ND	0	0	0	0
NN	0.516	0	0	0
PD	0.985	0.512	0	0
PN	0.508	0.985	0.515	0
	Student 20			
	ND	NN	PD	PN
ND	0	0	0	0
NN	0.5	0	0	0
PD	0.703	0.413	0	0
PN	0.351	0.703	0.5	0
	Student 78			
	ND	NN	PD	PN
ND	0	0	0	0
NN	0.52	0	0	0
PD	0.895	0.49	0	0
PN	0.464	0.891	0.518	0
	Student 83			
	ND	NN	PD	PN
ND	0	0	0	0
NN	0.531	0	0	0
PD	0.895	0.522	0	0
PN	0.486	0.914	0.543	0

Note. The ND denotes the Noun-Only & Directional data; the NN is the Noun-Only & Non-Directional data; the PD indicates the Pronoun-Edited & Directional data; and the PN means the Pronoun-Edited & Non-Directional data.

For all samples, matrix correlations between the directional and non-directional were around 0.5 because the relation R_{ij} is the same as R_{ji} in non-directional data, $R_{ij} = R_{ji}$, whereas R_{ij}

is different from $R_{ji,}$ in directional data. In a specified directional or non-directional condition, even student 20 had a higher correlation of 0.703 between the noun-only and pronoun-edited data (see Table 5). It was concluded that a noun-only and non-directional data model is sufficient to describe a useful concept map. Next, the proximity data (adjacency data) of the benchmark was created.

5.2 1st review: the proximity data

The proximity data obtained from the benchmark, ALA-Reader, and T-MITOCAR were compared. As Table 6 shows, the benchmark model had the greatest number of concepts and relations followed by those of T-MITOCAR and ALA-Reader.

Table 6

The Numbers of Concepts and Relations of the Benchmark Model, ALA-Reader, and T-MITOCAR

	Concept			Relation		
	B	A	T	B	A	T
Min	16	0	8	18	0	7
Max	54	23	19	64	35	46
Average	33	5	13	39	6	23

Note. Note. Sample size *N* = 28. The B denotes the benchmark model; the A is the ALA-Reader; and the T means the T-MITOCAR.

In order to further investigate the effect of concept mapping technologies, a one-way within subjects (or repeated measures) ANOVA was conducted for the number of concepts and relations, respectively. The reason to use a repeated measures ANOVA was that participants' responses are subjected to the three technologies, and the concepts and relations distilled from each of these technologies want to be compared.

As Table 7 describes, a repeated measures ANOVA with a Greenhouse-Geisser correction determined that the mean number of concepts and relations differed statistically significant between the concept mapping technologies [*F*(1.35, 36.42) = 188.49, *P* < 0.01; *F*(1.49, 40.23) = 98.54, *P* < 0.01, respectively]. Post hoc tests using the Bonferroni correction revealed that the numbers of concepts and relations from the benchmark model significantly differed from those of ALA-Reader and T-MITOCAR (*p*= 0.01 for both concepts and relations) (see Table 8).

Table 7.

Summary of Within-Subjects ANOVA

	Sum of Squares	*df*	Mean Square	*F*	Partial Eta Squared
Concept	11731.52	1.35	8697.42	188.49**	.875
Relation	15477.24	1.49	10387.14	98.54**	.785
Error	1680.48	36.42	46.143		
	4240.76	40.23	105.41		

Note. **p < 0.01. Greenhouse-Geisser values were used because the data violated the assumption of sphericity.

Table 8

Bonferroni Comparison for Week of Weight Measurement

Comparisons	Mean Weight Difference	Std. Error	95% CI	
			Lower Bound	Upper Bound
B vs. A	28.21**	0.64	24.59	31.84
B vs. T	19.71**	1.04	14.84	24.59
T vs. A	8.50**	0.71	5.95	11.05

Note. **p < 0.01

Interestingly, in the case of ALA-Reader, one and three cases both of which are expert' responses, deviated from the range of the concept and relation, respectively (see Figure 3). This result implied that ALA-Reader is more sensitive to the assessment context in terms of shared terms.

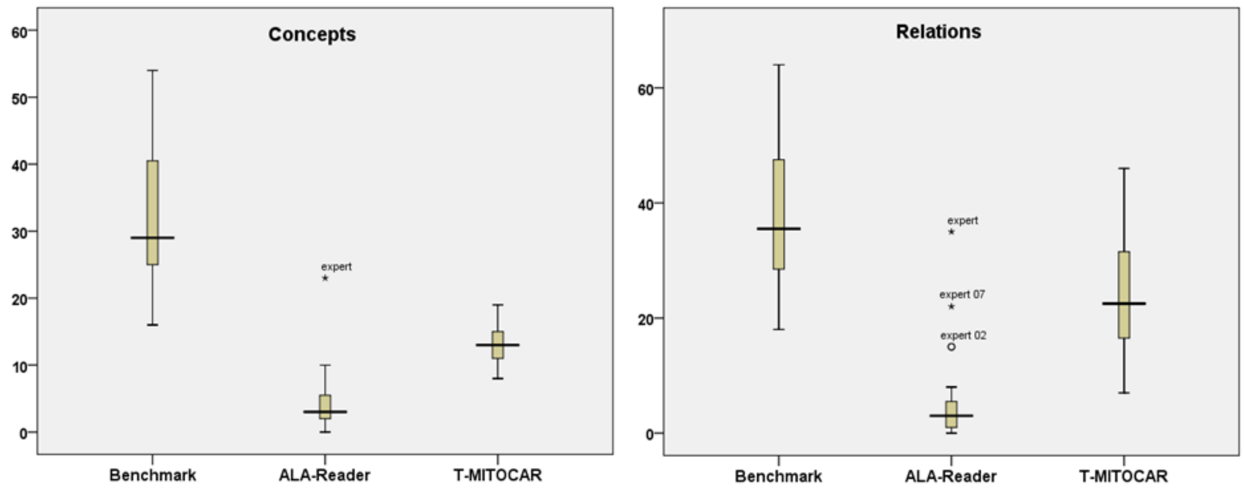

Figure 3. Boxplots of the numbers of concepts and relations

As can be seen, the numerical similarities of concepts and relations revealed that ALA-Reader and T-MITOCAR provided less information in terms of the number of concepts and relations (see Table 9).

Table 9

Similarities of the Numbers of Concepts and Relations of ALA-Reader or T-MITOCAR with those of the Benchmark

	ALA-Reader		T-MITOCAR	
	Concept	Relation	Concept	Relation
Min	0.000	0.000	0.200	0.179
Max	0.426	0.547	0.773	0.969
Average	0.126	0.117	0.432	0.584

Note. Sample size N = 28.

To examine the associations among the methods, correlation-coefficient analyses were conducted using word count, concepts, and relations. The benchmark and ALA-Reader had high correlations in their concepts and relations, r = 0.730 and 0.746, p < .01, respectively (see Table 10). In spite of a large difference in the numbers of concepts and relations, the ALA-Reader model was highly associated with the benchmark.

Table 10

Correlations among Concepts and Relations in the Benchmark, ALA-Reader, and T-MITOCAR and Word Count of Each Sample

	WC	BC	BR	AC	AR	TC	TR
WC	-						
BC	0.526**	-					
BR	0.572**	0.949**	-				
AC	0.224	0.730**	0.735**	-			
AR	0.365	0.701**	0.746**	0.966**	-		
TC	-0.044	0.174	0.206	0.099	0.051	-	
TR	-0.014	0.033	0.130	-0.033	-0.009	0.876**	-

Note. Sample size *N* = 28. ** *p* < .01.
WC (Word Count); BC (Benchmark Concept); BA (Benchmark Relation); AC (ALA-Reader Concept); AR (ALA-Reader Relation); TC (T-MITOCAR Concept); and TR (T-MITOCAR Relation).

It was assumed that the larger volume of text response in general represents more concepts and relations. A reliable tool should be sensitive to the volume of text representation. Although ALA-Reader and T-MITOCAR limit the number of concepts to no more than 30, it was expected that the numbers of distilled concepts and relations of the tools are to some extent associated with the volume of words in responses. Therefore, correlations with word count were investigated. The results showed that only the benchmark model has associations with word count in terms of concepts and relations, $r = 0.526$ and 0.572, $p < .01$, respectively (see Table 10).

Table 11

Simple Regression Analyses Investigating Linear Associations between Word Count and Concept

	Benchmark			ALA-Reader			T-MITOCAR		
	B	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Word Count	4.77	1.52	.53*	4.55	3.89	.22	-1.27	5.64	-.04
R^2		.25			.05			.00	
F		9.92*			1.37			.051	

* *p* < .05. ** *p* < .01.

Table 12

Simple Regression Analyses Investigating Linear Associations between Word Count and Relation

	Benchmark			ALA-Reader			T-MITOCAR		
	B	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Word Count	4.20	1.18	.57*	4.36	2.18	.37	-.13	1.75	-.01
R^2		.30			.13			.00	
F		12.62**			3.99			.00	

* *p* < .05. ** *p* < .01.

Subsequent regression analyses demonstrated that word count explained a significant proportion of the variance in the number of concepts and relations in the benchmark model, $R^2 = 0.25$, $F(1, 26) = 9.92$, $p < .01$ and $R^2 = 0.30$, $F(1, 26) = 12.62$, $p < .01$, respectively (see Tables 11 and 12).

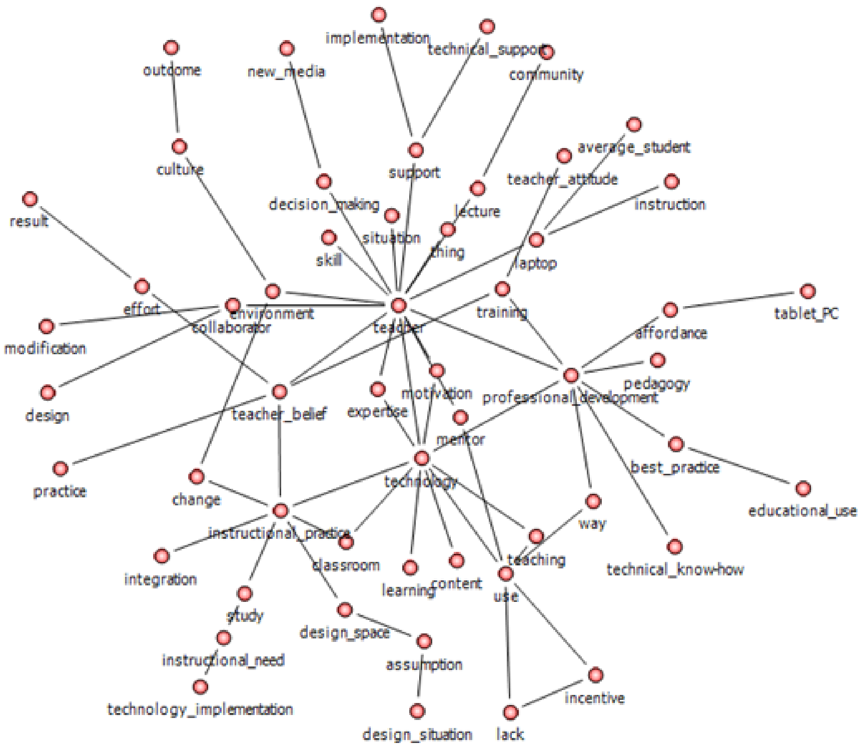

Figure 4a. Concept maps of the reference model created by the benchmark

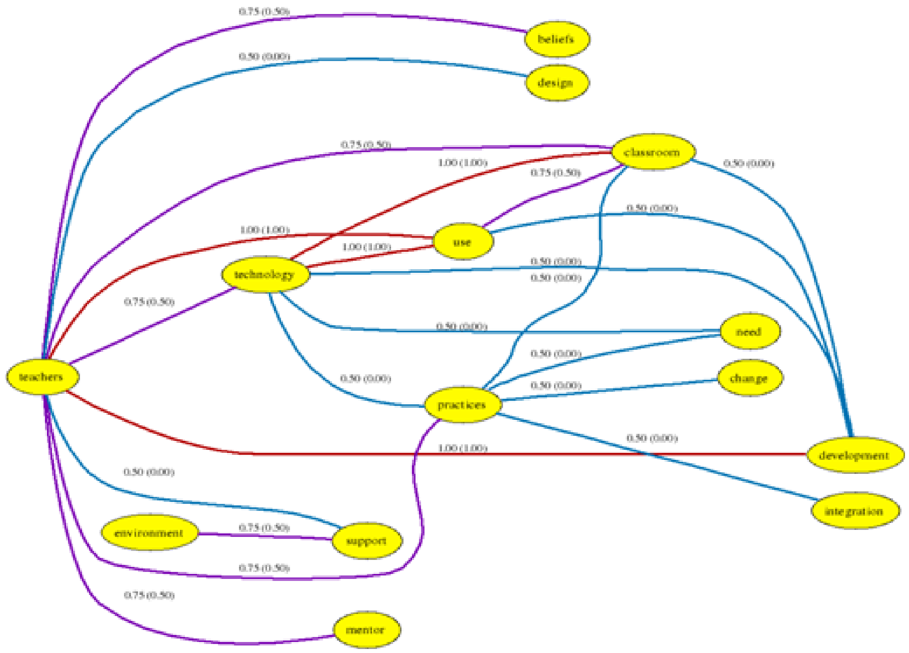

Figure 4b. Concept maps of the reference model created by T-MITOCAR

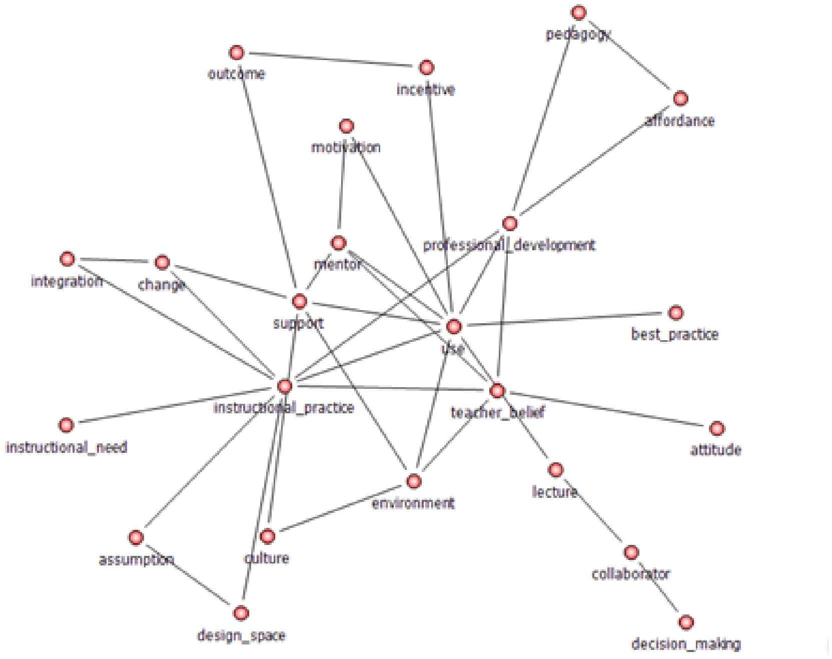

Figure 4c. Concept maps of the reference model created by ALA-Reader

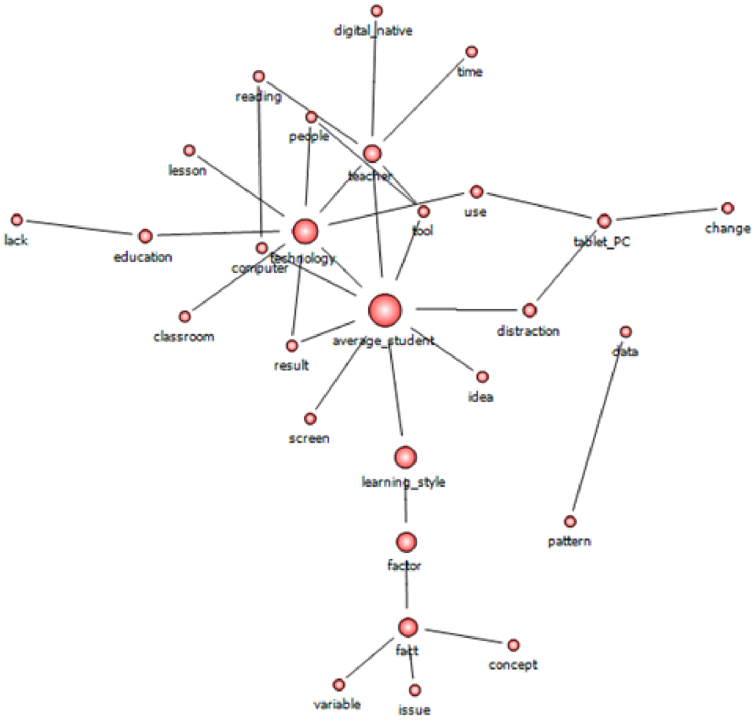

Figure 5a. Concept maps of the student 83 created by the benchmark

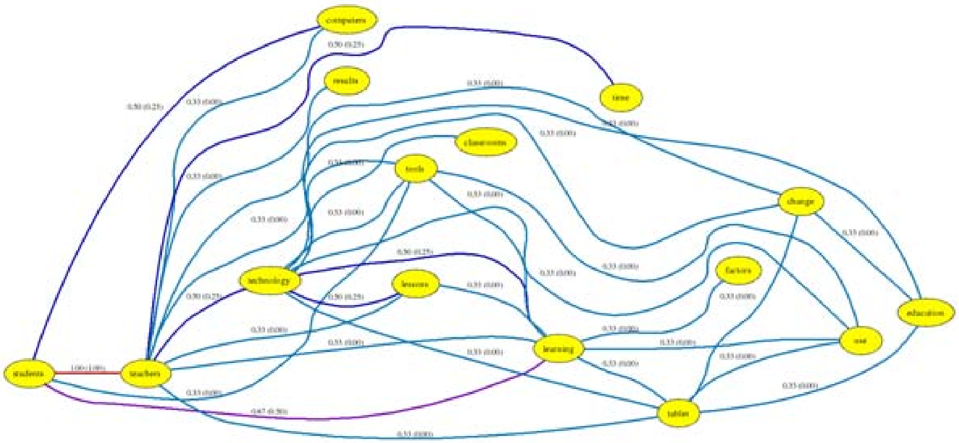

Figure 5b. Concept maps of the student 83 created by T-MITOCAR

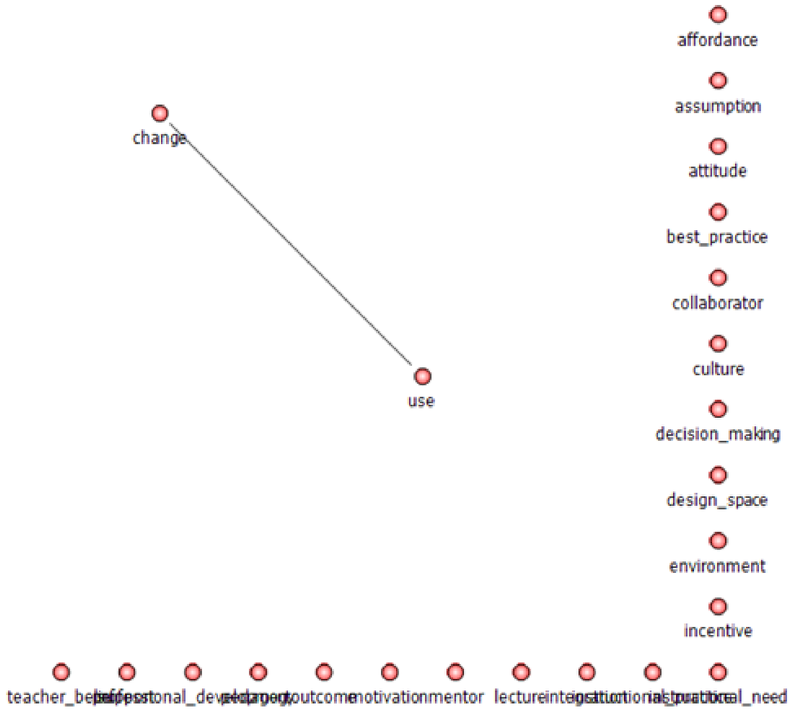

Figure 5c. Concept maps of the student 83 created by ALA-Reader

5.3 2nd review: the concept maps

For the reference model and student 83, concept maps drawn from the competing methods were visually investigated. Regarding the reference, although the benchmark created a more cohesive and informative concept map (see Figure 4a), for two reasons all other concept maps were highly connected (see Figure 4b and 4c): (a) the reference response was written carefully in a cohesive manner; and (b) T-MITOCAR and ALA-Reader let all elements of the concept maps connect technically, assuming all concepts are linked in the mind.

In contrast, the concept maps of student 83 substantially differed from one another. The benchmark differentiated the reference model from the concept map of student 83 (see Figure 5a). In contrast, T-MITOCAR yielded a concept map (see Figure 5b) more complex than that of the reference model, while the concept map produced by ALA-Reader was simple, capturing only two concepts (see Figure 5c).

5.4 3rd review: Concept map parameters

Another numerical similarity and correlation study was performed regarding the structural parameters of concept maps: (a) diameter; (b) gamma (cluster coefficient); and (c) cohesive subgroups. In reference to Table 13, the benchmark model produced the largest and most complex structure (average diameter = 7.04 and subgroup = 14.82). The T-MICOCAR model had an average diameter and subgroup value similar to those of the benchmark (average diameter = 6.00 and subgroup = 13.07). In contrast, the highest average gamma was identified in the ALA-Reader model, which was probably affected by the smaller structure (gamma = 0.34 and subgroup = 1.61). According to Table 14, parameter values of the T-MITOCAR model were much closer to those of the benchmark than those of the ALA-Reader model.

Table 13

Descriptive Statistics of Structural Parameters of Concept Maps

	Min	Max	Mean	SD
Benchmark				
Gamma	0.00	0.32	0.18	0.09
Diameter	5.00	12.00	7.04	1.60
Subgroup	6.00	29.00	14.82	5.70
ALA-Reader				
Gamma	0.00	1.0	0.34	0.43
Diameter	0.00	7.00	2.04	0.10
Subgroup	0.00	10.00	1.61	2.20
T-MITOCAR				
Gamma	0.12	0.59	0.25	0.10
Diameter	3.00	9.00	6.00	1.44
Subgroup	8.00	19.00	13.07	3.21

Note. Sample size *N* = 28.

Table 14

Numerical Similarities of Structural Parameters with those of the Benchmark Model

	ALA-Reader			T-MITOCAR		
	Min	Max	Average	Min	Max	Average
Gamma	0.00	0.80	0.15	0.00	0.98	0.58
Diameter	0.00	0.86	0.28	0.33	1.00	0.81
Subgroups	0.00	0.39	0.09	0.50	1.00	0.73

Note. Sample size *N* = 28.

Table 15

Correlations among Structural Parameters of the Benchmark, ALA-Reader, and T-MITOCAR Including Word Count

	WC	BG	BM	BC	AG	AM	AC	TG	TM	TS
WC	-									
BG	.020	-								
BM	-.026	-.300	-							
BC	.363	-.126	.648**	-						
AG	.386*	.442*	-.215	.065	-					
AM	.054	-.079	.493**	.626**	-.026	-				
AC	.076	.005	.331	.630**	.092	.864**	-			
TG	.003	-.041	.078	-.034	.147	-.020	-.022	-		
TM	.074	-.029	.209	.334	-.165	.063	.094	-.623**	-	
TS	-.044	-.317	.375*	.410*	-.397*	.210	.072	-.445*	.617**	-

Note. Sample size *N* = 28. * *p* < .05. ** *p* < .01.
WC (Word Count); BG (Benchmark Gamma); BM (Benchmark Diameter); BC (Benchmark Cohesive subgroups); AG (ALA-Reader Gamma); AM (ALA-Reader Diameter); AC (ALA-Reader Cohesive subgroups); TG (T-MITOCAR Gamma); TM (T-MITOCAR Diameter); and TS (T-MITOCAR Structure Measure).

Correlation analyses were conducted (see Table 15). In contrast to concepts and relations, most structural parameters had no relation with the word count. This result implied that the structure of a concept map has features distinctive from the frequencies of concepts and relations. The gamma score was in a negative relationship with the diameter and subgroup scores across all models. The gamma score did not appear simple to interpret in an assessment situation. The diameter and subgroup parameters were closely correlated across the benchmark, ALA-Reader, and T-MITOCAR.

Similar to concepts and relations, in spite of the ALA-Reader model's lower numerical similarity values, its structural parameters correlated more highly with the benchmark than did those of the T-MITOCAR's. For example, BC was correlated with AC, *r* = 0.630, p < .01, while the r-value of TS was 0.419, with a *p* < .01.

Table 16

Descriptive Statistics of Similarity Measures of Concept Maps

	Min	Max	Mean	SD
Benchmark				
Surface	0.281	0.984	0.590	0.184
Graphical	0.714	1.000	0.921	0.107
Gamma	0.000	0.942	0.644	0.245
Concept	0.119	0.570	0.284	0.100
Proposition	0.000	0.277	0.073	0.067
Balance	0.000	0.529	0.222	0.151
ALA-Reader				
Surface	0.029	0.629	0.142	0.151
Graphical	0.167	1.000	0.347	0.208
Gamma	0.000	0.777	0.234	0.286
Concept	0.087	0.435	0.190	0.112
Proposition	0.000	0.425	0.127	0.116
Balance	0.000	1.292	0.635	0.466
T-MITOCAR				
Surface	0.318	1.000	0.707	0.188
Graphical	0.500	1.000	0.835	0.138
Gamma	0.440	0.965	0.751	0.155
Concept	0.000	0.519	0.197	0.118
Proposition	0.000	0.360	0.061	0.087
Balance	0.000	0.697	0.223	0.214

Note. Sample size *N* = 28.

5.5 4th review: Concept map scores (similarity measures)

The final review was of six concept map scores obtained by measuring similarities between the student model and the reference model. Table 16 summarized the descriptive statistics of six measures. The average surface, graphical, and gamma scores were above 0.5 in the benchmark and T-MITOCAR, while the other three (concept, proposition, and balance) had low similarities, ranging from 0.061 to 0.284. That is, overall, the samples were above the half levels of the reference in terms of surface, graphical, and gamma scores but were not in concept, proposition, and balance scores. All scores of ALA-Reader were very low except for the balance measure (*m* = 0.635). Those scores resulted from the very small concept map sizes, in particular many of the student concept maps, affected by the constraint on analyzed concepts.

Table 17

Numerical Similarities of the Similarity Measures between the Benchmark and ALA-Reader or T-MITOCAR

	ALA-Reader			T-MITOCAR		
	Min	Max	Average	Min	Max	Average
Surface	0.004	0.639	0.018	0.345	0.973	0.721
Graphical	0.009	1.000	0.332	0.500	1.000	0.840
Gamma	0.000	0.967	0.293	0.000	0.994	0.738
Concept	0.003	0.991	0.571	0.000	0.993	0.626
Proposition	0.000	0.903	0.286	0.000	0.900	0.409
Balance	0.000	0.937	0.218	0.000	0.990	0.443

Note. Sample size *N* = 28.

Similar to the earlier investigations, the scores of T-MITOCAR had a high numerical similarity of at least more than 0.7 with those of the benchmark in terms of surface, graphical, and gamma (see Table 17). As for the concept, proposition, and balance scores, their numerical similarities were moderate, ranging from 0.409 and 0.626. In contrast, the scores of ALA-Reader had a low similarity, ranging from 0.018 to 0.332, with the exception of concept, which was 0.571. When the ranges of distributions were reviewed, the ALA-Reader and T-MITOCAR models appeared similar. A majority of the samples yielded relatively small-sized concept maps when processed via the ALA-Reader.

Table 18

Ranks of the seven experts in the Benchmark Data

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7
Surface	8 (5, 8)	4 (2, 4)	5 (2, 1)	18 (7,20)	3 (6,21)	2 (4,16)	1 (1,2)
Graphical	1 (1, 1)	1 (3, 8)	1 (3, 8)	1 (3,26)	1 (3,21)	1 (2,21)	1 (3,1)
Gamma	21 (12, 9)	3 (3, 23)	17 (2, 6)	8 (12,16)	25(12,17)	6 (1,13)	11 (4,10)
Concept	3 (4, 17)	2 (1, 6)	1 (3, 1)	8 (7,14)	4 (6,12)	6 (5,21)	4 (1,1)
Proposition	5 (11, 10)	7 (3, 6)	1 (2, 1)	2 (4,8)	2 (19,16)	6 (12,18)	2 (1,2)
Balance	5 (18, 8)	11 (11, 5)	2 (9, 2)	1 (7,6)	3 (19,16)	7 (17,18)	3 (8,3)

Note. Sample size *N* = 28. In the parentheses, the first is the rank in the ALA-Reader data and the second is the rank in the T-MITOCAR's.

To investigate the measurement accuracy of the three approaches, all 27 samples were ranked according to a single concept map score for the individual samples. Expert responses were expected to be ranked at the top. Overall, in the benchmark model, expert responses fell into the upper ranks of the list (see Table 18). The ranks of ALA-Reader and T-MITOCAR varied but ALA-Reader provided rankings closer to those of the benchmark than T-MITOCAR. There was no pattern in the gamma ranks of all three models.

Despite the lack of numerical similarity, the surface and graphical scores had a significant correlation only between the benchmark and ALA-Reader, *r* = 0.724 and 0.412, *p* < .05 (see Table 19). The correlations of concept and propositional scores were significant overall across the three approaches. The balance score had only a moderate association between the benchmark and T-MITOCAR, *r* = 0.489, *p* < .01. On the whole, ALA-Reader generated concept map scores better associated with those of the benchmark than T-MITOCAR, while T-MITOCAR had a better association in terms of proposition and balance scores (see Table 19).

Table 19

Correlations of the Concept Map Scores between the Benchmark and ALA-Reader or between the Benchmark and T-MITOCAR

	ALA-Reader	T-MITOCAR	ALA-Reader and T-MITOCAR[a]
Benchmark			
Surface	.724**	.093	.416*
Graphical	.412*	-.222	.057
Gamma	.277	-.145	-.096
Concept	.815**	.696**	.520**
Proposition	.555**	.654**	.634**
Balance	-.048	.489**	.025

Note. Sample size *N* = 28. * *p* < .05. ** *p* < .01. Pearson *r* was applied to the similarity measures
a. The correlations of each measure between ALA-Reader and T-MITOCAR.

Table 20

Correlations of the Similarity Measures of Concept Maps in the Benchmark

	1	2	3	4	5	6
1. Surface	-					
2. Graphical	.338	-				
3. Gamma	-.233	-.278	-			
4. Concept	.583**	.093	-.034	-		
5. Proposition	.512*	.272	-.013	.870**	-	
6. Balance	.400*	.264	-.027	.723**	.924**	-

Note. Sample size *N* = 28. * *p* < .05. ** *p* < .01. Pearson Correlation Coefficients in the lower diagonal.

Lastly, in regard to the benchmark, correlations of its concept map scores were examined (see Table 20). The surface score was moderately correlated with the concept and proposition score, *r* = 0.583 and 0.512, p < .05, respectively. In contrast, no significant correlation was identified with structural parameters such as the graphical and gamma scores. The balance score was highly associated with the propositional score, $r = 0.924$, $p < .01$. Those results demonstrated that concept map scores account for different features of concept maps.

6. Conclusion

6.1 Research findings

This study assumed that an individual student's understanding is meaningfully elicited via a natural language approach. Two state-of-the-art technologies, ALA-Reader and T-MITOCAR, were selected because they were consistent with the initial assumption. In order to validate the technologies, an alternative method was established as a benchmark.

It was believed that linguistic knowledge representation should be open-ended in terms of concepts and should be directional in terms of relations. In addition, it was assumed that editing pronouns in text responses helps obtain more accurate and descriptive concept maps. Creating a benchmark drew on distilling semantic relations from responses because eliciting linguistic semantic structures was assumed to be a better way to visually represent concept maps.

The noun-only data was not significantly different from the pronoun-edited data, which was aligned with Clariana and colleagues' [17,48] suggestion. In addition, there was little difference between the directional and non-directional approaches Thus, it was concluded that a simple noun-only and non-directional approach was sufficient for creating a benchmark.

Two findings are summarized through the numerical similarity and correlation analyses amongst the benchmark, ALA-Reader, and T-MITOCAR:

- The benchmark model created the most descriptive concept maps in terms of the number of concepts and relations, followed by T-MITOCAR and ALA-Reader.

- The ALA-Reader model had smaller numerical similarities, although it yielded higher correlations with the benchmark model than the T-MITOCAR model in terms of proximity data, concept map parameters, and concept map scores.

Those results are probably affected by the constraints introduced by each technology. ALA-Reader and T-MITOCAR can only analyze at most 30 concepts in a concept map. Moreover, ALA-Reader pre-defines a particular set of terms to be used in text analyses. Accordingly, ALA-Reader yielded the smallest concept maps. The higher correlation between the ALA-Reader and benchmark models is in part explained by their methodological similarity. That is, both use adjacency rather than distance data.

When considering the sharp drop in the numbers of concepts in student samples, ALA-Reader will be more appropriately applied in a setting in which a set of key terms are intentionally introduced or adequately exposed before and during assessment activities, which may yield a closer association with the benchmark model than occurred in this study. T-MITOCAR is only applicable to cases having more than 350 words. Accordingly, the technology seems on the whole adaptable to data reduction utilizing a large volume of text. Although T-MITOCAR yielded data numerically more similar to those of the benchmark, their associations with the benchmark were somewhat lower than those of ALA-Reader. This result requires further study in regard to the association between adjacency data (the benchmark) and distance data (T-MITOCAR). Table 21 summarizes the characteristics of the three concept map analysis technologies.

Table 21

Summary of Concept Mapping Technologies

	The Benchmark	T-MITOCAR	ALA-Reader
Approach	Open-ended (No constraint on words in a response and concepts in a concept map)	Partially Open-ended (No constraint on words but requiring more than 350 words in a response; at most 30 concepts analyzed in a concept map)	Close-ended (No constraint on words in a response but only predefined concepts-up to 30-analyzed in a concept map)
Complexity	Complex	Medium	Simple
Relation Judgment	Semantic	Distance	Adjacency
Data (Matrix)	Adjacency data	Distance data	Adjacency data
Direction of Relation	Directional/Non-directional	Non-directional only	Non-directional only
Concept Map	Descriptive/Natural	Abstractive/ Constrained	Abstractive/ Constrained
Targeted assessment tasks	Ill-structured complex problem	Ill-structured complex problem	Structured complex problem (in the sense that key concepts are introduced before an assessment)
Instruction before assessment	Not necessary	Not necessary	Desired (Students are expected to learn key terms through instruction on a task)
Key advantage	Descriptive/naturalistic concept maps provide formative information for individualized instructional support.	It is beneficial to obtain key concepts and relations from a larger volume of text such as essay analysis.	Accurate diagnosis is expected, providing a set of key terms are introduced.

6.2 New opportunities

The benchmark model was explored initially, and concept maps were processed manually. In spite of those limitations, the benchmark model provided some opportunities: First, its concept maps were much more descriptive than those of the other two models (see Table 21). When the purpose of assessment is to provide formative feedback and instructional support, descriptive information of students' status is essential.

Second, the benchmark model was able to more capably distinguish better concept maps from maps of lesser quality. For example, expert responses were accurately identified and ranked at the top, which to a large degree resulted from the modified similarity formula suggested in this study.

Third, the benchmark model has no constraints on the number of words and used terms. Thus, this approach can deal with diverse assessment contexts. For example, when gathering data, the instructor clearly asked the students to write responses of more than 350 words. However, only one third of the students met the requirement. It is natural that diverse levels of students in a classroom provide diverse volume of responses. Accordingly, an assessment technology should cover all responses.

6.3 Suggestions

In regard to further development of the benchmark model, there are two suggestions for future studies: First, the methods for drawing semantic relations from written responses need to be more specific and algorithmic. Second, a set of combined rules is required to assess the progress of student learning based on multiple concept map scores. For example, the six concept map scores were not often at the same level. They were not correlated in some of the pairs because concept map scores account for the different features of concept maps. Lastly, developing an automated assessment technology embedding the benchmark model is required. That technology would enable a teacher to have a better sense of students' learning and provide them with elaborated feedback and support.

References

[1] R.B. Clariana, Multi-decision approaches for eliciting knowledge structure, in: Computer-Based Diagnostics and Systematic Analysis of Knowledge, 2010: pp. 41-59.

[2] V.K. Narayanan, Causal mapping: An historical overview, in: V.K. Narayanan, D.J. Armstrong (Eds.), Causal Mapping for Research in Information Technology, Idea Group Inc (IGI), Hershey, PA, 2005: pp. 1-19.

[3] J.D. Novak, A.J. Cañas, The origins of the concept mapping tool and the continuing evolution of the tool, (2006).

[4] J. Spector, T.A. Koszalka, The DEEP methodology for assessing learning in complex domains, Syracuse University, Syracuse, NY, 2004.

[5] P. Pirnay-Dummer, D. Ifenthaler, J. Spector, Highly integrated model assessment technology and tools, Educational Technology Research and Development. (2009).

[6] F. Dochy, M. Segers, P. Van den Bossche, D. Gijbels, Effects of problem-based learning: a meta-analysis, Learning and Instruction. 13 (2003) 533-568.

[7] D.H. Jonassen, K. Beissner, M. Yacci, Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge., Structural Knowledge: Techniques for Representing, Conveying, and Acquiring Structural Knowledge. (1993).

[8] A. Newell, H.A. Simon, Human problem solving, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[9] M.S.R. Segers, An alternative for assessing problem-solving skills: The overall test., Studies in Educational Evaluation. 23 (1997) 373-98.

[10] D. Gijbels, F. Dochy, P. Van den Bossche, M. Segers, Effects of problem-based learning: A meta-analysis from the angle of assessment, Review of Educational Research. 75 (2005) 27-61.

[11] R.M. Thomas, High-stakes testing: coping with collateral damage, Psychology Press, Mahwah, NJ, 2005.

[12] D. Ifenthaler, Relational, structural, and semantic analysis of graphical representations and concept maps, Educational Technology Research and Development. (2008).

[13] A.C. Jeong, jmap, (2008).

[14] D.L. O'Connor, T.E. Johnson, Measuring team cognition: Concept mapping elicitation as a means of constructing team shared mental models in an applied setting, in: Pamplona, Spain, 2004.

[15] R.B. Clariana, P. Wallace, A comparison of pair-wise, list-wise, and clustering approaches for eliciting structural knowledge in information systems courses, International Journal of Instructional Media. 36 (2009) 287-302.

[16] E.M. Taricani, R.B. Clariana, A technique for automatically scoring open-ended concept maps., Educational Technology Research & Development. 54 (2006) 65-82.

[17] R. Clariana, P. Wallace, V. Godshalk, Deriving and measuring group knowledge structure from essays: The effects of anaphoric reference, Educational Technology Research and Development. 57 (2009) 725-737.

[18] R.W. Schvaneveldt, ed., Pathfinder associative networks: studies in knowledge organization, Ablex Publishing Corp., Norwood, NJ, 1990.

[19] S. Kalyuga, Assessment of learners' organised knowledge structures in adaptive learning environments, Applied Cognitive Psychology. 20 (2006) 333-342.

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

[21] J.M. Spector, V.P. Dennen, T.A. Koszalka, Causal maps, mental models and assessing acquisition of expertise, Technology, Instruction, Cognition and Learning. 3 (2006) 167–183.

[22] P.N. Johnson-Laird, The history of mental models, in: K.I. Manktelow, M.C. Chung (Eds.), Psychology of Reasoning: Theoretical and Historical Perspectives, Psychology Press, New York, NY, 2005: pp. 179-212.

[23] P.N. Johnson-Laird, Mental models and thoughts, in: K.J. Holyoak (Ed.), The Cambridge Handbook of Thinking and Reasoning, Cambridge University Press, Cambridge, Mass., 2005: pp. 185-208.

[24] A. Newell, Unified theories of cognition, Harvard University Press, Cambridge, MA, 1994.

[25] N. Schlomske, P. Pirnay-Dummer, Model based assessment of learning dependent change during a two semester class, in: Kinshuk, D. Sampson, J.M. Spector (Eds.), IADIS International Conference Cognition and Exploratory Learning in Digital Age 2008, Freiburg, Germany, 2008: pp. 45-53.

[26] R.W. Schvaneveldt, F.T. Durso, T.E. Goldsmith, T.J. Breen, N.M. Cooke, R.G. Tucker, et al., Measuring the structure of expertise, International Journal of Man-Machine Studies. 23 (1985) 699-728.

[27] R. Axelrod, Structure of decision : the cognitive maps of political elites, Princeton University Press, Princeton N.J., 1976.

[28] K. Carley, M. Palmquist, Extracting, representing, and analyzing mental models, Social Forces. 70 (1992) 601-636.

[29] Quillian, Semantic memory, in: M.L. Minsky (Ed.), Semantic Information Processing, The MIT Press, Cambridge, MA, 1985: pp. 216–270.

[30] A.M. Collins, E.F. Loftus, A spreading-activation theory of semantic processing, Psychological Review. 82 (1975) 407-428.

[31] V.J. Shute, A.C. Jeong, J.M. Spector, N.M. Seel, T.E. Johnson, Model-based methods for assessment, learning, and instruction: Innovative educational technology at Florida State University, in: M. Orey (Ed.), Educational Media and Technology Yearbook, The Greenwood Publishing Group, New York, NY, 2009.

[32] V.J. Shute, D. Zapata-Rivera, Using an evidence-based approach to assess mental models, in: Understanding Models for Learning and Instruction, Springer, New York, NY, 2008: pp. 23-41.

[33] A.L. Brown, Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings, The Journal of the Learning Sciences. 2 (1992) 141-178.

[34] N.M. Seel, Epistemology, situated cognition, and mental models: “Like a bridge over troubled water,” Instructional Science. 29 (2001) 403-427.

[35] J.G. Greeno, Situations, mental models, and generative knowledge, in: H.A. Simon, D. Klahr, K. Kotovsky (Eds.), Complex Information Processing: The Impact of Herbert A. Simon, Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ, 1989: pp. 285–318.

[36] A. Garnham, Mental models as represenhtions of discourse and text, Halsted Press, New York, 1987.

[37] A. Garnham, Mental models and the interpretation of anaphora, Taylor & Francis, Philadelphia, 2001.

[38] K.A. Coronges, A.W. Stacy, T.W. Valente, Structural comparison of cognitive associative networks in two populations, Journal of Applied Social Psychology. 37 (2007) 2097-2129.

[39] K.A. Hutchison, Is semantic priming due to association strength or feature overlap? A microanalytic review, Psychonomic Bulletin & Review. 10 (2003) 785-813.

[40] S. Wasserman, K. Faust, Social network analysis: methods and applications, Cambridge University Press, Cambridge, MA, 1994.

[41] A.A. Rupp, S. Sweet, Y. Choi, Modeling learning trajectories with epistemic network analysis: A simulation-based investigation of a novel analytic method for epistemic games, (2010).

[42] R. Schvaneveldt, F.T. Durso, D.W. Dearholt, Network structures in proximity data, in: G.H. Bower (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory, Academic Press, San Diego, CA, 1989: pp. 249-284.

[43] Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W., Evidence-centered design of epistemic games: Measurement principles for complex learning environments, The Journal of Technology, Learning, and Assessment. 8 (2010).

[44] D.W. Shaffer, D. Hatfield, G.N. Svarovsky, P. Nash, A. Nulty, E. Bagley, et al., Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning, International Journal of Learning and Media. 1 (2009) 33-53.

[45] M.B. Curtis, M.A. Davis, Assessing knowledge structure in accounting education: an application of Pathfinder Associative Networks, Journal of Accounting Education. 21 (2003) 185-195.

[46] J.R. McClure, B. Sonak, H.K. Suen, Concept Map Assessment of Classroom Learning: Reliability, Validity, and Logistical Practicality., Journal of Research in Science Teaching. 36 (1999) 475-92.

[47] M.A. Ruiz-Primo, S. Schultz, M. Li, R.J. Shavelson, On the cognitive validity of interpretations of scores from alternative concept mapping techniques, University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), 1999.

[48] R.B. Clariana, P. Wallace, A computer-based approach for deriving and measuring individual and team knowledge structure from essay questions, Journal of Educational Computing Research. 37 (2007) 211-227.

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

[50] D. Knoke, J.H. Kuklinski, Network analysis, Sage Publications, Newbury Park, 1982.

[51] P. Hage, F. Harary, Structural models in anthropology, Cambridge University Press, Cambridge, MA, 1983.

[52] D. Ifenthaler, I. Masduki, N.M. Seel, The mystery of cognitive structure and how we can detect it: tracking the development of cognitive structures over time, Instr Sci. (2009).

[53] T.E. Goldsmith, K. Kraiger, Applications of structural knowledge assessment to training evaluation, in: J.K. Ford, S.W.J. Kozlowski (Eds.), Improving Training Effectiveness in Work Organizations, Psychology Press, Mahwah, NJ, 1997: pp. 73-95.

[54] P.R. Monge, N.S. Contractor, Theories of communication networks, Oxford University Press, New York, NY, 2003.

[55] P. Pirnay-Dummer, D. Ifenthaler, Automated knowledge visualization and assessment, in: D. Ifenthaler, N.M. Seel, P. Pirnay-Dummer (Eds.), Computer-Based Diagnostics and Systematic Analysis of Knowledge, Springer, New York, NY, 2010: pp. 77-116.

[56] C.M. Goodman, The Delphi technique: a critique, J Adv Nurs. 12 (1987) 729-734.

[57] C. Hsu, B.A. Sandford, The Delphi Technique: Making Sense of Consensus, Practical Assessment Research & Evaluation. 12 (2007).

[58] C. Okoli, S.D. Pawlowski, The Delphi method as a research tool: an example, design considerations and applications, Information & Management. 42 (2004) 15-29.

[59] M.T.H. Chi, R. Glaser, M.J. Farr, The nature of expertise, L. Erlbaum Associates, Hillsdale, NJ, 1988.

[60] A.M. Colman, E. Shafir, Tversky, Amos, in: N. Koertge (Ed.), New Dictionary of Scientific Biography, Charles Scribner’s Sons, Farmington Hills, MI, 2008: pp. 91-97.

[61] A. Tversky, E. Shafir, Preference, Belief, and Similarity: Selected Writings, 1st ed., The MIT Press, London, England, 2003.

[62] A. Tversky, Features of similarity, Psychological Review. 84 (1977) 327-352.