Text Summaries or Concept Maps: Which Better Represent Reading Text Conceptualization?

TRISTAN E. JOHNSON*1, 2, PABLO N. PIRNAY-DUMMER3, DIRK IFENTHALER3, ANNE MENDENHALL1, SELÇUK KARAMAN4 AND GERSHON TENENBAUM2

Learning Systems Institute1 and the Department of Educational Psychology and Learning Systems, Florida State University, USA, Department of Educational Science, University of Mannheim, Germany3 and Ataturk University, Turkey4

PO Box 3062540, Tallahassee, Florida 32306-2540
Phone: (850) 644-8770, Fax: (850) 644-4952

Capturing students’ mental models has been proposed as a viable means to measure students’ understanding and conceptualization of given learning materials. Mental models are usually represented by either short text summary or a graphical map (i.e., concept map). This study aimed at testing which learner’s mental representation associates higher with three criteria: original text, expert concept map, and expert text summary. HIMATT, a mathematical framework proven to share a sound reliability of mental model in both semantic and graphical formats, was used to elicit the association between students’ mental models and the three criteria reference models following studying two book chapters. The findings indicate that across all association indices, students’ text summary elicitations were stronger than the students’ concept map elicitations over the three criterion-reference models. Moreover, stronger similarities emerged for the four structure indices (surface, graphic, structure, and gamma) than for the three semantic indices (concept, proposition, and balance) within the text summary and concept map formats. The results are attributed to students’ strong familiarity with written representation of the learning materials rather than creating concept maps. Furthermore, the results indicate that reading a text stored in long-term memory and retrieving and representing it into a concept map is harder than retrieving and representing it in a written text. Further research must clarify to what extent the practice of transforming written materials into graphical maps will improve the validity of using concept maps for mental conceptualization.

Keywords: mental models, elicitations, concept maps, text summaries, semantic measures, graphical measures, text conceptualization, reference models, HIMATT, convergent validity

*Corresponding author: tejohnson@fsu.edu
This research was supported by NSF grant #SES 0624110. The content does not necessarily represent the policy of the NSF and there should be no assumption of endorsement by the Federal Government.
Much effort was devoted to the development of a theoretical foundation of mental models (e.g., Gentner & Stevens, 1983; Johnson-Laird, 1989; Norman, 1983; Spector, Dennen, & Koszalka, 2006), and their instructional application (Anzai & Yokoyama, 1984; Mayer, 1989; Seel, 1995, 2003). However, there are still a number of concerns as to their validity (Ifenthaler, 2008; Ifenthaler & Seel, 2005; Seel, 1999), i.e., which form of expression (visual or contextual - descriptive) better represents what one comprehends from a learning text.

One essential question concerning the assessment of knowledge and mental models is which methodology should be used, one that uses visual representation (i.e., concept map) or one that consists of a written text (i.e., a summary). Many authors consider concept maps to be an adequate format of externalization for analyzing complex knowledge structures (Johnson, Ifenthaler, Pirnay-Dummer, & Spector, 2009; Novak, 1998). Concept maps seem preferable to classical knowledge tests, such as multiple-choice tests for the purpose of representing linked knowledge by means of network-like visualization. On the other hand, there are strong arguments indicating that natural language representations are a good method for assessing knowledge and mental models (Pirnay-Dummer, Ifenthaler, & Rohde, 2009).

The main purpose of the current study is to test the convergent validity for text summaries and concept maps using reference representations and book chapters. Specifically, we compare the similarities and differences of the graphically and textually elicited knowledge structures and their semantics during a graduate level systems analysis course.

**ELICITATION OF INTERNAL REPRESENTATIONS**

A representation is internal to the mind, and for obvious reasons not directly observable (Fodor, 1987, 2003; Ifenthaler, 2010b; Jonnassen & Cho, 2008; Pinker, 1994; Pirnay-Dummer, 2006; Seel, 1991). Representations are widely viewed as having a language-like syntax, and a compositional semantic (see Carruthers, 2000; Fodor, 2003; Margolis & Laurence, 1999; Pinker, 1994; Strasser, 2010). A mental model (in the sense of internal representations) is a representation of a thing, ideas or more generally, an ideational framework. It relies on language and uses symbolic pieces and processes of knowledge to construct a heuristic for a situation, which is instantiated by the world, or an internal process resembling the world, (e.g., a mental simulation) (see Johnson-Laird, 1983; Schnotz & Preuss, 1997; Seel, 1991). Its purpose is heuristic reasoning,
which leads either to intention, planning, behavior, or to a reconstruction of cognitive processes (see Piaget, 1976).

Mental models are fundamental to problem solving, including complex ones (Ceci & Ruiz, 1992; Jonassen, 2000; Just & Carpenter, 1976; Spector, 2006). The construct of a mental model is itself complex. There is no direct correlate in behavior, which represents the structure and operation of mental models; therefore, the assessment of a mental model is methodologically challenging as it involves all the metaphors of the mind (e.g., memory, attention, and motivation). The assessment of mental models necessitates multiple operationalizations for assumptions making, and predictions' steadiness (Jonassen & Cho, 2008).

A correlate of the internal representation is an external representation, which is constructed to support learning (Ifenthaler, 2010b) and assessment (Ifenthaler, 2008, 2010a; Johnson et al., 2009; Pirnay-Dummer, 2006, Pirnay-Dummer, Ifenthaler & Spector, 2010). The constructs of external representations are essentially representations of representations. The first representation mapping is a not-yet specified set of functions derived from the world to the mind. The second mapping is an additional set of functions used for the transition of the representations from the mind to an observable object in the world. Both functions certainly contain heuristics, and may be idiosyncratic to a not-yet specified extent. Thus, both transitions require interpretation, and interpretation requires a language structure (Montague, 1974). How the interpretation is carried out, and which residual artifact it produces (both ways) depends on the method that is used. Thus, there is a need to look at the construct from different perspectives. Using different human interpretation modules may help to solve a good part of this methodological problem, and shed more light on their interrelation (e.g., their convergent validity). HIMATT (Highly Integrated Model Assessment Technology & Tools) was used in this study to satisfy this need.

HIMATT is a web-based application, which provides an experimental framework for different perspectives on external representation. All HIMATT modules are implemented to run on a web server with a browser-based user interface that conducts the analysis, the comparisons, as well as generates the graphical output, and has a research management interface to layout experiments. It includes two different methods for assessing external representations and several functions for analysis and comparison (e.g., to monitor the progress of regular learners compared to experts). Pirnay-Dummer et al. (2010) provided a detailed overview of the methodology. The assessment strategy is based on T-MITOCAR (Text Model Inspection Trace of Concepts and Relations). Any text of a certain length can be graphically visualized by the T-MITOCAR software (Pirnay-Dummer,
T-MITOCAR tracks the association of concepts from a text directly to a graph, using an operational heuristic. Closer relations tend to be presented more closely within a text. This does not necessarily work within single sentences since syntax is more expressive and complex. However, text that contains 350 or more words, can be used to generate associative networks as graphs from text, and the heuristic becomes very stable at this point. The benefits and boundaries of the association net approach can be seen from linguistic and psycholinguistic research on syntax (e.g., Brill, 1995; Frazier, 1999; Just & Carpenter, 1976; Mitchell, 1994) and formal semantics (Helbig, 2006; Jackendoff, 1983; Fodor, 1987; Link, 1979, 2002; Montague, 1974; Pollio, 1966; Turner, 1983; Waldo, 1979). The external representation process is carried out automatically in multiple computer linguistic stages. The output graph contains the associations and their weights, which can be used for further analysis. The output graph is an association network rather than a concept map. However, it shares the concepts and the propositional structure with concept maps. The links are associations rather than annotated functions of semantic hierarchy. However, semantic hierarchy can rarely be found in concept maps made by learners who are only briefly trained in how to construct them (see Almond, Steinberg, & Mislevy, 2002; Hoz, 2009).

**PURPOSE**

As new technologies are created to meet various methodological challenges, there is a need to establish their convergent validity by using comparison-type techniques allowing similarity measures to emerge from the use of both concept maps and text summaries. Given that there are multiple elicitation techniques to generate the external representations, the main question that we are interested in relates to the level of trustfulness when regular learners are compared to a reference model that is based on either the original text or an external representation of an expert concept map or an expert text summary.

Accordingly, the purpose of the study is to compare and cross-validate two external representations of students’ models with a specific reference model, such as an expert model. This procedure provides validity evidence using multi-methods triangulation approach. Moreover, estimates can be given to what extent the mode of the individual interpretation during the externalization process plays a part in the assessment outcomes. Hence, we assume that the similarities between the reference representations and the learner’s text summaries or concept maps do not differ.
METHOD

Participants
Participants included 21 graduate students enrolled in a graduate analysis course at a major university in the southeastern United States. The study took place as a graded activity. Five of the participants were male and 16 were female. The average age of the participants was 35.33 (SD = 10.35). The majority of the students had a background in instructional systems or human resource development. Eight participants had previous experience with Performance System Analysis. Participation in the construction of concept map and executive summary activities was required for course completion.

Task
As part of the study, students were asked to create an executive summary and a concept maps for two chapters readings. Reading materials were selected chapters from HPI Essentials (Piskurich, 2002), and Handbook of Human Performance Technology 3rd Edition (Pershing, 2006).

For the first task, participants read chapter one from HPI Essentials (Piskurich, 2002). The chapter titled, “What is HPI? What makes a Performance consultant? How can you tell if you already are one?” introduces readers to the role of human performance improvement (HPI) consultants. The chapter is about key properties and overall processes of HPI. In this context, the chapter illustrates basic concepts such as performance, behavior, and accomplishment. The chapter also includes the role of HPI as a systematic process of designing interventions to close performance gaps.

For the second task, participants read chapter two titled, “Business analysis: The driving force behind the HPI process.” This reading was derived from the “Handbook of human performance technology” (Pershing, 2006). The chapter covers the conceptual foundation and development of human performance technology in terms of how the disciplines and theories affected HPI.

Procedure
For each task, participants read a chapter and then wrote an executive summary and created a concept map of the chapter. An executive summary is a document that is approximately 250 – 1000 words describing, in the participants own words, the key points and strategies presented in the chapter. After completing their executive summary participants created a concept map. A concept map is a diagram that illustrates the relationship between variables.
Participants were asked to prepare text summaries and concept maps for both readings. The participants were familiar with the terms and requirements for writing their summaries; therefore, only brief instructions for the executive summary were provided. They were asked to write at least 250 words as a summary, and were notified not to use quotes from the chapter or plagiarize the content. Students were asked to write the summaries using their own words. Participants were less familiar with how to create concept maps and instructions were provided. Students were asked to practice creating a concept map to help familiarize themselves with the task. They were asked to use Cmap Software (http://cmap.ihmc.us/conceptmap.html) to create concept maps. Guideline and e-mail support were provided to help them install and use the tool. They were given three weeks to create their concept maps and text summaries.

Reference Models
For the reference models, three experts (course instructors) created concept maps and text summaries from the same readings. After reviewing both the maps and the summaries together, they modified the representations until they reached an agreement on what should be included in the concept map and the text summary for each reading.

Analysis

Data Manipulation
After the data was collected, the model similarity analyses were carried out using HIMATT. Each student model (concept map and text summary) was paired to each of the reference models, i.e., original text, expert concept map, and expert text summary (see Figure 1). Similarity analysis results of these pairs were compared statistically. There are 6 similarity pairs between the two students’ representations and three reference representations similarities (see Figure 1). For example, similarity set number one pairs the students’ concept map model with the original text reference model, while similarity of set number two pairs students’ text summary with the original text.

Calculation of Similarity Indices
Within HIMATT, the main aggregation method uses the T-MITOCAR algorithm to aggregate individual models to group models. This method is stable in aggregating text (Pirnay-Dummer & Ifenthaler, 2010). The main reason for the stability is the linear association measure of T-MITOCAR, which makes it possible to translate a group text (e.g., all the answers within a group to a given task)
into a single graph by treating the text source as one. Once assessed, HIMATT can compare model structures on a graph theoretical level. Over the years seven measures have shown a stable representation of different structural and semantic constructs. Although they can be subsumed either under a notion of semantics or structure, they measure different entities. Depending on an individual research questions, some of them may be left out. However, they do not substitute each other (i.e., they are not parallel). All measures are transformed to similarities. The similarity is either calculated on a frequency measure or on sets (e.g., sets of concepts or propositions). For any pair of frequencies $f_1$ and $f_2$, the similarity is derived by

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

which results in a measure of $0 \leq s \leq 1$, where $s = 0$ is complete exclusion, and $s = 1$ is identity. The other measures collect sets of properties from the graph (e.g. the vertices = concepts or the edges = relations). In this case, the Tversky (1977) similarity property applies for the given sets $A$ and $B$:

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

where $\alpha$ and $\beta$ are weights for the difference quantities which separate $A$ and $B$. This can be used to balance different aspects of the assessment (e.g., if the
condition under which an expert model was assessed does not for pragmatic reasons match the condition of the learners). Unless there is a methodological reason for such a balancing, they are usually equal ($\alpha = \beta = 0.5$). The four structural and two semantic measures are defined as follows (see Pirnay-Dummer et al., 2010): (1) the surface measure (see Ifenthaler, 2010a) - compares the number of vertices within two graphs. It is a simple and easy way to calculate values for surface complexity, (2) the graphical matching (see Ifenthaler, 2010a) - compares the diameters of the spanning trees of the graphs, and is an indicator for the range of conceptual knowledge. It corresponds with structural matching, as it is also a measure for structural complexity only, (3) the density of vertices measure (also often called “gamma”) (see Pirnay-Dummer, 2006) describes the quotient of terms per vertex within a graph. Since both graphs, (a) those which connect every term with each other term (everything with everything), and (b) graphs which only connect pairs of terms can be considered weak models, a medium density is expected for most good working models, (4) the structural matching measure (see Pirnay-Dummer, 2010) - compares the complete structures of two graphs without regard to their content. This measure is necessary for testing all hypotheses that make assumptions about general features of structure (e.g., assumptions which state that expert knowledge is structured differently from novice knowledge), (5) concept matching (see Pirnay-Dummer, 2006) - compares the sets of concepts (vertices) within a graph to determine the use of terms. It counts how many concepts are alike. This measure is especially important for different groups operating in the same domain (e.g. using the same textbook). It determines differences in language use between the models, (6) the propositional matching (see Ifenthaler, 2010a) - value compares only fully identical propositions (concept-link-concept) between two graphs. It is a measure for quantifying semantic similarity between two graphs, and (7) the balanced semantic matching uses both concepts and propositions to match the semantic potential between the knowledge representations (Pirnay-Dummer & Ifenthaler, 2010).

The individual measures usually correlate differently. There are significantly higher correlations within each classification (convergent, structure between $r = .48 - .79$ and semantics between $r = .68 - r = .91$), and lower correlations between them (divergent, between $r = -.24 - r = .36$). The density of vertices (gamma) usually stands alone, and only rarely correlates with the other structural measures because it accounts for a different feature of structure (correlations between $r = .37 - r = .38$). Pirnay-Dummer et al. (2010) provide a full validation study conducted on $N = 1,849,926$ model comparisons in 13 different subject domains ranging from common knowledge to scientific subject domains.
Statistical Analysis

The analysis used to construct-validate the mental representation methods (i.e., concept maps and/or text summaries) consists of four within-subjects (WS) factors: Chapter (1 and 2), students’ model representations (concept map vs. text summary), reference model representation (original text, expert concept map, and expert text summary), and index (surface, graphic, structure, gamma, concept, proposition, and balance). These factors were used in a Repeated Measures (RM) Multivariate Analysis of Variance (MANOVA) using similarities as the dependent variable (see Figure 2).

RESULTS

The RM MANOVA applied to the similarity value revealed that the chapter effect was non-significant ($p > .05$), thus the two-way and 3-way interaction effects are reported across chapters. The second factor, students’ model representations (concept map vs. text summary) resulted in a significant effect, Wilks’ $\lambda = .31$, $F(1,14) = 31.14$, $p = .001$, $\eta^2 = .69$. Students’ text summaries had higher similarities ($M_S = .47$) across all reference models than students’ concept maps ($M_S = .36$). The reference model representation (original text, expert concept map, and expert text summary) effect resulted also as a significant factor, Wilks’ $\lambda = .05$, $F(2,13) = 131.27$, $p = .001$, $\eta^2 = .95$. On average, original text and expert text summary had higher similarities with students’ model representations than expert concept map ($M_S = .46$ and .44 VS $M_S = .34$, respectively).
Testing the 2-way effects indicated a students’ model representation by reference model representation significant effect, Wilks’ $\lambda = .08$, $F(2, 13) = 72.34$, $p = .001$, $\eta^2 = .92$. This effect is shown in Figure 3.

As noticed, when students’ concept maps are contrasted with reference model representations, the similarity values emerged with original text, expert concept map, and expert text summary were small to moderate (i.e., $M_s = .39$, 34, and .36, respectively). However, when students’ text summaries are contrasted with reference model representations, only the expert concept map was small to moderate ($M_s = .34$), while the original text and expert text summary resulted in moderate to high similarity values ($M_s = .54$ and .53, respectively).

Finally, the 3-way, students’ model representation by reference model representation by index resulted in a significant effect, Wilks’ $\lambda = .013$, $F(12, 3) = 18.29$, $p = .02$, $\eta^2 = .99$. This effect is shown in Figure 4.

The similarity values presented in Figure 4 revealed that similarity values of students’ model representation and reference model representations resulted in some similar and some different patterns comprising the 7 indices. Students’ concept maps shared moderate to strong ($M_s = .32 - .70$) similarities with the original text, expert concept map, and expert text summary only for the indices surface, graphic, structure, and gamma, but less with the index of concept ($M_s = .10 - .20$). Students’ concept maps shared very poor similarity with the index of proposition ($M_s = .00 - .01$).
and the index of balance ($M_s = .03 - .07$). Similarly to Students’ concept map, Students’ text summary shared a stronger similarity values with surface, graphic, structure, and gamma derived from the original text, expert concept map, and expert text summary ($M_s = .42 - .85$). However, students’ text summary shared some similarity with the index of concept ($M_s$ original text $= .35$, and $M_s$ expert text summary $= .32$), and balance ($M_s$ original text $= .37$, and $M_s$ expert text summary $= .33$), but not with the proposition index ($M_s = .00 - .14$).

**DISCUSSION**

The current study compared and cross-validated graphically (concept maps) and textually (text summary) elicited mental representations, and tested the student models’ similarities to specific reference models including the original text, expert concept map, and expert text summary. This aim stems from the need to clarify the adequacy of concept map (Johnson et al., 2009; Novak, 1998), and natural language (Pirnay-Dummer, Ifenthaler, & Rohde, 2009) as methods of eliciting mental conceptualization following the studying of two textbook chapters.

The analysis and comparisons of the externalized representations consisted of seven measures suggested by the HIMATT toolset (Pirnay-Dummer & Ifenthaler, 2010). The automated quantitative analysis generated structural and semantic measures. The structural indicators were surface, graphical, gamma, and structural matching. The semantic indicators were concept, propositional, and balanced
propositional matching. These indicators describe and track changes in students’ representations and compare the graphically and textually elicited knowledge. Two distinct chapters were used to elicit the association coefficients, but proved to have similar association coefficients, thus the findings generalize across the two chapters.

The findings indicate that across all the association indices students’ text summary elicitations were stronger than the student concept map elicitations over the three criterion reference models: original text, expert concept map, and expert text summary (see Figure 3). Furthermore, stronger similarities emerged for the four structural indices (surface, graphic, structure, and gamma) than for the three semantic indices (concept, proposition, and balance) (see Figure 4). Students’ text summaries were more similar in structure to their references criteria than student concept maps. Moreover, text summaries were more similar across all indices to the original text and the expert text summary, but much less similar to the expert concept map (see right side of Figure 4). Thus, students’ conceptualization in a text summary format can be used as a valid representation of conceptualization of a written text. A concept map representation is limited in scope for students who lack familiarity and knowledge to transfer written text into concept map format.

These findings are attributed to the students’ familiarity with writing, and their limited ability in establishing concept maps. Most students (~90%) testified that they lack experience in concept mapping, and found it a difficult challenge. For most of them the elicitations of concept maps were the first and second attempt-sever. Thus, it is assumed that students must be familiarized with the procedures of creating a conceptual map that represents a written text; otherwise, drawing conclusions from a graphical representation is questionable at best.

Several authors linked internal and external representations (see Ifenthaler, 2010b, Ifenthaler, 2008, 2010a; Johnson et al., 2009; Pirnay-Dummer, 2006, Pirnay-Dummer et al., 2010), and claimed that external representations are representations of representations. They further postulated two processes: one that represents the text in the mind (e.g., memory structure), and the other that transfers this representation into an observable object in the real world (e.g., retrieval route). The findings of the current study pertain to the second postulate that consists of specific heuristics not precisely specified. However, both transitions (graphical and natural language externalization) require interpretation, which in turn requires a language structure (Montague, 1974). Thus it is not surprising that the associations between student text summaries and the reference models were stronger than the associations between student concept maps and the reference models: Different languages are compared. Consistent with this interpretation, the results support the notion that stored information in a natural language code
is harder to be retrieved (externalized) in a mapping format than in a written format because of the higher congruence between the memory representation and the retrieval format. In other words, geometrically analogue information may be conveyed easier on a graph, while a complex relational semantic structure with different depths of interpretation is easier to construct in natural language, maybe just because students are used to it - at least more than to represent it on a graph.

The associative differences between the structural and semantic indices are partially attributed to the mathematical computation of these indices. The structural indices are based on associative networks, while the semantic indices are much more restrictive. For example, when HIMATT calculates a structural similarity, the procedure does not consider similar structure with similar terms, but only the structural similarity. Assuming that a structure has three nodes linked to a single node (i.e., star structure), the calculation is independent of the node terms. In contrast, for semantic similarity, the concept matching index determines if any of the node terms are identical. The propositional matching index determines if there are any link nodes that bind identical terms; a much more restrictive criterion causes an overall lower level of similarity independent of the student and reference model formats. Thus, the semantic matching measures within the used technologies aim at exact terms, and are therefore more restricted than the context-free indices of structure.

An additional aim of this study was to test the validity of the reference model formats. More specifically, we attempted to determine whether expert models are valid in both formats (concept map and text summary). To meet this aim, we included a third reference model format, which consisted of the entire texts that students were asked to read. The original text was used to test the notion that the experts' text summary and experts' concept map models are similar to the original text on all the seven HIMATT indices. The findings indicate that students' text summaries were stronger associated with the original text and the expert text summary and to a much lesser degree with expert concept map (see Figure 4). In contrast, students' concept maps were similarly associated with original text, expert text summary, and expert concept map except with the surface index. These findings further support the notion of the two stage cognitive process of semantic and graphic codes. The two reference criteria, original text and expert text summary, are in a written representation format, and thus associate closer to students' text summary. In contrast, expert concept map are graphical in nature, and thus differ in essence to the students' text summary. However, the structural indices associations between student text summary and expert concept map are within the range of 0.40 - 0.62 indicating sound similarities, and thus still acceptable criterion validity. Furthermore, the structural indices associations between
student concept maps and the expert concept map are within the range of 0.30 - 0.70 indicating that similar conceptual codes results in sound association coefficients and acceptable criterion validity. A within and between expert groups analyses (Ifenthaler, 2010a; Pirnay-Dummer, 2006) support the structural framework reported in this study, and comparison can be found within mental model theory, in particular within the successive model completion (e.g., Seel, 1991; Brachman, Fikes & Levesque, 1983).

To summarize, the use of concept maps and text summaries as a means of capturing students’ mental models has been proposed as a viable means to measure students’ understanding and conceptualization of specific domain knowledge (Johnson et al., 2009; Novak, 1998). Yet, the assessment methods to conduct such an analysis have not been sufficiently established. HIMATT was developed as a reliable tool for measuring learning and conceptual development (Pirnay-Dummer & Ifenthaler, 2010). The current study’s results provide criterion validity to the written text format of eliciting conceptualization from written learning texts. Graphical presentations are less valid for students who are unfamiliar with mental transformation from written text to mapping formats at this point. However, similar research must explore the assertion that learning and experience in “written text to vision” transformation of concept mapping may yield different results, and may enhance learning and reasoning in learners.

REFERENCES


