Concept mapping assessment in medical education: a comparison of two scoring systems

Daniel C West,1 Jeanny K Park,1 J Richard Pomeroy2 & Jonathan Sandoval2

Background Concept mapping has the potential to measure important aspects of a student’s evolving knowledge framework in a way that conventional examinations cannot. This is important because development of an elaborate and well-structured knowledge framework is a critical step toward becoming an expert in a particular field. Little is known about the best way to score concept maps in the setting of medical education. Therefore, as a preliminary step in addressing this question, we compared two different scoring systems for validity: a structural method based on the organization of a map’s hierarchical structure and a relational method based, not on structure, but on the quality of each individual map component.

Methods A total of 21 paediatric resident doctors completed concept map training, drew a preinstruction concept map about ‘seizures’, completed a seizure education course, and then drew a postinstruction seizure map. Two raters using both structural and relational methods scored each map.

Results Structural scores increased significantly after instruction and were higher in more experienced residents, but relational scores were not significantly different. Interrater scoring reliability for both methods ranged from moderate to strong, but was greater using the relational scoring method.

Conclusions These data suggest that scoring systems for evaluating concept maps in postgraduate medical education may need to account for structural features of maps, if scores are to reflect changes in the developing knowledge frameworks of resident doctors. More research to further evaluate reliability and validity is critical prior to any future use of concept mapping assessment in medical education.

Keywords *Concept formation; education, medical, graduate/*standards; educational measurement; reproducibility of results.

Medical Education 2002;36:820–826

Introduction

According to constructivist learning theory we learn, not simply by memorizing new facts and concepts, but also by assimilating that information into a pre-existing framework of knowledge.1 Supporting this theory is empirical evidence that development of an elaborate and well-structured knowledge framework, at least in part, determines how knowledge is used for decision making and problem solving and is a critical step toward becoming an expert in a particular field.2–4 Despite its importance, assessment of a student’s evolving knowledge framework is often neglected, probably because it is difficult to measure.5

Concept mapping has the potential to measure aspects of a student’s evolving knowledge framework. In concept mapping, the student links together a series of concepts related to a particular topic (called concept-links and cross-links) producing a two-dimensional, hierarchically organized diagram representing their knowledge framework.6 Concept mapping is now frequently used as a learning and organizational tool in a variety of classroom and professional settings. However, it is rarely used as an assessment method because there is limited information regarding its validity and reliability, particularly in medical education.

The validity of concept mapping assessment is largely dependent on how maps are interpreted or scored. The two most commonly investigated scoring methods are structural and relational. The structural method assigns weighted scores based on the organization of a map’s

1Department of Pediatrics, School of Medicine, University of California, USA
2Division of Education, University of California, USA
Correspondence: Daniel C. West, MD, Department of Pediatrics, University of California, Davis, 2516 Stockton Boulevard, Sacramento, California 95817, USA. Tel.: 00 1 916 734 2428; Fax: 00 1 916 734 0342; E-mail: dcwest@ucdavis.edu
hierarchical structure, concept-links, and cross-links. 6 In contrast, the relational method takes into account the quality or importance of each individual concept-link, without regard to the overall structure of the map. Using the structural scoring method, we recently demonstrated that concept maps could be a valid measure of the knowledge frameworks of pediatric residents at various stages of postgraduate training.7 In that study, we observed that as residents’ knowledge and experience increased, maps grew more complex and difficult to score. Relational scoring methods may have potential advantages over structural methods when scoring complex maps, because they place greater importance on the correctness and overall quality of an individual concept-link and are simpler to use, and thus, potentially more reliable. Evidence in other educational settings also suggests that relational scoring methods may be more valid.5,8 At present, we do not know which scoring method, structural, relational, or otherwise, is most valid and best reflects a student’s or resident’s knowledge framework as represented in a concept map. However, it is critically important that we determine the most valid method if concept mapping assessment is ever to be used on a larger scale in medical education.

In this article, we investigate whether the scoring of concept maps using a relational method is a valid measure of the knowledge frameworks of pediatric resident doctors during postgraduate medical training.8 As a measure of predictive validity, we re-scored concept maps from our earlier study using a relational method and compared relational scores of residents before and after instruction and at different levels of training with previously validated structural scores.

Methods

Subjects

This study was performed in the setting of postgraduate pediatric training at the University of California, Davis. Our institutional human subjects review committee approved the study and the investigators obtained written informed consent. Twenty-one pediatric residents enrolled (9 first year residents, 8 second year residents, and 4 third year residents).

Study design

Details of the study design have been described previously.7 Briefly, all study subjects participated in a standardized concept map training exercise after which they were given 30 minutes to draw independently, and without study aids, a preinstruction concept map about ‘seizures’ on a blank sheet of paper. Residents then participated in a triple session (1 hour per session) seizure education course consisting of ‘case-based’ instruction about the diagnosis and management of seizure disorders in children. Residents then drew a postinstruction concept map about seizures under identical conditions as before.

Concept maps

We used a modification of the ‘hierarchical’ method of concept mapping developed by Novak.6 To make a map, a resident links together a series of concepts (defined as a perceived regularity in events or objects designated by a label) using arrow lines with a statement (proposition) written above the line describing how the concepts are related (referred to as a concept-link).6 In the preinstructional map shown in Fig. 1, the concept seizure linked to nonfocal by the proposition ‘can be classified as’ is an example of a concept-link. Map hierarchy is indicated by the direction of the arrow in the concept-link and the arrangement of concepts in the map with more general concepts at the top and more specific concepts below or extending outward. Arrangement of a series of related concept-links in a hierarchical chain represents a domain of knowledge. The relationship between different domains of knowledge is indicated by cross-links, which are also connected using...
Figure 1 Pre-instruction and post-instruction concept map from the same resident (approximately 50% reproduction of hand-written concept maps).
arrow lines labelled with a phrase describing the relationship. In Fig. 1, the concept focal linked to CT/MRI by the phrase needs is an example of a cross-link. The final component is examples, which is linked to the related concept in the most subordinate position on the map and not enclosed in a circle.

**Scoring methods**

Two different raters, blinded to the identity of the map author, independently scored each map using both the structural and relational methods. Maps were first scored using the structural method and subsequently using the relational method. Both raters were certified by the American Board of Paediatrics and completed standardized concept map scoring training as previously described.7

**Structural method**

The structural scoring method has been described previously.7 The theoretical basis for the weighted scoring scale is based on the depth of thought or understanding required to form the category and focuses on the relative position of each concept in the map.6 Briefly, maps were scored in 4 categories with the following point assignments for each valid component:

- concept-link (2 points each);
- cross-links (10 points each);
- hierarchy (5 points each);
- examples (1 point each), and
- invalid components (0 points).

Scores for each component were added together to produce a total structural score for each map by each rater. Raters’ relational scores were added together to create a standard relational score for each map.

**Data analysis**

All statistical analyses were performed using STATA version 6.0 (College Station, Texas, USA) and because of the small sample size, nonparametric methods to compare rank order were used. As a measure of interrater reliability, the correlation between each rater’s total structural and total relational scores for each map was tested using Spearman rank correlation. The remaining analyses were performed using standard structural and standard relational scores. As a measure of predictive validity, we compared prestructural with poststructural map scores with each scoring method using the Wilcoxon sign rank test. To assess differences based on level of training, we compared the map scores of more experienced residents (second and third years, n = 12) with those of less experience (first years, n = 9) using the Two-sample Wilcoxon rank sum test.

**Results**

All 21 participants completed all phases of the study. Interrater correlations were strong using the relational method (preinstructional, r = 0.96; postinstructional, r = 0.91) and moderate to strong using the structural method (preinstructional, r = 0.65; postinstructional, r = 0.84).

**Scores before and after instruction**

Figure 1 shows a partial reproduction of a hand-written preinstruction and postinstruction concept map from the same individual. Qualitative comparison of the maps before and after instruction revealed incorporation of new concepts into maps, further differentiation of existing concepts, new cross-link relationships, and conversion of examples into concepts. Table 1 summarizes the median standard relational and structural scores on pre- and postinstruction maps. Relational scores did not change significantly after instruction (P = 0.14), while structural scores increased (P = 0.004).

**Scores by level of training**

Table 1 also summarizes the median standard structural and relational scores based on training level. On the preinstruction maps, relational scores of second and
third year residents were not significantly different to those of first year residents \((P = 0.72)\), while structural scores were higher for second and third year residents \((P = 0.03)\). After instruction, there was no significant difference based on level of training (relational, \(P = 0.37\); structural, \(P = 0.18\)).

Discussion

Previously, we demonstrated that structural concept map scores of resident doctors increased significantly in situations where knowledge frameworks were expected to change (i.e. after instruction) and were significantly higher in residents with more training compared with those with less.\(^7\) These findings are similar to those reported using structural scoring methods in other educational settings and suggest that structural concept map scores are a valid measure of expected change and differences in the knowledge frameworks of resident doctors.\(^9,10\) In the present study, we found that relational scores of the same concept maps did not detect predicted changes in resident knowledge frameworks after instruction or predicted differences based on level of training. Thus, although interrater reliability of the relational scoring method was high, this method failed to demonstrate the validity seen using the structural scoring method. These findings imply that the relational method may be less sensitive than the structural method in measuring changes and differences in concept maps, at least as applied in this limited setting in postgraduate medical education.

Other published reports using relational scoring methods arrive at different conclusions to those of our study. In one particularly relevant study in the setting of a college biology course, investigators compared the validity and reliability of the relational method to other scoring methods, including the structural method.\(^8\) As in our study, these investigators found that the relational scoring method had greater interrater reliability than the structural method. In contrast to our findings, they concluded that the relational method was most valid based on the degree of similarity of scores generated by each method with scores of an expert reference map. Other investigators have tested variations of the relational scoring approach, such as rating the ‘goodness’ of each concept map link from high school physics students or the ‘correctness’ as they related to a biology course curriculum.\(^11,12\) Although no comparisons with structural scoring were made in these reports, these investigators showed that map scores correlated with multistep problem-solving examination scores in physics and multiple choice examination scores in biology. These observations appear to contradict the findings in our study; however, the discrepancies could be due to the use of substantially different measures of validity or subtle variations in scoring methodology. Furthermore, these studies were performed in markedly different educational settings compared with our study (secondary/college versus postgraduate medical education), raising the question of whether students’ concept maps could change in different ways depending on the material to be learned or the students’ previous training.

Table 1 Comparison of structural and relational scores of the same concept maps

<table>
<thead>
<tr>
<th></th>
<th>Structural method</th>
<th></th>
<th>Relational method</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Score (Range)*</td>
<td>P value</td>
<td>Score (Range)*</td>
</tr>
<tr>
<td><strong>Comparison of scores before and after instruction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All residents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preinstruction map</td>
<td>21</td>
<td>303 (179–508)</td>
<td>0.004†</td>
<td>320 (99–444)</td>
</tr>
<tr>
<td>Postinstruction map</td>
<td>21</td>
<td>387 (171–725)</td>
<td></td>
<td>321 (135–538)</td>
</tr>
<tr>
<td><strong>Comparison of scores by level of training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preinstruction maps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second/third year residents</td>
<td>12</td>
<td>349 (184–508)</td>
<td></td>
<td>307 (187–444)</td>
</tr>
<tr>
<td>First year residents</td>
<td>9</td>
<td>263 (179–400)</td>
<td>0.03‡</td>
<td>320 (99–431)</td>
</tr>
<tr>
<td>Postinstruction maps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second/third year residents</td>
<td>12</td>
<td>467 (171–715)</td>
<td></td>
<td>335 (135–538)</td>
</tr>
<tr>
<td>First year residents</td>
<td>9</td>
<td>343 (196–725)</td>
<td>0.18‡</td>
<td>314 (222–513)</td>
</tr>
</tbody>
</table>

* Data are presented as median (range) standard scores (see method section for explanation of scoring methods).
† For comparison by Wilcoxon signed rank test.
‡ For comparison by Two-sample Wilcoxon rank sum test.
It is possible that the seizure educational course in our study did not improve relational concept mapping scores because the instruction was ineffective. Alternatively, it is possible that the concept mapping assessment did not measure areas influenced by the instruction. These possibilities are less likely because we have previously shown, using qualitative comparison of pre- and postinSTRUCTION concept maps, that maps changed after instruction. Observed changes included incorporation of new concepts into maps, further differentiation of existing concepts, and enhanced cross-linking between domains of knowledge. In addition, ineffective instruction could not account for the lack of difference in relational scores of residents at different levels of training.

Because of the small sample size, our study had limited power to detect differences in scores between groups; thus, it is possible that with a larger sample size, significant differences in relational scores of smaller magnitude might become apparent. In addition, the small sample size required that we use nonparametric methods of statistical analysis. Consequently, we were unable to perform a more robust analysis (e.g. G-theory analysis) of other factors that influence reliability and estimate the extent to which concept mapping assessment could be generalized to other groups of learners and raters. The need to use nonparametric methods of analysis also prevented us from determining interrater agreement (i.e. the degree to which each rater’s map score was in absolute agreement); therefore, we cannot be certain that one rater did not commit systematic errors that could account for our findings. To minimize the impact of our limited ability to analyse reliability and interrater agreement, the study was designed so that both raters scored all maps, and similarly to other investigations, we used the sum of both raters’ scores on each map for analysis, rather than each rater’s individual score. We then assessed interrater reliability by determining the extent to which raters’ scores were proportionally in agreement using Spearman rank correlation. By this measure, the interrater reliability in our study was similar to that reported in other investigations of concept mapping assessment.

Concept mapping assessment holds great promise as an evaluation tool in medical education because it has the potential to measure the evolving knowledge frameworks of students and residents as they progress from novice to expert doctors in a way that conventional examinations cannot. Furthermore, it could provide medical educators with a unique tool to uncover distortions of students’ understanding of content and to identify errors of omission. However, developing valid and reliable methods of scoring concept maps is critical to their future utility. The limited size and scope of our study require that we use caution when drawing conclusions; nevertheless, our findings have important implications for any future use of concept mapping assessment in medical education. Our data suggest that scoring systems for evaluating concept maps in postgraduate medical education need to account for the structure of expressed relationships between and among concepts (i.e. structural features of maps), if scores are to reflect changes in the developing knowledge frameworks of resident doctors. More research is needed to confirm these findings and to establish more rigorously the reliability of concept mapping assessment in medical education before it should be used as an assessment tool outside the investigational setting.

Contributors

All the authors participated in concept development and research design. JRP developed and taught study subjects concept mapping. JKP and DCW scored the concept maps. DCW performed the data analysis with review by JS. All authors participated in data interpretation. DCW wrote the manuscript with critical review and revision by all authors.

Acknowledgements

We thank LuAnn Cruys, Debbie Barnes, and Sally Colvin for administrative support and Constance Bowe, MD, Department of Neurology, University of California, Davis, for teaching the seizure education course. Importantly, we thank the paediatric residents at UC Davis for enthusiastically participating in this study.

This paper was presented in part at the American Education Research Association meeting, Seattle, Washington, USA, April 11, 2001 and at the Paediatric Academic Societies Joint Meeting, Baltimore, Maryland, USA, April 28, 2001.

Funding

This work was supported in part by grants from the Children’s Miracle Network Research Fund at the University of California, Davis, USA.

References


Received 29 June 2001; editorial comments to authors 17 October 2001; accepted for publication 14 April 2002