International Journal of Engineering Education Vol. 34, No. 3, pp. 1025–1037, 2018
Printed in Great Britain

0949-149X/91 $3.00+0.00

Assessment of Conceptual Knowledge using a Component-Based Concept Map Scoring Program*

MARY KATHERINE WATSON
Associate Professor, Civil and Environmental Engineering, The Citadel, The Military College of South Carolina. 171 Moultrie Street, Charleston, SC 29409, USA. E-mail: mwatson9@citadel.edu

ELISE BARRELLA
Assistant Professor, Department of Engineering, Wake Forest University, 455 Vine Street, Winston-Salem, NC 27101, USA.
E-mail: barrelem@wfu.edu

JOSHUA PELKEY
Senior Product Manager, VMware AirWatch, 1155 Perimeter Center West #100, Sandy Springs, GA 30338, USA.
E-mail: joshpelkey@gmail.com

Conceptual understanding is an important prerequisite for engineering competence. Concept maps, which capture the content and structure of knowledge, can be used to assess conceptual knowledge, although cumbersome scoring methods limit their use. A literature review was conducted to summarize concept map scoring methods and automated scoring programs. While quantitative, component-based methods prevailed in the literature, no program was available to automate this method. Thus, the goal of this project was to present and evaluate a component-based computer program for scoring concept maps. The program automates application of the traditional scoring method in which number of concepts, highest hierarchy, and number of cross-links are counted as indicators of knowledge breadth, depth, and connectedness, respectively. A sample of concept maps (n = 78) was scored by two judges and the computer program. High agreement (Krippendorff's alpha > 0.80) between manual and automated scores was observed for number of concepts and number of cross-links. Although less than acceptable agreement between manual and automated scores was observed for highest hierarchy, the two measures of knowledge depth were highly correlated (Spearman's rho > 0.5). Ultimately, the computer program's measure of knowledge depth was termed longest path, while judges' measure of knowledge depth was termed longest hierarchy. Overall, the computer program can be used to rapidly, precisely, and reliably score concept maps to aid in assessment of conceptual knowledge.

Keywords: assessment; concept maps; conceptual knowledge; educational technology

1. Introduction

Conceptual knowledge is an important facet of engineering competence, which allows practitioners to deviate from established heuristics to create innovative designs. Conceptual knowledge is factual, structured, and interrelated. Rittle-Johnson [1] describes that conceptual knowledge includes "understanding of principles governing a domain *and* the interrelations between units of knowledge in a domain [1]" (pg. 2). Starr [2] describes that conceptual understanding must be "deep" and "rich with connections" (pg. 408).

Concept maps, which are graphical tools for organizing knowledge, encourage students to transcribe their own knowledge networks into a tangible construct that can be viewed by others. Consequently, concept maps have been used to both enhance [3] and assess students' conceptual knowledge in a variety of engineering domains. To construct a concept map, students identify and arrange related concepts and use directive and descriptive linking lines to show relationships between those concepts [4–6]. The basic unit of a concept map is a proposition, which includes two concepts joined by a descriptive linking line. Hierarchies are defined by propositions that include the concept map topic. The level of hierarchy is the number of concepts in the longest path down a hierarchy. Cross-links, which are important for representing knowledge connectedness, are descriptive linking lines that create propositions by joining two concepts from different map hierarchies [6, 7].

Consider Fig. 1, which is a concept map about houses. The concept map has four nodes, or concepts—"walls," "floors," "foundations," and "concrete." It also has three hierarchies (A, B, and C) defined by the first-level concepts "walls," "floors," and "foundations," respectively. Hierarchies A and B are Level 1 hierarchies because they only contain one concept, while Hierarchy C is a Level 2 hierarchy because it contains two concepts – "foundations" and "concrete." The proposition "foundations support floors" is a cross-link because it connects concepts from Hierarchies B and C.

As theoretically-based constructs that capture the

* Accepted 21 January 2018.

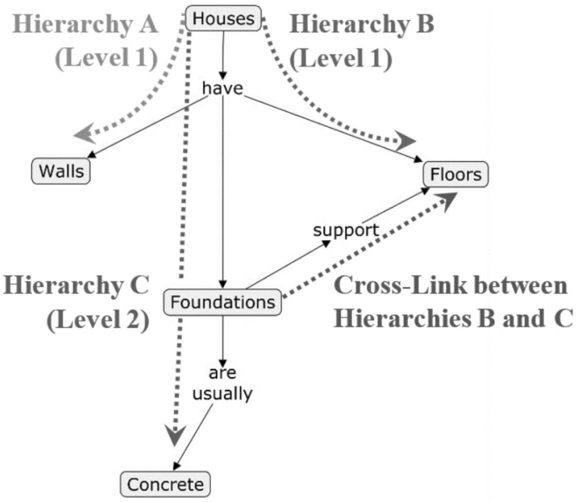

Fig. 1. Basic concept map components (Adapted from [8]).

content and structure of knowledge, concept maps are becoming more commonly used as assessment tools. One major advantage of concept maps is ease of construction, as they are often simpler to create than essays, presentations, or posters [5, 9]. Consequently, the simplicity of concept maps allows students to focus on their understanding of the material, rather than on development of the construct. A second advantage is that they can be used as assessments of ill-defined, rapidly-changing, and/or subjective domain areas where traditional objective assessments (e.g., multiple choice tests) are difficult to develop. For instance, concept maps have been used to assess understanding related to sustainability [10, 11]. Even still, concept maps have been used to assess disciplinary knowledge as a whole, such as in civil [12], industrial [7], and chemical [13] engineering. Despite the advantages of concept maps, their application as assessment tools remains somewhat limited due to the difficult and time-consuming nature of scoring the constructs [7, 8, 14, 15].

The goal of the study was to create and evaluate a computer program to aid in rapid scoring of concept maps, thereby making them more feasible as classroom assessment and research tools. The objectives were to: (1) conduct a comprehensive literature review to inform design of the program, (2) analyze inter-rater reliability of automated and human-generated scores for a sample of concept maps, and (3) analyze correlations between automated and human-generated scores as a measure of convergent validity. Ultimately, the scoring program was developed to interface with CmapTools, a free concept mapping software. Concept maps can be quickly imported, recreated, and analyzed using Python language data structures and the NetworkX software package. The new scoring program can be used to quickly and reliably score concept maps to facilitate assessment of conceptual knowledge in a variety of domains.

2. Review of concept map scoring

While there is much literature on concept maps, a systematic review was conducted on scoring methods, as per Borrego, Foster, and Froyd [16], in order to inform the design of a new automated scoring program.

2.1 Guiding questions and inclusion criteria

The goal of the systematic review was to gather information to create a program that would respond to the needs of the concept map users, while not duplicating existing scoring programs. Consequently, the two guiding questions for the literature review were: (1) What types of scoring methods are most common for analyzing concept maps? (2) What automated programs are currently available?

Several inclusion criteria were specified to aid in identifying records that address the guiding questions: (1) the study uses concept maps to assess understanding in any domain; (2) the study presents a reproducible, quantitative or qualitative, method for analyzing concept maps; (3) the study was published during 1990 to 2016 (without restriction to geographical area); and (4) the study is published in English.

2.2 Searching, screening, and appraising

Several databases were searched in order to discover papers that met the inclusion criteria (Fig. 2). Specifically, Education Resources Information Center (ERIC), Academic Search Complete (ASC), and the American Society for Engineering Education (ASEE) PEER Document Repository were explored. For the ERIC and ASC databases, the search terms used were [concept map AND scoring] present anywhere in the text. Initially, 50 and 26 records were retrieved from the two databases, respectively. Duplicate records were excluded and a total of 67 records were retained for abstract screening. For the ASEE PEER search, the search terms used were [“concept map” + scoring]. Initially, 208 records were identified. Only those with a relevance score of 0.04 or above were retained from the search because preliminary examination indicated that papers with lower relevance scores were not pertinent.

In total, 120 records were retained for abstract screening. Of the ERIC/ASC records, 55 were deemed potentially relevant, while 12 were excluded for not presenting a concept-map-based assessment. When screening the ASEE PEER records, it was

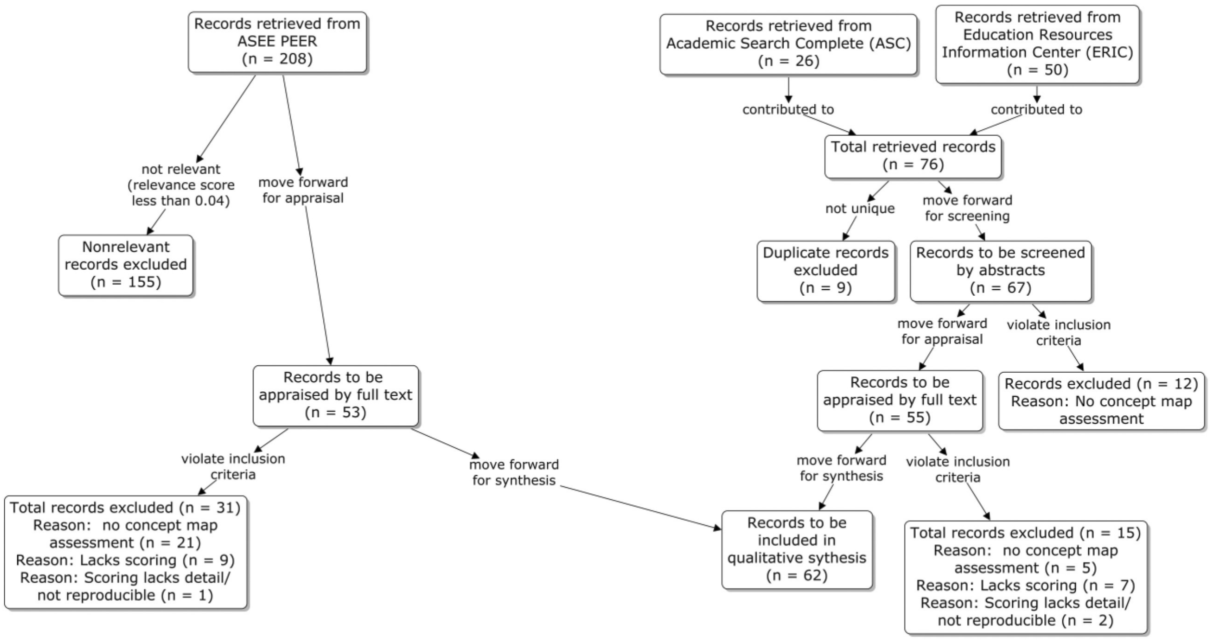
Records retrieved from ASEE PEER (n = 208)
not relevant (relevance score less than 0.04)
move forward for appraisal
Nonrelevant records excluded (n = 155)
Records to be appraised by full text (n = 53)
violate inclusion criteria
move forward for synthesis
Total records excluded (n = 31) Reason: no concept map assessment (n = 21) Reason: Lacks scoring (n = 9) Reason: Scoring lacks detail/ not reproducible (n = 1)
Records retrieved from Academic Search Complete (ASC) (n = 26)
Records retrieved from Education Resources Information Center (ERIC) (n = 50)
contributed to
contributed to
Total retrieved records (n = 76)
not unique
move forward for screening
Duplicate records excluded (n = 9)
Records to be screened by abstracts (n = 67)
move forward for appraisal
violate inclusion criteria
Records to be appraised by full text (n = 55)
Records excluded (n = 12) Reason: No concept map assessment
move forward for synthesis
violate inclusion criteria
Records to be included in qualitative sythesis (n = 62)
Total records excluded (n = 15) Reason: no concept map assessment (n = 5) Reason: Lacks scoring (n = 7) Reason: Scoring lacks detail/ not reproducible (n = 2)

Fig. 2. Summary of search, screening, and appraisal phases of literature review.

noted that many abstracts were either absent or lacking enough detail to compare against inclusion criteria. Consequently, all ASEE PEER records moved forward for full-text appraisal.

Overall, 108 records were appraised by their full texts. Of the ERIC/ASC and ASEE PEER records, 15 and 31 records were excluded, respectively. Reasons for exclusion included lack of a concept-map-based assessment, lack of a scoring method, or presentation of non-reproducible scoring methods. During the initial full-text appraisal, the lead author began extracting key text from each record that would aid in later classification of scoring techniques. From the first round of coding, 13 different groups of scoring methods were outlined.

2.3 Synthesis of scoring techniques

The goal of the synthesis process was to categorize and quantify frequency of application for the numerous scoring methods that were identified in the 62 retained records. First, the previously-extracted excerpts were coded against the 13 pre-identified categories. After this second round of coding, the categories were condensed into eight categories.

2.3.1 Concept map scoring categories

As illustrated in Table 1, scoring methods were classified as quantitative, qualitative, and/or comparison to expert maps, and then assigned to one of eight sub-categories.

Three categories of quantitative scoring methods were identified. First, the ‘‘counting components’’ category captures approaches that count concepts, links, cross-links, and hierarchical properties, many of which were originally proposed by Novak and collaborators [6, 17]. ‘‘Composite metrics’’ encompasses scores that are computed using multiple basic components. For instance, Jablokow et al. [18] computed *complexity*, which is the ratio of total links compared to total concepts. Finally, those scores classified as ‘‘proximity/similarity’’ primarily result from computer algorithms that provide measurements of similarity between links and/or concepts across two sets of concept maps, often based on proximity of nodes [19, 20].

Four categories of qualitative scoring methods were identified. First, the ‘‘holistic rubric/rating’’ category includes methods that require raters to make a single judgement about concept quality. For example, Koul, Clariana, and Salehi [21] applied a holistic rubric developed by Kinchin and Hay [22] that guides raters in making an overall score based on concept map structure (e.g., spoke, chain, net). Second, the ‘‘proposition rating’’ category includes a variety of primary trait rubrics that require judges to rate the quality of propositions using a provided scale. Perhaps the most commonly used was that of Ruiz-Primo and Shavelson [23] who proposed ratings of *excellent*, *good*, *fair*, *don’t care*, and *inaccurate*/*invalid* for proposition scoring. Also, several analytic rubrics, which require judges to use a provided scale to rate a variety of performance dimensions, were identified. A popularly-cited rubric was that of Besterfield-Sacre et al. [7] who presents a three-point scale for rating the

Table 1. Eight categories used for coding of scoring methods

Category	Description/Examples	Sample Citation
Quantitative		
1. Counting components	Number of concepts, Number of cross links	[25]
2. Composite metrics	Density, Complexity	[18]
3. Proximity/similarity	Minimum valence, Pathfinder index/score	[19]
Qualitative		
1. Holistic rubric/rating	Single rating of concept map as a whole	[26]
2. Proposition rating	Rating each individual proposition	[23]
3. Analytic rubric	Rating several concept map performance dimensions	[7]
4. Coding concepts or links	Categorizing concepts and/or linking phrases	[24]
Comparison to Expert Maps	Calculating scores in reference to expert map	[27]

comprehensiveness, *organization*, and *correctness* of concept maps. Also, some authors opted to use emergent [24] or *a priori* categories [8] to code concepts and/or linking descriptions.

Finally, the "comparison to expert maps" category was created to capture methods that compare any score to an expert construct. For example, counting only those student-provided links that also appeared in an expert concept map would be included in both the "traditional" category, as well as the "comparison to expert map" category.

2.3.2 *Final coding of records*

The third round of coding entailed re-reading full-text records and classifying each scoring method against the final eight categories. Within the 62 records that met all inclusion criteria, the most cited methods were those that involved "counting components." For the predominately engineering-related studies published in ASEE PEER, use of the "counting components" category was higher (90.9%) than other records (70.0%) (Table 2). More specifically, 57.5% of all records included a count of nodes, while 45.0% included a count of links (Table 3). Perhaps number of nodes and number of links were common because they are fairly easy to count by hand, even for complex concept maps. Other common components used to analyze concept maps were the number of cross-links (30.6%) and highest hierarchy (29.0%) (Table 3).

Rubrics were also highly cited as scoring tools (Table 2). Most commonly, analytic rubrics were used to guide judges in rating performance dimensions (38.7%). The ASEE PEER records favored the use of analytic rubrics, most commonly the Besterfield-Sacre et al. [7] rubric. Next, proposition rating using primary trait rubrics were also often used by

Table 2. Eight types of scoring methods applied in scoring records

	Percentage (%)		
Scoring Method	Of retained ASC/EWRI records	Of retained ASEE PEER records	Of total retained records
Quantitative			
Counting components	70.0	90.9	77.4
Composite metrics	5.0	22.7	11.3
Proximity/similarity	17.5	9.1	14.5
Qualitative			
Holistic rubric/rating	7.5	13.6	9.7
Proposition rating	30.0	45.5	35.5
Analytic rubric	27.5	59.1	38.7
Coding concepts or links	20.0	31.8	24.2
Comparison to Expert Maps	27.5	18.2	24.2

Table 3. Specific components used to analyze concept maps

	Percentage (%)		
	Of retained ASC/EWRI records	Of retained ASEE PEER records	Of total Retained records
Number of nodes	57.5	54.5	56.5
Number of propositions/links	45.0	36.4	41.9
Number of cross-links	32.5	27.3	30.6
Highest hierarchy	30.0	27.3	29.0
Number of hierarchies	7.5	13.6	9.7

researchers (35.5%), including those from engineering (59.1%). Use of holistic rubrics were among the least cited methods among all records (9.7%), as well as engineering studies (13.6%).

2.4 *Synthesis of scoring programs*

Six concept map scoring programs were discovered in the literature. Overall, these programs were cited in a total of 9 records, which represents 14.5% of all total retained records. All of the programs discovered use either proximity/similarity analysis or proposition rating to score concept maps (Table 4).

2.4.1 *ALA-Mapper*

ALA-Mapper and the Knowledge Network and Orientation Tool for the Personal Computer (KNOT) software are two tools that are used in series to score concept maps. First, ALA-mapper converts concept map data into proximity data, which are often arrays that include links (or lack of links) between all terms or actual pixel distance between terms. Next, KNOT uses the proximity data to create a pathfinder network for each concept map. A pathfinder network is an alternative representation of knowledge structure that resembles a concept map without link labels. Using pathfinder network analysis, metrics can be calculated to compare the relatedness of two concept maps or two groups of concept maps. Often times, student concept maps are compared to expert concept maps. First, common similarity captures the number of links (usually without regard to linking terms) shared by two networks. Second, configural similarity (or neighborhood similarity) is the ratio of the intersection of the two networks compared to the union of the two networks [28]. Four reviewed records used ALA-Mapper (formerly S-Mapper) and KNOT [19, 21, 29, 30].

2.4.2 *CRESST HyperCard®*

CRESST HyperCard® is an applet that allows for individual and collaborative concept mapping that also has a built-in scoring feature. Students construct "closed" concept maps that include only pre-provided concepts. The semantic content score captures the similarity between student and expert links (common similarity), while the organization structure score captures neighborhood (or configural) similarity [31]. Thus, the output from CRESST HyperCard® and KNOT are similar.

2.4.3 *TPL-KATS—concept map*

TPL-KATS is a software that allows for creation and subsequent scoring of concept maps. Students create concept maps using instructor-defined concepts connected either by provided or unique linking phrases. When the instructor creates the concept list, he or she assigns valence values, which describe the strength of association between any two concepts. Thus, expert conceptions about the importance of concept relationships are needed, although this information does not need to be in the form of a concept map. Three types of scores are available for concept maps. The shortest path quantifies the number of links encountered on the shortest path between two concepts. The minimum valence describes the minimum instructor-defined valence score on the shortest path between two concepts, while the average valence describes the mean valence score on the shortest path between two concepts [32].

2.4.4 *Robograder*

Robograder is an automatic scoring feature that is built in to Concept Map Connecter, a concept map creation tool. First, students use the tool to create concept maps using provided concepts. For scoring, instructors must provide a comprehensive set of correct and incorrect propositions along with corresponding scores, ranging from –2 to 2. For instance, a correct proposition can be scored as "superior" (+2) or "acceptable" (+1). One unique feature of Robograder is that it uses an online thesaurus, to "amplify" the scoring matrix [33].

2.4.5 *The concept map tool*

The Concept Map Tool is a web-based tool that includes a rule-based grading system. Students are able to construct concept maps using only the concepts and link labels that are provided in the program by the instructor. Three sets of rules are used to score the concept maps against an expert concept map and provide students with immediate feedback to support iterative improvements in knowledge. First, proposition rating rules award six levels of positive or negative points based on how student propositions compare to expert propositions. Next, path rules are used to award points based on hierarchical structure, as compared to the expert concept map. Finally, set rules are used to compare concepts between the student and expert concept maps. Credit is given for common concepts, while negative credit is awarded for concepts that appear in the expert concept map, but not the student construct [34].

2.4.6 *Lin et al. program*

Lin et al. [35] created a protocol for scoring open-ended concept maps. Similarity Flooding Algorithm (SFA) is used to determine similarity between student and expert concept maps. The premise of SFA is that nodes are similar when their neighboring nodes are similar. During this process, WordNet®-based semantic similarity measurement is used to facilitate appropriate matching of nodes by considering words with similar meanings as matching. When comparing concept maps, one output is "absolute similarity" for the nodes that are considered matches between student and expert constructs.

2.5 Insights for new scoring program

Two important observations were made when comparing the syntheses of scoring methods and programs. First, it was clear that educators commonly use component-based scoring; however, no available program provides automated application of this method. Perhaps the traditional method is applied frequently because it is perceived to be objective and easiest to apply when relying on hand-scoring of concept maps. However, several authors have noted the difficulty of determining highest hierarchy and number of cross-links, which are measures of knowledge depth and connectedness, respectively [8, 36]. The Concept Map Tool includes counting of those concepts that also appear in an expert concept map, as well as a score of hierarchical similarity, although it does not provide estimates of highest hierarchy or number of cross links. As a result, it was decided to build a component-based scoring program to reflect the needs of concept map users.

Second, it was observed that all available programs are restrictive in their definition of "correct" knowledge, either through limiting concept use and/or use of expert concept maps. For instance, only two programs (Pathfinder/ALA Mapper and Lin and colleagues) allow students to include their own concepts during assessments. Furthermore, only two programs (Robograder and TPL-KATS—Concept Map) do not require the use of an expert concept map for scoring. Certainly, there are some domains for which knowledge content is well-defined and closed concept maps may be appropriate. However, one of the benefits of concept maps is the ability to unbiasedly capture student knowledge about broad and ill-defined topics. For some domains, like sustainability for example, no one expert's concept map could be considered absolutely correct, such that all other concept maps should be compared to it. Consequently, the need for open scoring of concept maps, without restriction of concepts or imposition of expert structure, was identified.

3. Development of concept map scoring program

Informed by the literature, it was decided to build a component-based program able to score concept maps without the use of concept restriction or expert concept maps. The traditional method was chosen as the basis for the program [7, 37]. The traditional method counts number of concepts (NC), highest hierarchy (HH), and number of cross-links (NCL) as indicators of knowledge breadth, depth, and connectedness, respectively. Adapted from prior work [17], a weighted scheme was used to calculate the total score: Total = (NH – NC) + 5*(HH) + 10*(NCL) [8].

A concept map scoring program [38] was developed using the Python programming language [39] and NetworkX [40], a Python software package for creating complex networks. The program requires input of concept maps created using CmapTools, a common concept map creation program [41]. It allows for scoring of multiple maps with a single execution. The scoring program is open-source and available on GitHub under the name Cmap-Parse [38].

To score the concept maps, the program first imports the list of concepts and linking phrases and rebuilds the graph as a directed multigraph using the NetworkX software package. NetworkX provides an extensive set of graph algorithms which can be leveraged to analyze and score the concept maps. Also, the root node (or concept map topic) is identified.

Several traditional scores are determined using methods provided in NetworkX. To calculate the

Table 4. Characteristics of available concept map scoring programs

	No	Automated Method(s)	Open concept selection?	Expert cmap?
Pathfinder Network Analysis[1]				
1. Pathfinder/ALA Mapper	4	Proximity/similarity	Yes	Yes
2. CRESST Applet	1	Similarity	No	Yes
Other Programs				
1. Robograder	1	Proposition rating	No	No
2. TPL-KATS—Concept Map	1	Proposition rating; Proximity	No	No
3. Concept Map Tool	1	Proposition rating; Component	No	Yes
4. Lin and colleagues	1	Similarity flooding algorithm	Yes	Yes

[1] DeFranco and Neill [20] also used a pathfinder network analysis; however, the program(s) used were not indicated. Consequently, that record is not included in this count.

number of concepts, the built-in 'order' method is used on the re-created graph which simply returns the number of nodes in the graph. Determining the highest hierarchy is completed by creating a list of all simple paths, or paths without repeated nodes, from the root node to all other nodes in the graph. The highest hierarchy is the longest path from the root node to any other node in the graph.

The number of cross-links is determined using a custom algorithm. To start, each first-level hierarchy concept is labeled with a distinct integer. An integer is also applied to the remaining concepts, based on whichever first-level concept is closest or has the shortest path. For example, if the shortest path from a concept originates from hierarchy labelled number one, then this concept also receives a label of one. Once all concepts are labelled with an integer, the number of cross-links is calculated as the sum of all propositions which span concepts with two different integers. The algorithm for determining the number of cross-links is based on a similar (time-consuming) method for hand-scoring [8].

4. Study methods

Concept maps generated by juniors in an interdisciplinary engineering program at James Madison University were analyzed. The traditional scoring method was applied by two trained judges, as well as the computer program. Agreement and correlation between judges and computer-generated scores was used to quantify the reliability of the program.

4.1 Collecting concept map data

Concept map assessments were used as a direct measure of student sustainability knowledge, as detailed in previous publications [42, 43]. Briefly, students completed a training session on how to construct concept maps and how to use CmapTools. Afterward, students created concept maps on the focus question: "What is sustainability?" Students used CmapTools, a free concept mapping software, to construct and organize their concept maps [41]. A sample student-generated concept map is provided in (Fig. 3), although submissions were visually of varying complexity.

4.2 Scoring concept maps using judges

After submission, two judges examined concept maps. Although the traditional method was designed to be objective, quantification of highest level of hierarchy and number of cross-links proved to be difficult for structurally-complex concept maps with many branches and cross-links, as has been previously reported [36]. For instance, when many cross-links are present, it can be difficult to tell where one hierarchy ends, and one begins. As a result, judges followed a systematic procedure that involved assigning concepts to a hierarchy, identifying cross-links, and counting the level of a hierarchy as the number of concepts in the longest path down a hierarchy [7] until a cross-link is reached (Fig. 4).

The judges applied the scoring method during two phases of scoring. First, judges individually counted the number of concepts, highest hierarchy, and number of cross-links. Afterward, judges compared their scores and discussed any discrepancies in order to arrive at a set of consensus scores. Inter-rater reliability of judges' individual scores was quantified using Krippendorff's alpha, which can be applied to all levels of measurements and any number of judges [44]. All Krippendorff's alpha

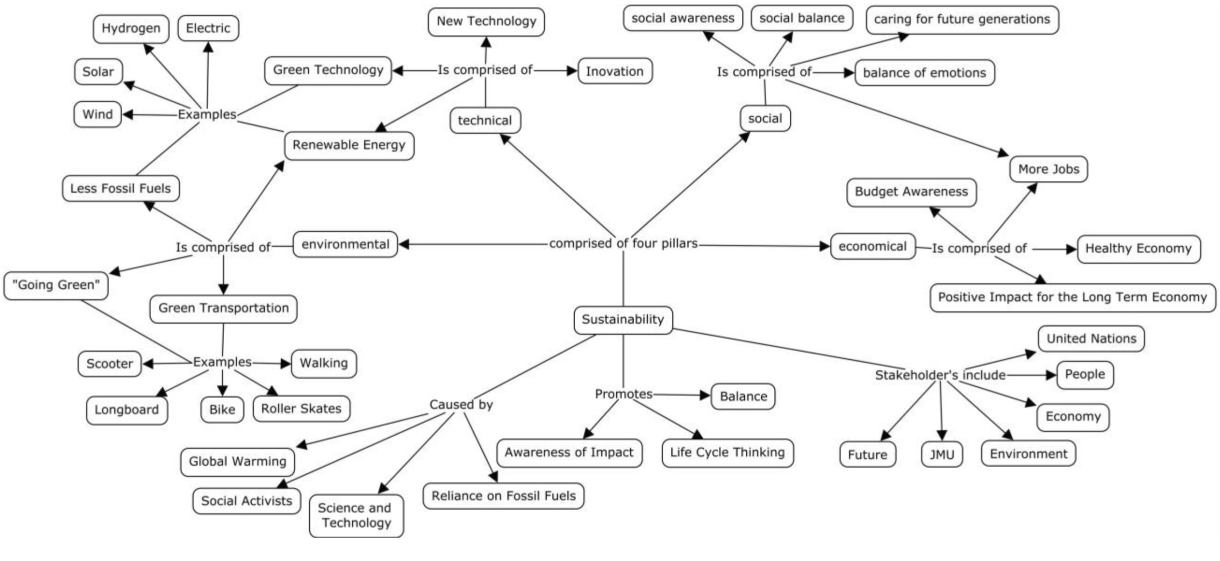

Fig. 3. Sample student-generated, sustainability-focused concept map.

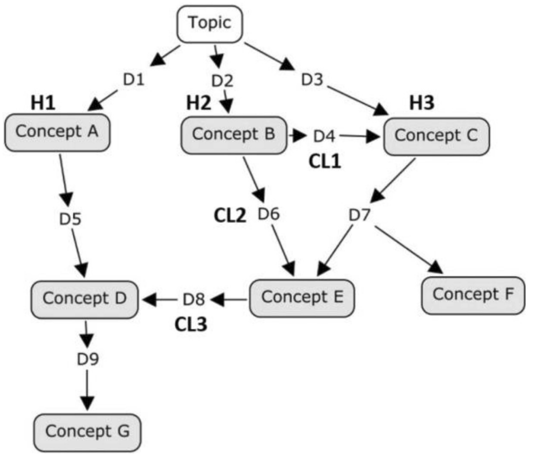

Fig. 4. Application of traditional scoring method. Scores assigned as follows: NC = 7; HH = 3; NCL = 3. Hierarchy 1 (A→D→G) Level = 3; Hierarchy 2 (B) Level = 1; Hierarchy 3 (C→E; C→F) Level = 2.

were above 0.80, which is classified as "adequately acceptable" [44, 45].

4.3 Scoring concept maps using computer program

Student concept maps were also scored using the computer program. Concept maps were exported using the "propositions as text" option in CmapTools, which creates a text file that includes every proposition in the concept map. The text file was then imported into the program and analyzed for number of concepts, highest hierarchy, and number of cross-links.

After initially analyzing the concept maps, the computer program generated a few errors that were addressed. First, some concept maps included concepts that were not connected to any other concept. These stand-alone concepts were deleted from the concept maps before further analysis. Second, some concept maps included concepts that were connected using linking lines that were not connected properly. These incorrect connections were adjusted before further analysis.

4.4 Statistical analysis

Statistical analyses were completed to compare scores generated by judges and the computer program. Inter-rater reliability of judges' consensus scores and the automated scores was determined using Krippendorff's alpha (α) [44, 45]. Scores with Krippendorff's α above 0.80 were designated as "adequately acceptable", while values above 0.67 were classified as "acceptable for exploratory research" [46]. Correlations between manual and automated scores were determined using Spearman's rho (ρ) and used as a measure of convergent validity (especially for highest hierarchy and number of cross-links, which were determined operationally somewhat differently between judges and the computer program). Significant correlations were identified as those exhibiting p-values less than or equal to 0.05. Effect sizes were classified according to Cohen [46] as small ($\rho < 0.30$), medium ($0.50 < \rho \leq 0.30$), or large ($\rho \geq 0.50$).

5. Results

Similar manual and automated scores were determined for concept maps (Table 5). The median number of concepts, which is an indicator of knowledge breadth, was identical for both scoring modes. Furthermore, the number of hierarchies was identical for both scoring modes. Even still, the median number of cross-links, which is an indicator of knowledge connectedness, was identical for both scoring modes. Results for highest hierarchy, which is an indicator of knowledge depth, was very close between the two scoring modes.

Krippendorff's α was used to quantify agreement between the judges and the computer program (Table 6). Adequately acceptable agreement ($\alpha \geq 0.80$) was observed for the number of concepts, number of hierarchies, number of cross-links, and total score. However, below adequate agreement ($\alpha < 0.67$) was observed for the highest hierarchy score.

Table 5. Median scores for concept maps analyzed by judges and the computer program (n = 78)

Score	No. Concepts	No. Hierarchies	Highest Hierarchy	No. Cross-Links	Total Score
Computer	23	4	5	4	97
Judges	23	4	4	4	91.5

Table 6. Agreement between judges' scores and computer program, as measured by Krippendorff's α (n = 78)

	Krippendorff's α	Interpretation
Number of Concepts	0.999	Adequately Acceptable
Number of Hierarchies	0.991	Adequately Acceptable
Highest Hierarchy	0.664	Not Acceptable
Number of Cross-Links	0.925	Adequately Acceptable
Total Score	0.942	Adequately Acceptable

Table 7. Correlation between judges' consensus scores and the computer program, as measured by Spearman's rho (n = 78)

	Spearman's ρ	Effect Size
Number of Concepts	0.999***	Large
Number of Hierarchies	0.990***	Large
Highest Hierarchy	0.742***	Large
Number of Cross-Links	0.958***	Large
Total Score	0.952***	Large

Spearman's ρ was used to quantify the degree of correlation between the judges and the computer program (Table 7). Correlations close to one, which signifies nearly perfect correlation, were observed for number of concepts, number of hierarchies, and number of cross-links. The lowest correlation, although still significant, was observed for highest hierarchy.

6. Discussion

6.1 Comparing judges' and computer program scores

Based on the sample of concept maps analyzed, acceptable agreement was established between most manual and automated traditional scores. Adequately acceptable agreement ($\alpha \geq 0.80$) was observed for number of concepts, number of hierarchies, number of cross-links, and total score. In fact, unacceptable agreement ($\alpha < 0.67$) was observed only for the highest hierarchy sub-score. Despite the low level of agreement, measures of highest hierarchy proved to be highly correlated. Thus, although the manual and automated highest hierarchy sub-scores were not identical, high correlation supports convergent validity. Consequently, it is expected that the manual and automated scores provide different measures of the same construct - knowledge depth.

Differences in highest hierarchy scores between the two modes were due to discrepancies in how this parameter was operationally defined. The highest hierarchy score is intended to measure knowledge depth. During manual scoring a convention was established to make determination of highest hierarchy feasible, especially for complex concept maps. Judges first assigned concepts to a hierarchy and counted the number of concepts in each hierarchy until they reached a cross-link. For example, in Fig. 5A, the highest hierarchy is Hierarchy 2 with four concepts (D → E → F → A). In the manual scoring method, counting of the length of Hierarchy 2 ended with concept A because a cross-link (F → A) had been reached. Thus, the hand-scoring approach essentially captured the *longest hierarchy.* In contrast, the computer program determined the highest hierarchy as the longest path from any first-level concept to any other concept in the map. For example, in Fig. 5B, the highest hierarchy would be six (D → E → F → A → B → C). This, perhaps more correct, definition of highest hierarchy (as *longest path*) was not feasible for hand scoring. Thus, both modes capture knowledge depth, although with different protocols.

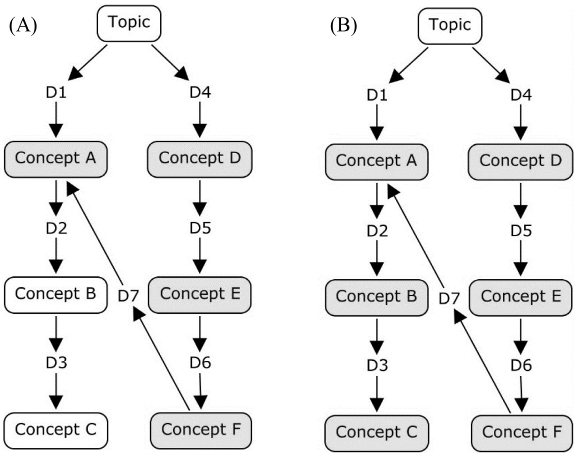

Fig. 5. Knowledge depth using (A) manual and (B) automated scoring.

Although high agreement on the number of cross-links was observed between manual and automated scoring modes, existing discrepancies were due to differences in assigning concepts to hierarchies. For many concept maps, it is possible for a concept to be located under multiple hierarchies (multi-hierarchical concepts). Consequently, assignment of the concept to one hierarchy or another can impact the number of cross-links. For instance, in Fig. 6A, concept F is assigned to hierarchy 2. As a result, there would be two cross-links in the concept map (D → F and E → F). Another alternative is to assign concept F to hierarchy 1, as shown in Fig. 6B. As a result, there would be only one cross-link in the concept map (E → F). While the assignment of concept F only changes the number of cross-links by one, the discrepancy can be greater for more complex concept maps.

As a judge, it can be difficult to determine which hierarchy the student intended a multi-hierarchical concept to belong. For multi-hierarchical concepts, judges occasionally used the context of the concept map to help in concept assignment. For instance, in one concept map, the judges assigned "natural resources" to a hierarchy defined by the first-level concept "environment" rather than the first-level concept "economy." However, this does impose the structure of the judges' own knowledge network onto the concept maps to be scored. In many cases when judges could not reasonably assign a concept to one category over another, the concept map was interpreted from left to right, as is depicted in Fig. 6B. Concept A defines Hierarchy 1 and any concept

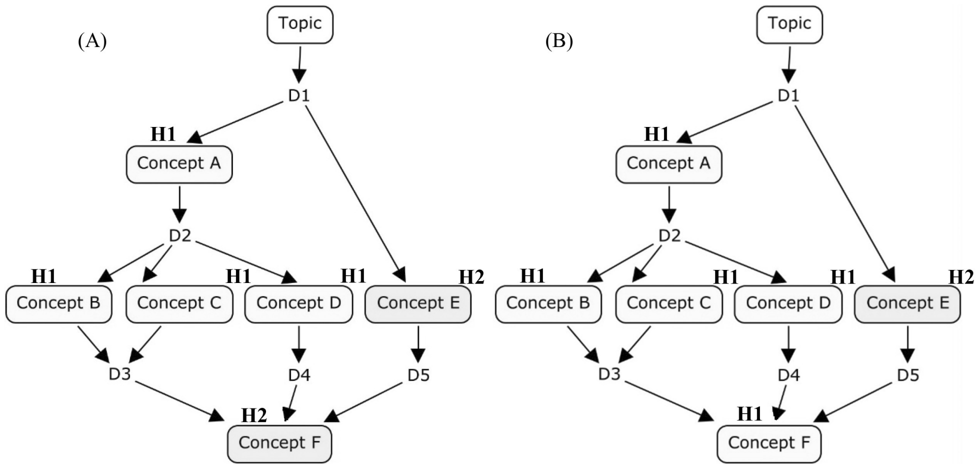

Fig. 6. Scoring a multi-hierarchical concept (concept F)

that exists along a path from concept A is assigned to Hierarchy 1. The computer program, however, does not consider the concept map content at all. Rather, it takes the latter approach in assigning concepts to a hierarchy. Consequently, the computer program overall is more systematic in its scoring of multi-hierarchical concepts..

6.2 Implications for assessment of conceptual knowledge

The computer program evaluated in this study can be used to rapidly and reliably analyze concept maps on a variety of topics to aid in assessment of conceptual knowledge. First, the computer program proved to enumerate most of the traditional sub-scores very similarly to trained judges. Indeed, the automated determination of highest hierarchy (as longest path) is likely a more insightful indicator of knowledge depth than the protocol of convenience operationalized by the judges. Second, the computer program produces more precise results than teams of judges. During hand scoring, different judges may assign concepts to different hierarchies, which can impact quantification of the highest hierarchy and number of cross-links. The computer program, however, will always report the same score for a given concept map. Finally, use of the computer program is much faster than hand scoring. It is estimated that scoring of the current 78 concept maps took each of the two judges four hours to complete. In contrast, it took approximately 10 minutes to convert the concept maps to text files and less than ten seconds for the program to compute traditional sub-scores. Thus, the computer program can aid in fast, reliable, and precise analysis of concept maps.

While the computer program facilitates rapid concept map scoring, quantitative output should be coupled with qualitative analysis to capture a complete view of students' conceptual knowledge. Ultimately, the program allows for quick quantitative measure of knowledge breadth (as number of concepts), knowledge depth (as longest path), and knowledge connectedness (as number of cross-links). However, no insight into the actual content of student knowledge is provide. For instance, when analyzing the sustainability-focused concept maps in the current study, the computer program was unable to determine whether students were including economic, environmental, and/or social concepts in their constructs. Furthermore, the program is unable to make subjective judgements about the correctness of students' concepts and linking lines. For quick content analysis, the authors have used word clouds to quickly compare concept maps between different groups of students [10]. To capture the correctness of knowledge, however, the authors often rely on trained judges. A broader consideration of the advantages and disadvantages of scoring methods is available in an earlier publication [8].

7. Conclusions

A study was conducted to create and examine a new computer program for rapidly analyzing concept maps to capture students' conceptual knowledge in a variety of domains. A comprehensive literature review on available scoring methods and automated programs was conducted to inform design of the new program. A set of student-generated concept maps were scored by two judges and the computer program to evaluate the reliability and validity of the new automated tool. The following conclusions were made based on the results.

1. The literature review supported that a program based on the traditional scoring method would be useful for educators and researchers. The

traditional method uses counts of the number of concepts, highest hierarchy, and number of cross-links as indicators of knowledge breadth, depth, and connectedness, respectively.

2. Acceptable inter-rater reliability was established for number of concepts and number of cross-links, which suggests that both scoring modes can reliably capture knowledge breadth and connectedness, respectively. While both scoring modes reliably captured the number of cross-links, the hand-scoring approach was cumbersome and time-consuming.
3. Low inter-rater reliability and significant Spearman correlations between manual and automated highest hierarchy scores suggest that scoring modes capture different elements of knowledge depth. The judges' protocol for *longest hierarchy* is one of convenience, while the computer program can quickly and systematically capture the *longest path* within the concept map.

Overall, the newly developed computer program can be used to rapidly analyze student conceptual knowledge, as captured in concept maps. Given that one of the primary barriers to use of concept maps is difficulty in scoring, availability of the computer program may make application of concept mapping assessments more feasible. However, given the strictly quantitative output of the program, supplemental scoring methods may still be needed to capture the content and/or correctness of student knowledge. Assessment and improvement of students' conceptual knowledge is important because deep understanding allows engineers to apply knowledge to new situations and ultimately devise innovative designs and solutions. To accomplish this feat, engineers must possess a deep conceptual understanding of engineering fundamentals so that they are able to critically analyze new scenarios and create tailored solutions.

Acknowledgements—Partial support for this study was provided by NSF #1463865 *Developing and assessing engineering students' cognitive flexibility in the domain of sustainable design*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. B. Rittle-Johnson, Promoting transfer: Effects of self-explanation and direct instruction, *Child Development*, **77**(1), 2006, pp. 1–15.
2. J. R. Star, Reconceptualizing procedural knowledge, *Journal for Research in Mathematics Education*, **36**, 2005, pp. 404–411.
3. G. W. Ellis, A. Rudnitsky and B. Silverstein, Using concept maps to enhance understanding in engineering education, *International Journal of Engineering Education*, **20**(6), 2004, pp. 1012–1021.
4. M. A. Ruiz-Primo and R. J. Shavelson, Problems and issues in the use of concept maps in science assessment, *Journal of Research in Science Teaching*, **33**(6), 1996, pp. 569–600.
5. M. Haugwitz, J. C. Nesbit and A. Sandmann, Cognitive ability and the instructional efficacy of collaborative concept mapping, *Learning and Individual Differences*, **20**(5), 2010, pp. 536–543.
6. J. D. Novak and A. J. Cañas, The theory underlying concept maps and how to construct and use them, Technical Report IHMC CmapTools, 2008, http://eprint.ihmc.us/5/2/TheoryUnderlyingConceptMaps.pdf.
7. M. Besterfield-Sacre, J. Gerchak, M. Lyons, L. J. Shuman and H. Wolfe, Scoring concept maps: An integrated rubric for assessing engineering education, *Journal of Engineering Education*, **93**(2), 2004, pp. 105–115.
8. M. K. Watson, J. Pelkey, C. R. Noyes and M. O. Rodgers, Assessing conceptual knowledge using three concept map scoring methods, *Journal of Engineering Education*, **105**(1), 2016, pp. 118–146.
9. J. McClure, B. Sonak and H. K. Suen, Concept map assessment of classroom learning: reliability, validity, and logistical practicality, *Journal of Research in Science Teaching*, **36**(4), 1999, pp. 475–492.
10. E. M. Barrella and M. K. Watson, Comparing the outcomes of horizontal and vertical integration of sustainability content into engineering curricula using concept maps, in W. L. Filho and S. Nesbit (eds), *New Developments in Engineering Education for Sustainable Development*, Springer International Publishing, 2016.
11. M. K. Watson, J. Pelkey, C. Noyes and M. Rodgers, Assessing impacts of a learning-cycle-based module on students' conceptual sustainability knowledge using concept maps and surveys, *Journal of Cleaner Production*, **133**, 2006, pp. 544–556.
12. A. R. Bielefeldt, First-year students' conceptions of sustainability as revealed through concept maps, *American Society for Engineering Education Annual Conference & Exposition*, New Orleans, LA, 2016.
13. S. Muryanto, Concept mapping: An interesting and useful learning tool for chemical engineering laboratories, *International Journal of Engineering Education*, **22**(5), 2006, pp. 979–985.
14. Y. Yin, J. Vanides, M. A. Ruiz-Primo, C. C. Ayala and R. J. Shavelson, Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use, *Journal of Research in Science teaching*, **42**(2), 2005, pp. 166–184.
15. G.-J. Hwang, P.-H. Wu and H.-R. Ke, An interactive concept map approach to supporting mobile learning activities for natural science courses, *Computers & Education*, **57**(4), 2011, pp. 2272–2280.
16. M. Borrego, M. J. Foster and J. E. Froyd, Systematic literature reviews in engineering education and other developing interdisciplinary fields, *Journal of Engineering Education*, **103**(1), 2014, pp. 45–76.
17. J. D. Novak and D. B. Gowin, *Learning How to Learn*, Cambridge University Press, New York, NY, 1984.
18. K. W. Jablokow, J. F. DeFranco, S. S. Richmond, M. J. Piovoso and S. G. Bilén, Cognitive style and concept mapping performance, *Journal of Engineering Education*, **104**(3), 2015, pp. 303–325.
19. R. B. Clariana, R. Koul and R. Salehi, The criterion-related validity of a computer-based approach for scoring concept maps, *International Journal of Instructional Media*, **33**(3), pp. 317–325.
20. J. DeFranco and C. Neill, Improving team performance: The cognitive style factor, *American Society for Engineering Education Annual Conference & Exposition*, Louisville, KY, 2010.
21. R. Koul, R. B. Clariana and R. Salehi, Comparing several human and computer-based methods for scoring concept maps and essays, *Journal of Educational Computing Research*, **32**(3), 2005, pp. 227–239.
22. 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, *Educational Research*, **42**(1), 2000, pp. 43–57.
23. M. A. Ruiz-Primo and R. J. Shavelson, Concept-map based

assessment: On possible sources of sampling variability, Center for Research on Evaluation, Standards and Student Testing, Los Angeles, CA, 1997, https://login.citadel.idm.oclc.org/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED422403&site=ehost-live.

24. K. Oliver, A comparison of web-based concept mapping tasks for alternative assessment in distance teacher education, *Journal of Computing in Teacher Education*, **24**(3), 2008, pp. 95–103.
25. X. Liu, The validity and reliability of concept mapping as an alternative science assessment when item response theory is used for scoring, *Annual Meeting of the American Educational Research Association*, New Orleans, LA, 1994.
26. E. Van Zele, J. Lenaerts and W. Wieme, Improving the usefulness of concept maps as a research tool for science education, *International Journal of Science Education*, **26**(9), 2004, pp. 1043–1064.
27. D. A. Schreiber and G. L. Abegg, Scoring student-generated concept maps in introductory college chemistry, *Annual Meeting of the National Association for Research in Science Teaching*, Lake Geneva, WI, 1991.
28. R. B. Clariana, Deriving individual and group knowledge structure from network diagrams and from essays, in D. Ifenthaler, P. Pirnay-Dummer, and N. M. Steel (eds), *Computer-based diagnostics and systemic analysis of knowledge*, Springer, New York, NY, 2010.
29. R. B. Clariana and E. M. Taricani, The consequences of increasing the number of terms used to score open-ended concept maps, *International Journal of Instructional Media*, **37**(2), 2010, pp. 163–173.
30. J. A. Rye and P. A. Rubba, Scoring concept maps: An expert map-based scheme weighted for relationships, *School Science & Mathematics*, **102**(1), 2002, pp. 33–44.
31. H. E. Herl, H. F. O'Neil, Jr., G. K. W. K. Chung, R. A. Dennis and J. J. Lee, Feasibility of an on-line concept mapping construction and scoring system, *Annual Meeting of the American Education Research Association*, Chicago, IL, 1997.
32. R. M. Hoeft, F. G. Jentsch, M. E. Harper, A. W. Evans Iii, C. A. Bowers and E. Salas, TPL-KATS—concept map: A computerized knowledge assessment tool, *Computers in Human Behavior*, **19**(6), 2003, pp. 653–657.
33. D. Luckie, S. H. Harrison and D. Ebert-May, Model-based reasoning: using visual tools to reveal student learning, *Advances in Physiology Education*, **35**(1), 2011, pp. 59–67.
34. B. E. Cline, C. C. Brewster and R. D. Fell, A rule-based system for automatically evaluating student concept maps, *Expert Systems with Applications*, **37**(3), 2010, pp. 2282–2291.
35. Y. Lin, A. M. Shahhosseini, M. A. Badar, W. T. Foster and J. C. Dean, Using conceptual mapping to help retain tribal knowledge, *American Society for Engineering Education Annual Conference & Exposition*, New Orleans, LA, 2016.
36. D. C. West, J. K. Park, J. R. Pomeroy and J. H. Sandoval, Critical thinking in graduate medical education: A role for concept mapping assessment?, *Medical Education*, **284**(9), 2000, pp. 1105–1110.
37. J. Turns, C. Atman and R. Adams, Concept maps for engineering education: A cognitively motivated tool supporting varied assessment functions, *IEEE Transactions on Education*, **43**(2), 2000, pp. 164–173.
38. J. Pelkey, C*map-Parse*, GitHub, 2016, https://github.com/joshpelkey/cmap-parse
39. Python, https://www.python.org/, accessed 3 January 2018.
40. NetworkX, Available: https://networkx.github.io/, accessed 3 January 2018.
41. A. J. Cañas, G. Hill, R. Carff, N. Suri, J. Lott, G. Gómez, T. C. Eskridge, M. Arroyo and R. Carvajal, CmapTools: A knowledge modeling and sharing environment, *Second International Conference on Concept Mapping*, Pamplona, Spain, 2004.
42. M. Watson, J. Pelkey, C. Noyes and M. Rodgers, Use of concept maps to assess student sustainability knowledge, *American Society for Engineering Education Annual Conference and Exposition*, Indianapolis, IN, 2014.
43. M. K. Watson, Assessment and improvement of sustainability education in civil and environmental engineering, Doctoral Dissertation Georgia Institute of Technology, 2013.
44. A. F. Hayes and K. Krippendorff, Answering the call for a standard reliability measure for coding data, *Communication Methods and Measures*, **1**(1), 2007, pp. 77–89.
45. K. Krippendorff, *Content analysis: An introduction to its methodology*, 2nd edn, Sage Publications Inc., Thousand Oaks, CA, 2004.
46. J. Cohen, A power primer, *Psychological bulletin*, **112**(1), 1992, pp. 155–159.

Mary Katherine Watson received BS and MS degrees in biosystems engineering from Clemson University, Clemson, in 2007 and 2009, respectively. She completed her PhD in civil and environmental engineering at The Georgia Institute of Technology (Georgia Tech), Atlanta, in 2013. She is currently an Associate Professor of civil and environmental engineering at The Citadel in Charleston, SC. Previously, she was a Graduate Assistant in the Center for the Enhancement of Teaching and Learning at Georgia Tech. In the area of engineering education, she has received five best paper awards from the American Society for Engineering Education (ASEE). Also, she has received eight teaching awards. She was named the Young Civil Engineer of the Year by the South Carolina Section of the American Society of Civil Engineers (ASCE). Her technical research interests are in the area of sustainable biotechnology. Dr. Watson is an active member of several organizations. She is currently the chair for the Committee on Effective Teaching for the Civil Engineering Division of ASEE. She is also an active member of ASCE.

Elise Barrella earned a BS in civil engineering from Bucknell University, Lewisburg, PA in 2006. She then received a Master of City and Regional Planning degree in 2008 and a PhD civil engineering (transportation systems) in 2012 from Georgia Tech, Atlanta. She is an Assistant Professor and Founding Faculty Member of the Department of Engineering at Wake Forest University in Winston-Salem, NC. She was Assistant Professor of engineering from 2012–2017 at James Madison University (JMU) in Harrisonburg, VA and prior to that was a Graduate Assistant in the Infrastructure Research Group and at the Center for Quality Growth and Regional Development at Georgia Tech. She has published journal articles, conference proceedings, and book chapters in the areas of transportation systems and engineering education. Dr. Barrella is an active member of ASEE and Women's Transportation Seminar and serves in technical committee leadership roles for Transportation Research Board of the National Academy of Sciences. She is also involved in local government as a volunteer on engineering and planning committees.

Joshua Pelkey earned a BS in computer engineering from Clemson University in 2008. He then received an MS degree in electrical and computer engineering from Georgia Tech, Atlanta in 2010. He is currently a Senior Product Manager at VMware AirWatch in Atlanta, GA. Previously, he held positions of Product Manager and Technical Consultant at

AirWatch. He enables organizations across a wide variety of industries to secure, monitor, manage, and support their entire fleet of mobile assets with cutting-edge enterprise mobility management technology. In the area of engineering education, he specializes in the use of concept maps as assessment tools and has published work in the Journal of Engineering Education and the Journal of Cleaner Production. His technical research interests are in the area of wireless communication and computer network simulation. Mr. Pelkey holds several certifications related to product management. He is a Certified Scrum Product Owner and also holds a Pragmatic Marketing Certification (PMC-III). Related to his work on enterprise mobility, he holds a patent entitled "Enforcement of Proximity Based Policies" (US 9584964 B2).