Improved Method for Creating Criterion Maps for Automatic Mind Map Analysis

Amber Franklin
Department of Speech Pathology
and Audiology
Miami University
Oxford, Ohio 45056
Email: frankla@miamioh.edu

Ryan Sunderhaus, Chris Bell, and Peter Jamieson
Department of Electrical and
Computer Engineering
Miami University
Oxford, Ohio 45056
Email: jamiespa@miamioh.edu

***Abstract*—In this work, we continue our study on analyzing student created mind maps automatically by providing a new methodology to select the technical vocabulary that students use in their mind maps. The basis of our previous experiments is an instructor chooses a set of twenty words used within the course that will be the set of words to test in mind maps. The instructor then creates their own mind map with this set of words, which is called the criterion map. Next, students create their mind maps using the same twenty words, and the criterion map and student map are analyzed with each other using various algorithms to produce metrics that quantify how similar the two maps are. When this activity is repeated longitudinally over a semester we can show that students are learning if their metrics of similarity are improving over time. One challenge, however, is which twenty words should be selected by the teacher. Similarly, will the set of twenty words impact the quality of observed learning. In 2011 and 2012, we collected our mind map data based on twenty words selected with no methodology (random). In 2013 and 2014, we created a methodology where approximately forty words are initially chosen, and these forty words are reduced down to 20 by creating a larger mind map and picking the words that have low connectivity. The hypothesis here is that less connectivity in the mind map will make it easier for the student to create their own quality maps. Our results show that this new methodology improves the arithmetic average of one of our best comparison metrics for all data points by a worse case of 2.4% better and best case 75% better.**

I. Introduction

Over the past few years, we have been exploring how mind maps as a class assessment technique (CAT) [1] can be used over a semester to measure student learning. The broad goal is to be able to use computers to provide learners with automatic feedback about aspects of their learning; more narrowly, the technique we are investigating allows computers to give students automatic feedback about their understanding of the relationships between the technical vocabulary used and introduced in a course.

Briefly, the methodology to do this is to have students create closed mind maps (a closed mind map uses words that are pre-determined) two to three times over a semester with the same set of technical words used in the respective course. Then using algorithms to analyze these maps and produce similarity metrics between an experts map and the students map, we can show if a student is improving, and also, we can provide the student with details about which vocabulary words they are wrongly associating with other vocabulary words.

We have studied this methodology with respect to the quality of the similarity metric used ([2], [3], and [4]) and we have studied how this methodology applies to different courses [5]. One key question that the community and some of our collaborators ask about these previous studies was how do you choose the words that are to be included in the experts mind map - the criterion map. This is the question addressed in this paper by providing a new method for creating the 20 words in the CAT and evaluating if our new methodology improves the overall similarity results.

Our hypothesis for out new method to create the 20 technical terms to be used in our mind map CAT is that the less densely connected a mind map is the easier it will be for the student. Since the CAT will be easier for a student to create we expect that the results will show more distinct difference between measurements at the various stages of a semester long study. We tested our methodology over four years using the ad hoc or random method in 2011 and 2012, and we used our new methodology in 2013 and 2014 for the same class.

Our results show that our new way of creating the technical vocabulary improves the class wide average distinction between metrics by a worse case of 2.4% better and best case 75% better. These results show that this improves the automatic measurements, but we note that this does not necessarily mean the students are learning better. However, the goal of this work is to provide automatic feedback to the student, and therefore, the benefits of these results is that a machine could automatically pick the set of words for the activity to start off as easier, but then as the student improves make the CAT more challenging.

The remainder of this paper is organized as follow: Section II describes what are mind maps and technical vocabulary assessment challenges; additionally, we examine research in multiple choice tests since a mind map is arguably a large scale multiple choice test in terms of creating relationships. Section III describes both the methodology for the mind map CAT experiment and the new way we propose to create the set of 20 vocabulary words. Section IV shows our results, and finally, section V provides a discussion and conclusion.

II. Mind Maps as a Class Assessment Technique

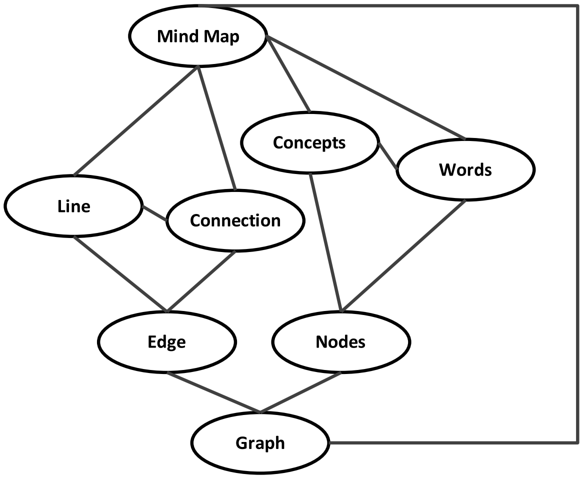

Fig. 1. Example of a mind map on the relationship between mind maps and graphs

Mind maps are visual representations of simple one-to-one relationships between words/concepts [6]. They can be used as CATs [1] allowing teachers to evaluate student understanding during class, and then, maps can be examined to provide feedback to their students. However, this is a time consuming activity for the teacher. Figure 1 shows an example mind map that expresses the authors understanding of mind maps and how they relate to a mathematical structure called graphs. The words/concepts that are in a mind map are the nodes of a graph (circled bubbles), and the connecting lines between these words are edges of a graph. Therefore, mind maps are also graphs meaning that they can be represented and analyzed algorithmically by computational machines.

There is continued interest in mind maps as a pedagogical tool that can help structure learning [7] as well as a vast and rich data set that can provide other insights [8]. In our past studies we have provided background on mind maps and scoring them, and our updated references:

1) comparing the scores on tests to the technique [9]
2) having two independent experts score (sometimes with a rubric) the mind map on a scale two times with one week delay and compare correlation of ratings [10]
3) using structures and frameworks to identify redundancies and troubling portion of a map [11]
4) using a large data-set of mind maps for deep fact finding of interrelated topics [12]

The types of mind maps we use in this study are called closed, which means they have a limited and predetermined set of words (nodes) [13]. Our scoring technique uses a criterion map which is a mind map created by the teacher/expert as a golden model that can be compared against the student's mind maps [14].

A. Mind Maps and Multiple Choice Tests

In our design of these experiments, an interesting question emerged, "Are closed mind maps just another form of multiple choice vocabulary tests with multiple answers?" In some ways a mind map exercise is a multiple-multiple choice test.

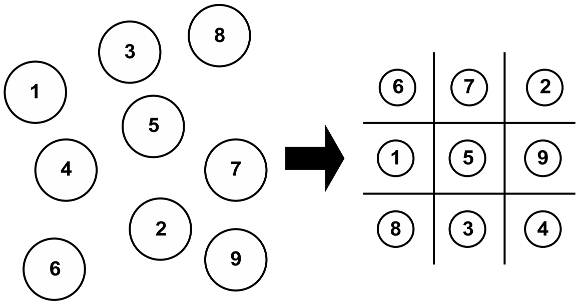

Fig. 2. How a number game of picking 3 numbers that make 15 is an isomorph of tic-tac-toe board.

The answer is they are similar, and this relates to the idea of mathematical isomorphs. More specifically, an isomorphic problem has multiple presentation formats at the surface level, but are the same problem underneath such as the tic-tac-toe problem figure 2 [15]. Isomorphic problems have been of interest to cognitive psychologists, and they have been used to help us understand strategical approaches people take to solving problems [16], [17].

Multiple choice tests are designed with respect to a *stem*, the problem, an *answer*, and a number of *distractors* [18]. The mind map in our case makes the *stem* a word (each word in turn will be a stem) and the edges that connect all related words which are the *answer(s)*, and therefore, all non-connecting words can be considered *distractors*.

Researchers have investigated the question of how similar should *distractors* be to the *answer*. For example, Mitkov *et. al.* looked at how similarity measures can be used to pick *distractors* and *answers* in multiple choice tests [19]. Similarly, Turney looked at a method to calculate "relational" similarity (correspondence between relations) between certain types of multiple choice questions [20]. This research is far broader than we can discuss in this paper and spans fields such as linguistics [21] to statistics [22], but the ideas and the terminology will be useful in describing this work.

In some way, our new methodology focuses on reducing the number of *answers* and increasing the number of *distractors*. This may seem bad for the learner and it would be if the similarity between *distractor* and *answer* was close, but we believe the *distractors* in this case are different enough to not confuse the learner.

III. Mind Map CAT - Experimental Methodology

As we have used our methodology in many studies, we will briefly review the experiment in first in section III-A, and then we will describe in section III-B our new methodology for selecting the words used in our closed mind map CATs.

A. Experimental Method

Our experimental methodology starts with the assumption that *most* students learn in a course over a semester, and this learning includes an improved understanding of the technical vocabulary as related to that course. The technical vocabulary is not necessarily the most important learning objective for a course, and in terms of emphasis and assessment, the vocabulary might be a periphery outcome called a "worth being familiar with" by Wiggins and McTighe's simple taxonomy [23]. This is the case for the course being investigated in this paper.

From this starting assumption, the idea is that over a semester a mind map CAT that captures a portion of the technical vocabulary being used in the class should improve as students make better connections between words as they understand more. In multiple choice test vocabulary, we might state that over a semester the number of correct answers (correct edges) increases while the number of distractors (incorrect edges) decreases. These trends can be captured by a computer that can measure this. However, the "right" answers need to exist, which is captured in what was defined earlier - the criterion map. A criterion map is a min map created by the expert with the same technical vocabulary words that contains the information of correct relationships between words. Therefore, a student's mind map and a criterion map are inputted into a machine that uses an algorithm(s) to measure and produce a metric of measurement that reflects similarity between the two maps. In a future CATs the new student created mind map is created and again compared to the criterion map to create a new metric that will show trends of improving or worsening similarity between the two maps - an improvement suggests learning based on our original assumption.

We have investigated a number of similarity metrics to see which best captures learning, and which metrics can be used to provide the students useful feedback [2], [3], [4]. Our two best metrics that satisfies these two requirements are:

1) Match Metric at a node and edge level - The *match metric* is an edge by edge comparison between the student mind map and the criterion map. The nodes in our graphs are uniquely identified by a label (the term written in the bubble), and this allows us to compare the two graphs in linear time. During this comparison a number of statistics are recorded about the differences including missing nodes ($MissN$), extra edges ($ExtraE$), and matching edges ($MatchE$) where the comparison is the student map as compared to the criterion map. The *match metric* is a combination of these statistics:

$$MatchMetric = \frac{MatchE}{MissN + ExtraE + MatchE} \tag{1}$$

This equation results in a number between 0 and 1. The number is interpreted where as it approaches 1 indicates there is more similarity between the two graphs.

2) Match Metric at a graphlet level - The general form of our graphlet based match-metric is:

$$GraphletMatch = \frac{\sum_{g_i}^{i\epsilon M} \frac{n_{s_{g_i}}}{n_{G_{g_i}}}}{g_{total}} \tag{2}$$

where g_i is a graphlet type (graphlets, which, are "a connected network with a small number of nodes" [24] and these small graphs are non-isomorphic induced subgraphs of a larger graph) from g0 to g29, M is the set of included graphlets in the metric (as in a selection of the graphlets that have meaning derived from the previous section), $n_{s_{g_i}}$ is the total number of matched metrics in of that type of graphlet, $n_{G_{g_i}}$ is the total number of that type of graphlet in the criterion map, and g_{total} is the total number of graphlets (as in $\sum_{g_i}^{i\epsilon M} n_{G_{g_i}}$). This metric will produce a number between 0 and 1 where as the number approaches 1, the student's mind map is closer to the criterion map.

Both metrics 1 and 2 are useful in showing trends in student learning, and both can be used to show student low-level feedback on their misunderstandings, but the 2nd metric can provide more interesting feedback on deeper relationships a student might be missing. For example, a *triangle* graphlet (g2) shows a relationship of three ideas. This is more significant than a one-to-one relationship and is important to make a student aware of this [25]. However, since this study focuses mainly on trends and not detailed feedback, we report all our results based on metric 1 noting that metric 2 has similar behavior.

The steps of each experiment are:

1) Select 20 technical vocabulary words
2) The expert creates a criterion mind map
3) Students create their own mind maps as a 10 minute CAT before the course and after exams
4) The maps are compared with the criterion map using a similarity metric

In this method, one question that comes up is why 20 words? The reason for this is more logistical than technical in that 20 words can be made into a mind map in the 10 minutes allocated. This gives the student approximately 30 seconds per word or around 2 seconds to compare the current word with other words to determine if there is a relationship, which seems reasonable. Using our methodology we could experimentally evaluate if less or more words than 20 has some impact on our results, but we do not believe valuable ideas would emerge from an experiment like this.

Another caveat with this experimental methodology is the question of is the criterion mind map good and how much subjectivity is there in creating such a map since it is a reflection of the individual's understanding (even though the teacher is the expert). This reality is a further concern for these experiments given that both the context of how the mind map is being thought about and that the English language is full of exceptions and meanings will impact results. For example, imagine the words "apple" and "frog" are part of the vocabulary. In the context of grammar, these might be related in a mind map since they are nouns; however, in the context of biological classification, they might not be related in terms of vegetable and animal, or are related in terms of living organisms. In regard to this, we believe the way forward is in the domain of big data and using many criterion maps and student mind maps to create large data-sets that provide interesting feedback and aren't just simply comparing how similar maps are. We, however, do not have the resources to explore such possibilities at present.

Finally, in terms of experimental design the creation of mind maps is done on paper. This requires an additional step to convert these paper graphs into digital graphs that can be represented in the computer. This data encoding is done by hand.

B. New Method to Create a Criterion Map

This work looks at improving steps 1 and 2 of the experimental method by providing a better way to select the 20 technical words and creating the criterion map. By better we do not mean how an expert should create their mind map, but instead what are some graph properties of the criterion map such that the resulting student created mind maps will show more differentiation and more similarity with respect to the match metric for the CAT activities.

We start, here, by defining the degree of a node and the density of a graph before stating our hypothesis. The degree of a node is the number of adjacent nodes, or the number of edges that connect to other unique nodes. For example, in figure 1 the degree of the "Mind Map" node is 5 and the degree of the "Connection" node is 3. The average degree is the total number of edges ($|E|$) divided by the total number of nodes ($|V|$). For this example, the graphs average degree is 12 divided by 8 or 1.5 . The *density* of a graph is defined as the total number of edges ($|E|$) divided by the maximum number of edges that a graph could have ($.5*|V|*(|V|-1)$). For the graph in figure 1, the maximum number of possible edges is 28 and the number of edges is 12, so the density is 12/28 or, approximately, 0.43.

Our hypothesis is that a lower average degree and graph density of the criterion map will result in a set of technical vocabulary words that will be easier for students to create their own mind maps with and their mind maps will show a more clear learning progression for our experiment. The extreme of this would be choosing 20 words that have no relationship at all, and a slightly less extreme version of this would be selecting words that have a linear relationship, which result in a graph that looks like a line. Our goal is not to remove all deeper relationships in the graph, but to make the connectivity of the graph lower so that the resulting criterion map has complexity (in terms of connectivity), but is an attempt to be on the lower end of complexity.

In our early versions of this work (2011 and 2012), we selected the 20 words from a course in an ad hoc method, which was basically random. For example, the course is a Digital System Design course at Miami University. This is a 200 level course that introduces students to the concepts of how transistors can be organized into gates, gates into useful combinational and sequential circuits, and useful circuits into larger systems that can control and compute. Key divisions in this course is between combinational and sequential circuits (where combinational circuits are taught in the first part of the course and sequential circuits in the second part of the course), number systems and calculations (introduced in the middle of the course), and the how to create Verilog designs as opposed to schematic designs. Therefore, with these key divisions we attempted to pick words from all of these domains to provide a good sample set of words for the course.

TABLE I. CLASSIFICATION OF 20 WORDS

Year	Comb.	Seq.	Verilog	Numbers	Other
2011	7	4	2	3	4
2012	7	3	2	3	5
2013	5	3	1	5	6
2014	7	4	2	4	3

Table I shows how the 20 technical words used in the mind map CATs are distributed into these groupings over each of the years this experiment was done. The "Other" category in column 6 is for some big concept words that don't fit into any one category; for example, digital circuits are designed with a goal of making them fast, area efficient, and power efficient, and therefore, these concepts would be classified as "Other" since they apply to many of the previous ideas.

In our ad hoc method, the main goal was to have words sampling from all the above categories, and not too much emphasis in one particular area. Once the words were selected then the expert would create their mind map, and the only restriction was that the criterion map used all the words and the graph was fully connected (meaning there was a connection path between all nodes via other nodes and edges).

The new methodology to create the 20 words that we propose has the following steps:

1) Select approximately 40 words (roughly double the target word count goal) with a reasonable selection (6 or more words) from each of the above categories
2) Have the expert create a mind map with all 40 words
3) Calculate the node degree (number of edges connecting to that node) for each node in the graph
4) In a greedy fashion remove nodes with the highest degree until you have 20 words. In cases of ties, remove words that keep the graph fully connected, remove a word that has a very similar to pair or opposite (for example, "one" and "zero") or comes from the same category. If this does not exist then randomly select which word to remove.

In table I, we can see that this methodology keeps the final categorized word count for our new methodology (2013 and 2014) similar to that of the our earlier counts (2011 and 2012).

The goal of this methodology is to create a criterion map that consists of less connected nodes as measured by average node degree and graph density. Table II shows how the new methodology changes the connectivity of the graphs in our experiment from the 2011 and 2012 ad hoc criterion maps to the 2013 and 2014 new method criterion maps. We also include the initial density and degree for the 40 word criterion map as shown in columns 4 and 5. The key observation here is that our new methodology creates criterion maps that are less connected than in previous years as expressed by the degree and density measurements (a smaller number for these metrics means less connectivity, which means less edges or word to word relationships).

This new methodology is more directed, but is not perfect. For example, it is possible (though it didn't occur) that the methodology results in what is called a disjoint graph (not fully connected). This isn't necessarily a problem, but it was a property that we didn't want. Also, for a highly connected

TABLE II. AVERAGE DEGREE AND DENSITY OF CRITERION MAPS

	20 Word Criterion Map		40 Word Criterion Map	
Year	Average Degree	Graph Density	Average Degree	Graph Density
2011	2.35	0.25	-	-
2012	2.85	0.30	-	-
2013	1.45	0.15	5.02	0.26
2014	1.10	0.12	4.56	0.24

set of 40 words, even this methodology might not result in a significantly less connected graphs. In our cases, the initial 40 word maps had a node degree distribution ranging from 1 to 11 and Table II shows how the 40 word criterion maps original average degree and density were high, but were significantly reduced by using the method.

IV. RESULTS

To investigate our hypothesis, we collected data for students who participated in our study (IRB approved) from 2011 to 2014. In each year, we followed the experimental methodology as described above. We performed the CATs on our digital system design class at the beginning of the class *Pre*, after exam one *ExamI*, and after exam two *ExamII* with the exception of missing a data collection point in 2013 for *ExamI*. Depending on absences the total population at each sample point deviates slightly from the population size reported. Finally, the changes to the course from 2011 to 2014 were small and the same professor (Dr. Jamieson) taught all of these sections. However, each year Dr. Jamieson did make efforts to improve the course focusing much of his attention on helping students improve their Verilog design ability.

TABLE III. ARITHMETICALLY AVERAGED MATCH METRIC RESULTS FROM 2011 TO 2014

Year	Population	Pre	ExamI	ExamII
2011	40	0.202	0.290	0.295
2012	41	0.160	0.240	0.271
2013	47	0.233	-	0.302
2014	58	0.229	0.377	0.473

Table III shows the results based on the arithmetically averaged match metric for each year at each measurement stage. Column 1 and 2 show the year and population of students who participated in the study (again noting that each sample point had a population that was equal to or less than this number based on attendance). Columns 3 through 5 shows the arithmetically averaged match metric (metric 1 described earlier rounded to 3 decimal places) for each of the measurement stages comparing each student's mind map to the criterion map created that year.

One thing to note, which is more thoroughly investigated in previous papers, is that the match metric number is growing as the semester progresses, and we say that this trend shows that students are learning the technical vocabulary better as the semester progresses. This is not always the case in all courses as investigated in [5]. However, for the study course we have consistently seen these trends.

As a reminder, the match metric is a measurement between 0 and 1 that as it approaches 1 means that the student created mind map is more similar to the criterion map. In table III the size of the match metric when compared between the 2011 and 2012 ad hoc criterion map creation group to the 2013 and 2014 new method criterion map creation group we see that the 2013 and 2014 numbers are larger when comparing similar time of measurement (column by column *Pre*, *ExamI*, and *ExamII*). In 2014, the values are considerably larger where the *ExamI* match metric results are better than any other measurement except the same years *ExamII* result.

The distinction from the starting measurement to ending measurement is also more significant in 2014 than in any other year (almost a difference of 0.25). However, all other years differences are about 0.1. A distinguishing features for the 2014 year is that the criterion map has the lowest density and average degree compared to all other criterion maps in this study, and another factor might be the number of "Other" words (see Table I) is also the lowest. Our advice for others using this technique is to use our methodology as proposed in the previous section, but also keep the number of forest concepts (words that relate to ideas throughout the course) to a minimum to get the best results strictly from the perspective of measurement.

Comparing the percentage improvement of the match metric between the two groups, the lowest percentage improvement is 2.47% between *ExamII* in 2013 to 2011 and the greatest percentage improvement is 74.6% between *ExamII* in 2014 to 2011. In general, our new technique always improves the average class performance as measured by the match metric.

V. DISCUSSION AND CONCLUSION

In this work, we provided a new methodology to create criterion maps for automatically evaluating them with machine based algorithms that compare a student's mind map to a criterion map measured three times over a semester. Our results show that our new technique provides teachers with a clearer methodology to pick the 20 technical vocabulary words to include in this CAT, and that following our proposed methodology will end in better and more differentiated similarity measurement results. Additionally, our data suggests that picking words that are not high level concepts (forest concepts) that apply to many ideas introduced throughout the course and focusing on more distinct ideas (tree concepts) results in a more clear measurement of learning.

These results are useful in terms of the goal of automatic feedback on mind map CATs. However, this does not suggest that educators should not consider using complex mind maps as CATs in their classes. Depending on the learning objective and the goal of the activity complex and dense mind maps may be more valuable. Strictly from the perspective of machine analysis, graph connectivity is a parameter we can control for to make the activity simpler and easier for beginners. In a way, this is a parameter that could be adjusted by a machine to change the difficulty of the activity.

At this point in our study of how to automatically analyze mind maps to provide students with feedback on their learning, we believe that future directions should focus on systems that collect larger sets of data. We believe our methodology has been refined successfully to find the best techniques and a solid methodology for performing such measurements. This will also improve the "criterion map" since the idea of expert could also be crowd sourced and would improve the comparison map. To achieve these goals, the focus needs to be put on a web-based system that can collect this data, and we are not certain we are the best group to pursue these ideas.

References

[1] T. Angelo and K. Kross, *Classroom Assessment Techniques - A Handbook for College Teachers*. Jossey-Bass, 2003.

[2] P. Jamieson, "Using modern graph analysis techniques on mind maps to help quantify learning," in *Frontiers in Education Conference (FIE), 2012*, oct. 2012. [Online]. Available: http://www.users.muohio.edu/jamiespa/html_papers/fie_12.pdf

[3] ——, "More graph comparison techniques on mind maps to provide students with feedback," in *Frontiers in Education Conference (FIE), 2013*, oct. 2013. [Online]. Available: http://www.users.muohio.edu/jamiespa/html_papers/fie_13.pdf

[4] P. Jamieson and J. Eaton, "owards a better graphlet-based mind map metric for automating student feedback," in *to be published at 2015 ASEE Annual Conference and Exposition*, 2015.

[5] A. Franklin, T. Li, P. Jamieson, J. Semlak, and W. Vanderbush, "Evaluating metrics for automatic mind map assessment in various classes," in *Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE*, Oct 2015, pp. 1–8.

[6] T. Buzan and B. Buzan, "The mind map book how to use radiant thinking to maximise your brain's untapped potential," *New York: Plume*, 1993.

[7] M. J. Somers, K. Passerini, A. Parhankangas, and J. Casal, "Using mind maps to study how business school students and faculty organize and apply general business knowledge," *The International Journal of Management Education*, vol. 12, no. 1, pp. 1–13, 2014.

[8] H. Faste and H. Lin, "The untapped promise of digital mind maps," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1017–1026.

[9] I. Abi-El-Mona and F. Adb-El-Khalick, "The influence of mind mapping on eighth graders science achievement," *School Science and Mathematics*, vol. 108, no. 7, pp. 298–312, 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1949-8594.2008.tb17843.x

[10] E. Evrekli, D. Inel, and A. Galim, "Development of a scoring system to assess mind maps," *Procedia - Social and Behavioral Sciences*, vol. 2, no. 2, pp. 2330 – 2334, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187704281000371X

[11] P. Kedaj, J. Pavlíček, P. Hanzlík *et al.*, "Effective mind maps in e-learning," *Acta Informatica Pragensia*, vol. 3, no. 3, pp. 239–250, 2014.

[12] J. Beel, S. Langer, M. Genzmehr, and B. Gipp, "Utilizing mind-maps for information retrieval and user modelling," in *Proceedings of the 22nd Conference on User Modelling, Adaption, and Personalization (UMAP)*, 2014.

[13] H. E. Herl, H. F. O'Neil, G. K. W. K. Chung, and J. Schacter, "Reliability and validity of a computer-based knowledge mapping system to measure content understanding," *Computers in Human Behavior*, vol. 15, no. 3-4, pp. 315–333, May 1999. [Online]. Available: http://dx.doi.org/10.1016/S0747-5632(99)00026-6

[14] M. Ruiz-Primo, R. Shavelson, and S. Schultz, "On the validity of concept-map-based assessment interpretations: An experimental testing the assumption of hierarchical concept maps in science," University of California, Tech. Rep., 1997. [Online]. Available: http://research.cse.ucla.edu/Reports/TECH455.PDF

[15] J. Zhang, T. Johnson, and H. Wang, "Isomorphic Representations Lead to the Discovery of Different Forms of a Common Strategy with Different Degrees of Generality ," in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, 1998. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=93CB1A5E27B562C8B7BD109A1BF3A241?doi=10.1.1.139.9596&rep=rep1&type=pdf

[16] H. A. Simon and J. R. Hayes, "Understanding Process - Problem Isomorphs," *Cognitive Psychology*, vol. 8, no. 2, pp. 165–190, 1976.

[17] K. Kotovsky, J. R. Hayes, and H. A. Simon, "Why are some problems hard? evidence from tower of hanoi," *Cognitive psychology*, vol. 17, no. 2, pp. 248–294, 1985.

[18] S. J. Burton, R. R. Sudweeks, P. F. Merrill, and B. Wood, *How to prepare better multiple-choice test items: Guidelines for university faculty*. Brigham Young University Testing Services Utah, 1991.

[19] R. Mitkov, L. A. Ha, A. Varga, and L. Rello, "Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation," in *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics, 2009, pp. 49–56.

[20] P. D. Turney, "Similarity of semantic relations," *Computational Linguistics*, vol. 32, no. 3, pp. 379–416, 2006.

[21] M. Paxton, "A linguistic perspective on multiple choice questioning," *Assessment & Evaluation in Higher Education*, vol. 25, no. 2, pp. 109–119, 2000.

[22] M. Amo-Salas, M. d. M. Arroyo-Jimenez, D. Bustos-Escribano, E. Fairén-Jiménez, and J. López-Fidalgo, "New indices for refining multiple choice questions," *Journal of Probability and Statistics*, vol. 2014, 2014.

[23] G. P. Wiggins, J. McTighe, L. J. Kiernan, and F. Frost, *Understanding by design*. Association for Supervision and Curriculum Development Alexandria, VA, 1998.

[24] N. Przulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[25] I. M. Kinchin, D. B. Hay, and A. Adams, "How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development," *Journal of Educational Research*, vol. 42, pp. 43–57, 2001. [Online]. Available: http://www.personal.psu.edu/kmo178/blogs/kmorourke/qualitative%20approach%20to%20concept%20map%20analysis.pdf