Scoring guidelines for instructors & researchers

Three dimensions are helpful to understand concept map scoring: instructor uses, kind of componential and holistic criteria (consisting of level and mode), and frames of reference.

Instructor uses

Instructor uses refer to the *actions that instructors perform* to score concept maps. They range on a continuum between qualitative and quantitative instructor uses.

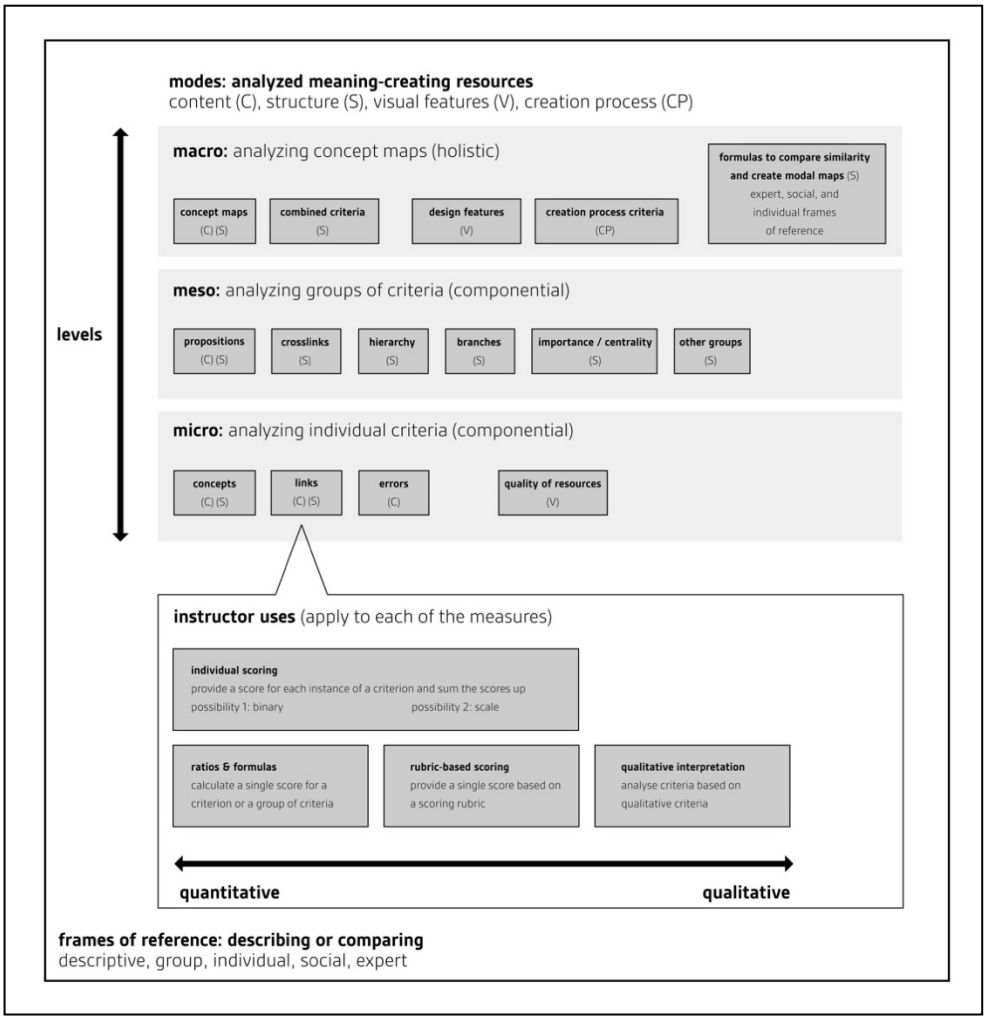

Figure 1: Framework of concept map scoring

Description of the instructor uses for scoring concept maps

Individually scoring each instance of a criterion

The first instructor use is to score the number of instances that a criterion appears, e.g., the number of propositions. These instructor uses provide a specified amount of credit for each

instance. Scoring can be performed by writing the scores next to each instance and then summing the scores up. There are two variations:
a) counting instances of criteria: This variation adds up all instances equally, e.g., by awarding 1 point for each correct proposition (Novak & Gowin, 1984).
b) scoring instances of criteria on a scale: This variation defines a scale with specified criteria and provides a varying amount of credit to each instance. An example would be to provide between 0 and 3 points for each proposition, depending on the quality (Yin et al., 2005).

Ratios and formulas

These instructor uses to score concept maps build on a ratio (e.g., divide the sum of correct propositions by the sum of all propositions; Ruiz-Primo et al., 2001b) or a formula to calculate a single overall score for a group of criteria.

Rubric-based scoring

Rubric-based scoring is applied to all criteria together to determine their overall quality. A rubric defines criteria and dimensions (Hafner & Hafner, 2003).

Qualitative interpretation

Finally, it is also possible to qualitatively evaluate a concept map.

How to decide on an instructor use for scoring concept maps

It is possible to apply each instructor use to each criterion. Often, instructor uses are combined to cover different relevant aspects and get a better overall picture of the quality of a concept map. The following sections include general suggestions for useful considerations before deciding on an instructor use for scoring concept maps. The sections on the different kinds of criteria contain specific information about how these instructor uses were applied to score concept maps in the studies of our systematic literature review.

Is it useful to judge the quality individual instances of criteria?

Sometimes, instructors and researchers are interested in scoring the individual building blocks of a concept map. In these cases, it is useful to decide whether they want to distinguish levels of quality (e.g., from low to high) or distinguish valid vs. invalid instances.
If distinguishing the quality of individual instances of criteria is important, consider *individual scoring on a scale* by defining a scale that sufficiently covers the expected levels of quality (e.g., 0-3 points) together with the criteria that distinguish each of these levels. Our systematic review of scoring criteria for concept maps found that, mostly, three or four levels are used. As a rule of thumb, higher numbers of levels allow for better differentiation of

quality, but also make the scoring procedure more complicated. After deciding on a scale, it should be applied to each instance before the scores are finally summed up.

If distinguishing valid vs. invalid instances of criteria is important, potential differences in the quality are usually discarded. Instead, instructors and researchers count all valid instances of a criterion in a concept map. Invalid instances of a criterion are most often discarded, although they can be included as a criterion of their own if needed (cf. section on errors below). Afterward, the sum of valid instances of a criterion is multiplied with the amount of credit that each instance should receive. Differentiating between the amount of credit allows to give more weight to criteria which are considered more important. For example, Novak and Gowin (1984) propose to score valid examples, propositions, hierarchical levels, and cross-links. They suggest awarding more credit to valid hierarchical levels (around 3 to 10 times) and cross-links (around 2 to 3 times) than to valid propositions, arguing that these criteria reflect progressive differentiation and integrative reconciliation (cf. section on theories of cognition and memory in main article).

Is it useful to score the overall quality?

Sometimes, instructors and researchers do not want to score individual buildings blocks, but instead judge the overall quality of a concept map. These cases are typically covered with rubric-based scoring. A rubric is defined as "*a coherent set of criteria for students' work that includes descriptions of levels of performance quality on the criteria*" (Brookhart, 2013, p. 4, italics in original). These criteria and related levels allow to describe concept maps and act as the foundation of assessment. Rubric-based scoring has the advantage of being quick. Furthermore, summing up individual scores can create similar scores for very different maps (Kinchin et al., 2000).

Is it useful to unite different criteria into a score?

Sometimes, instructors and researchers want to unite different criteria into a single score. Ratio- or formula-based scoring is well-suited for such purposes. For example, the sum of valid propositions compared to the sum of all propositions reflects the accuracy of learners (Ruiz-Primo et al., 2001b).

Criteria for scoring concept maps

There are three *levels* of criteria: micro (small, individual units; the building blocks of a concept map), meso (groups of units; the relations in a concept map), and macro (the entire map as a whole). The levels define the scope of the analysis. Criteria on the micro and meso levels are componential: different components of concept maps are considered independently.

Criteria on the macro level are holistic: concept maps are scored as integrated wholes. Very often, criteria from different levels are combined to get an overall picture of the map's quality. Furthermore, there are different *modes* of criteria. A "mode" describes how the concept map communicates the meaning it conveys (Bezemer & Kress, 2008). For example, text is a written mode of language, a spoken conversation is an oral mode of language, and an image is a visual mode of communication. Modes in concept mapping are:

- content (what a map communicates explicitly)
- structure (information about the layout and connections inside concept maps, often analyzed automatically with the help of graph theory)
- visual (design features of a concept map, e.g., colors and shapes, that could communicate additional information)
- creation (the process of how a map is built, especially useful to consult students)

Micro level

The micro level concentrates on the basic building blocks of a concept map. The criteria on this level can provide interesting insights, but do not consider the full potential and characteristics of concept maps: Concept maps create meaning by relating the building blocks on the micro level to one another, usually in propositions which are assumed to be the "basic unit of meaning" (Ruiz-Primo & Shavelson, 1996, p. 570). Therefore, micro level criteria are typically not used in isolation to fully assess a concept map. However, they can provide valuable information, nonetheless. There are three criteria on the micro level: concepts, links, and errors.

1) Concepts

Concepts are the terms used inside a concept map. They can be provided (learners can only use the terms from a given list), partially provided (learners can use the terms from a given list and add their own terms), or not provided (learners can use their own terms).

Scoring the content of concepts

Criteria that score the content of concepts fall into two varieties: scoring the quality of concepts and scoring categories of concepts. The first variation of concept criteria relates to the quality or accuracy of concepts to communicate the topic of the concept maps, for example by defining key terms that are scored individually (1 point for each mentioned key concept; Wallace & Mintzes, 1990), by defining a holistic scoring scheme of how well

concepts reflect the topic (Romero et al., 2017), or by calculating ratios of essential to secondary concepts (Calafate et al., 2009).

The second variation of concept criteria concentrates on categories of concepts instead of scoring every concept on its own. There are two types of categories of concepts:

***a) Functional categories*:** Functional categories refer to the role that a category of concepts has in the concept map. Examples are the most frequent functional category: They indicate whether a learner knows to what specific objects or events a concept belongs to. They can also indicate whether a learner is capable of transferring abstract knowledge to concrete objects or events. Thus, examples are individually scored with 1 point each in the component scoring approach proposed by Novak and Gowin (1984). Some papers use categories of concepts to identify which concept is the central one (interpreted as the most inclusive; Mendia & García, 2008) or use the central concept to identify the most prominent direct neighbors across a group of students, usually interpreting them as the most important associations of a topic (Wellbrock & Klein, 2014).

***b) Content categories*:** Content categories are semantic. They usually relate to the main interest of the concept mapping task or to different content areas that are relevant for the topic, for example the use of critical concepts (used as a measurement of content validity; Andrews et al., 2008). Categories of concepts are useful when a study is interested in scoring what ratio of concepts belong to a particular topic of interest, for example different areas of sustainability (Segalàs et al., 2010). As an alternative, content categories of concepts can serve as a standardization tool to facilitate comparison when students use various terms for similar content. For example, content categories can be used to redraw student maps with standardized terms (e.g., equating the terms “employee” and “worker”; Freeman & Urbaczewski, 2002). Such an approach is advisable if there is a large set of terms that relate to similar content, for example if learners have selected their own terms that might slightly differ.

Finally, a small number of papers score language features of concepts like counting the number of words or characters of the labels (Wei & Yue, 2017) or spelling (Romero et al., 2017).

Scoring the origin of concepts

In cases where a list of concepts was provided to learners and learners were allowed to add their own concepts, it is possible to score whether relevant concepts come from the provided list or not (DeFranco et al., 2012; Rivard & Straw, 2000).

Scoring the structure of concepts

As a simple criterion of the complexity of a concept map (Ifenthaler, 2010a), instructors and researchers might consider counting the number of concepts, a metric known as "order" of a network in graph theory (Benjamin et al., 2015).

2) Links

Links are the second basic building block of a concept map. They are usually scored structurally without considering the content of the link. Such structural scoring of links typically relies on scoring the number of links, referred to as "size" of a network in graph theory (Benjamin et al., 2015). As an alternative, Calafate et al. (2009) proposed the criterion "degree of meshness" (DM) which they define as the ratio of the total number of links to the minimum number of links.

The content of a relationship is usually scored with propositions. Propositions are semantic units of a link and two or more concepts (Novak & Gowin, 1984). Propositions allow to evaluate the quality of the content described in a relationship because they consider the relationship as a whole, defined by all relevant connected concepts and the link (Taricani & Clariana, 2006).

3) Errors

Errors are often scored implicitly in concept mapping when other criteria concentrate on valid instances (e.g., correct concepts or valid propositions) and discard invalid instances. However, errors can also be important on their own: they can show frequent misconceptions of learners (Kinchin et al., 2000). For example, in formative assessment, instructors could use this information about misconceptions to address specific areas of the topic that learners have not sufficiently mastered. An alternative is to individually score each error that learners made in their concept maps, for example by subtracting a specified number of points per error from the overall score (Terrio & Auld, 2002). Furthermore, instructors and researchers can distinguish different categories of errors, for example incomplete links, superfluous links, or links with reversed directions (Conradty & Bogner, 2008).

Finally, a small number of papers describe concept mapping tasks where learners had to identify correct or incorrect content in a given concept map. Their answers could then be scored individually (e.g., Corrêa et al. awarded 1 point for each correct identification; Corrêa et al., 2018) or as a ratio relating correctly identified errors with missed identification of errors (Correia et al., 2016).

Meso level

The meso-level criteria are groups of units. They are the most prominent criteria in scoring concept maps because they demonstrate the relationships between different components in a concept map – one of the fundamental characteristics of concept maps.

1) Propositions

Propositions are groups of concepts connected by a link. Propositions are the most frequent criterion used to score concept maps. They are referred to as the „basic unit of meaning" (Ruiz-Primo & Shavelson, 1996, p. 570) in concept maps. Most often, these are two concepts and one link, but there might be cases where more elements belong to a proposition. The important aspect is that a proposition is a semantic, meaningful unit. As a rule of thumb, propositions should be as simple and short as possible to fully describe a given semantic relationship (Cañas, 2009). It is very important that learners understand that a link label should explain the type of relationship that exists between concepts: It is the defining aspect of a meaningful proposition. Thus, linking labels can describe static and dynamic relationships between concepts, although concept maps have been criticized for largely focusing on static propositions (Safayeni et al., 2005).

Scoring the content of propositions

The most frequent criteria of scoring propositions are to evaluate their quality, their importance, or their category. Regarding quality of propositions, the typical instructor use is to score each proposition individually for quality, either using a binary choice (valid and invalid propositions) or by using a scale. Thus, instructors and researchers might ask themselves whether they want to concentrate on valid and invalid propositions or whether they need more levels to assess differences in the quality of propositions. If they decide to concentrate on valid and invalid propositions, the most frequent approach is to count all valid propositions and provide them with a specified number of points, mostly 1 point for each valid proposition (Novak & Gowin, 1984). An alternative is a salience score which can be defined as the number of valid propositions divided by the total number of propositions (Ruiz-Primo et al., 2001b), resulting in a ratio of valid propositions.

If instructors and researcher decide to assess differences in the quality of propositions, a range of methods is available:

- defining a scale with respective criteria for scoring the quality of propositions: 0 points for incorrect propositions, 1 point for partially incorrect propositions, 2 points

for correct but unscientific propositions, and 3 points for correct and scientific propositions (Yin et al., 2005)

- applying a "relational scoring" approach: use the flow chart below to assign scores for each proposition (McClure et al., 1999)

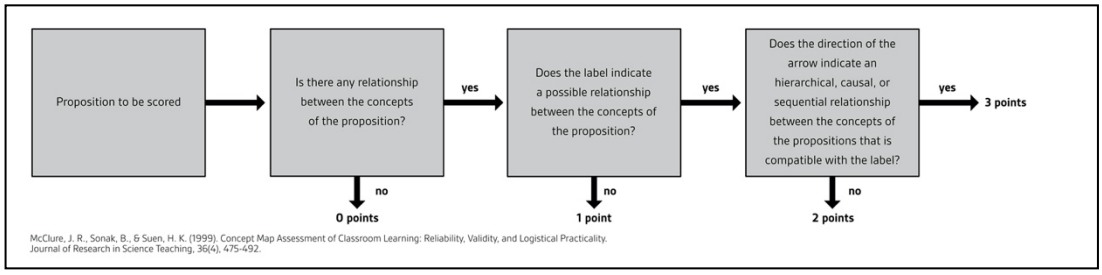

Figure 2: Relational scoring of propositions

- deciding to subtract points for incorrect propositions (Plomer et al., 2010)
- summing up all proposition scores into a "Proposition Accuracy Score" (Ruiz-Primo et al., 2001a)
- combining the quality of propositions in a "Proposition Quality Index" (Reiska et al., 2016):

$$Proposition\ Quality\ Index = 2x\ number\ of\ correct\ propositions + number\ of\ correct, but\ everyday\ propositions - number\ of\ incorrect\ propositions$$

Regarding importance of propositions, the methods described so far assume that all valid propositions are of equal importance. However, some propositions might be more important for understanding a topic than others. In these cases, it is possible to award learners with more points for valid important propositions than for valid less important propositions. One possibility is adding a weight from 0 to 1 to propositions which could then serve as a factor in calculating scores when comparing student map and expert map (Wu et al., 2012). An example of such a weighting factor would be 1 (for essential), 0.75 (for important), 0.5 (for medium importance), and 0.25 (for unimportant). Weighting can also be combined with scoring the quality of propositions. For example, an important good proposition might be scored as following: 0.75 (for "important" weighting factor) x 2 points (for "good" quality) = 1.5 points. Alternatively, it is possible to relate essential propositions to the total number of propositions (Schwendimann, 2014). Weighting propositions is facilitated in digital concept mapping tools that allow to define essential propositions and automatically include the weights as a parameter in calculating the scores (Shui-Cheng et al., 2002).

Regarding categories of propositions, our systematic literature review found two use cases. The first use cases for categories of propositions applies to cases where there is a large set of different propositions, for example when learners created concept maps following a low-directed task. When the task does not limit the set of linking terms, several problems can affect scoring, for example changing the direction of a link by using another linking term or passive constructions, using synonyms, or using general terms instead of specific terms (Strautmane, 2014). In these cases, it might be worthwhile to use categories of propositions for standardizing purposes, e.g., by specifying a set of defined terms (e.g., "is a", "is a part of"; Anohina-Naumeca et al., 2011) or regrouping propositions later (e.g., links indicating a causal relationship like "led to"; Herl et al., 1996).

The second use case for categories of propositions applies to cases where there are content aspects that make it useful to distinguish different areas of content. Examples of these areas are studies of interdisciplinarity of knowledge or studies of sustainability (where different domains are involved). In the following, we want to concentrate on three examples of using categories of propositions in scoring: the interdisciplinarity quality index, the category relevance & complexity index, and using semantic density & semantic gravity to distinguish types of knowledge.

- The *interdisciplinarity quality index* (IQI) was defined by Reiska et al. (2018) to examine interdisciplinarity as visualized in a concept map. It relies on the following steps:
    - classifying concepts into disciplines
    - grouping propositions into *disciplinary* (involve concepts from one discipline) and *interdisciplinary propositions* (involve concepts from different disciplines)
    - rating proposition quality from 0-2 points
    - calculating branch points (concepts that have more than two connections to other concepts)
    - calculating the IQI as follows:

$$\text{IQI} = \frac{\sum \text{correct interdisciplinary concepts}}{\text{maximum of correct interdisciplinary concepts}} + \frac{\sum \text{branch points}}{\text{maximum of branch points}} + \frac{\sum \text{proposition with high quality, scored 2 points}}{\text{maximum of propositions with high quality, scored 2 points}}$$

- The work on *category relevance & complexity index* has been introduced by Segalàs et al. (2010) and has been influential for concept mapping in areas of sustainability using a social frame of reference (cf. section of frames of reference). It relies on the following steps:

- assigning concepts in a concept map about sustainability to the categories "environmental", "social", "economic", and "institutional"
- calculating how many concepts belong to each category:

$$\text{concepts per category} = \frac{\text{number of concepts in a category}}{\sum_{i=1}^{i=\text{number of categories}} \text{number of concepts in a category}}$$

- calculating how many learners have included concepts of each category (in percent): $\text{percentage of learners} = \frac{\text{number of learners who include a certain category}}{\text{number of all learners in a study}}$
- creating a *category relevance* by multiplying these two criteria:

$$\text{cat. relevance} = \frac{\text{concepts per category i x percentage of learners with category i}}{\sum_{i=1}^{i=\text{number of categories}} \text{concepts per category i x percentage of learners with category i}}$$

- calculating the average number of concepts per learner
- using the following formula to calculate how many connections (relatively) exist between different categories:

$$\text{cat. connections} = \frac{\sum_{j=1}^{j=\text{number of learners}} \text{number of propositions between categories (for j)}}{\text{number of categories x number of learners}}$$

- finally, using these values to calculate the *complexity index*:

$$\text{complexity index} = \text{average number of concepts per learner x relative category connections}$$

- Kinchin et al. (2019) describe an instructor use that distinguishes propositions on two semantic dimensions (semantic gravity and semantic density) and used these to create categories of propositions (e.g., high semantic density and low semantic gravity). Afterward, they counted the occurrences of each of these categories and use these numbers to interpret which type of knowledge (e.g., procedural vs. declarative knowledge) is dominant in a concept map.

2) Cross-links

Cross-links are relations between different branches (Hao et al., 2010). They are interpreted as signs of meaningful learning, known as “integrative reconciliation”. Thus, learners start seeing connections between formerly disjunct areas (Ausubel, 1968) and create cross-links to indicate these connections. Therefore, cross-links can also be interpreted as criteria of interdisciplinarity because learners can relate different domains to each other (Himangshu-Pennybacker, 2016). Given their important role in concept maps, cross-links are typically scored higher than regular propositions or hierarchical levels, for example with 10 points for

each crosslink (Novak & Gowin, 1984). An alternative to scoring cross-links individually is to include cross-links in a holistic scoring rubric (Himangshu-Pennybacker, 2016) or to calculate the ratio of cross-links to concepts using the formula "cross-links / concepts x 100", which is interpreted as a criterion of interconnectedness (Martin et al., 2000).

3) Hierarchical levels

The number of hierarchical levels is a frequent criterion that is interpreted as progressive differentiation and integrative reconciliation, typically awarded with more points than a proposition (e.g., 5 points per hierarchical level; Novak & Gowin, 1984). Hierarchical levels are typically counted from the central concept outwards (in network-shaped concept maps) or from the top concept downwards (in hierarchical concept maps). However, chains of linked words do not count as hierarchical levels because they do not indicate structural knowledge (Novak & Gowin, 1984). As alternatives to counting hierarchical levels, our systematic literature review found the following criteria:

- Counting the numbers of concepts per hierarchical level as an estimation of their importance for the overall concept map (Jacobs-Lawson & Hershey, 2002)
- Creating a ratio of the number of concepts divided by the number of hierarchical levels (Schreiber & Abegg, 1991)
- Calculating the "Hierarchical Structure Score" (Brakoniecki & Shah, 2017): sum of the highest hierarchical level ("depth") and the number of concepts ("width") on the largest hierarchical level
- Scoring concepts by awarding a different amount of points depending on the hierarchical level of a concept (Ruben Pierre-Antoine & Mark, 2014)

4) Most important concepts

Sometimes, it is important to consider which concepts are the most important inside their proposition structure. Usually, the central concept is interpreted as the most important or most inclusive concept. It is the start for progressive differentiation and the foundation of meaningful learning (Novak, 2010). Most concept mapping tasks define the central concept as part of the focus question (usually the topic), but in some tasks, learners have to identify it themselves.

However, most approaches that examine the importance of concepts inside their proposition structure belong to the structural scoring approach. Relevant criteria that we identified in our systematic literature review are:

- *Specificity* (Morine-Dershimer, 1993) interprets the importance of concepts based on the content category they belong to. It is calculated as the number of concepts in a particular category divided by total number of concepts.
- *Centrality* (Morine-Dershimer, 1993) is related to the hierarchical levels. It is the number of levels that a particular concept category is away from central concept.
- Another instructor use to measuring the importance of concepts is *degree centrality.* It is calculated by counting the number of direct links that a concept has to other concepts (Clariana et al., 2013). Degree centrality is often calculated without considering the direction of the links in propositions. The rationale behind this decision is that the link direction can easily be changed, for example with using a passive verb or another verb to create the proposition (Krabbe, 2014). For example, the statements "cats and dogs belong to the group of mammals" and "the group of mammals contains cats and dogs" describe the same semantic relationship but would result in propositions with different directions. However, the direction of relationships can be maintained by distinguishing between *in- and outgoing degree* (Shallcross, 2016). In- and outgoing degrees are interpreted as whether a concept tends to be defining other concepts or whether it tends to be defined by other concepts (Reiss & Haussmann, 1990). Furthermore, it is possible to apply structural centrality to the entire concept map, called *graph centrality* which provides information about the shape of a concept map (Clariana et al., 2013).
- Another idea to consider the importance of concepts is to count concepts that act as *bridges* between parts of the map. These bridges can be defined as concepts that, if they were removed from the concept map, would result in splitting the concept map up into different sub maps. Austin and Shore (1995, p. 43) proposed a criterion called "connectivity" which they defined as "minimum number of components whose removal results in a concept map in which no concepts are related". Bernd et al. (2000) mentioned a criterion called "Einzelgewichtigkeit", that is to count how often it appears as a bridge in paths from terminal nodes.

5) Branches

A branch is a sub-tree in a concept map (Hao et al., 2010). It is sometimes used to score concept maps, most notably by counting the number of branches. Branches can also be scored by determining their depth, that is the number of concepts in the longest branch (Bielefeldt, 2016).

6) Other groups

Depending on the particular interest of scoring and the research question of the paper, it is appropriate to include other groups of criteria in scoring, mostly by investigating the structure of a concept map. For example, *cycles or loops* (with 3 or more concepts) can be valuable in systemic thinking because they indicate that learners have understood the dynamic influences on a topic (Luckie et al., 2011). Another example of potentially interesting groups in concept maps is to count the *number of sub-networks* (called "ruggedness" by Eckert; Eckert, 1998) or the *number of isolated concepts* (called "orphans" by Soika and Reiska; Soika & Reiska, 2014). There are also several advanced clustering algorithms available that help in identifying clusters of propositions (McGowen & Davis, 2019; Siew, 2018).

Macro level

Macro-level criteria consider the concept map as a whole. They are useful for a holistic overview of an entire concept map. These instructor uses for scoring concept maps are comparable to evaluating a text.

1) Scoring quality criteria of concept maps holistically

A frequent instructor use is to define a set of criteria that a high-quality concept map should meet. Most often, instructors base their evaluation on a scoring rubric that defines the criteria and levels of scoring (Hafner & Hafner, 2003).

The following list is a selection of rubrics that are frequently applied to score concept maps:

- The semantic scoring rubric focuses on evaluation of content elements by using six quality criteria (Miller & Cañas, 2008).
- The rubric created by Besterfield-Sacre et al. (2004) distinguishes three qualities: criteria comprehensiveness, organization, and correctness.
- The scoring rubric by de Sousa et al. (2019) builds on mixing content and structural quality criteria.

In cases where specific quality criteria are considered more important than others, instructors and researchers can include a weighting factor (Habib & Freiheit, 2007) that favors more important criteria.

Besides scoring of concept maps, rubrics are also useful as guidelines for learners. When provided to learners, rubrics can communicate learning targets and success criteria to learners

(Brookhart, 2013) and allow them to focus on relevant aspects while creating their concept maps. Other alternatives include reviewing the concept maps of others (Chen & Allen, 2017) or self-reflecting on the quality based on the scoring rubrics (Schwendimann & Linn, 2016).

2) Scoring the structure of concept maps

Concept maps can have different structures which can be interpreted in terms of different structural categories (Kinchin et al., 2000):

- *Spoke structures* show a core concept in the middle and various other concepts connected to this central core. They are typically interpreted as learning in progress (students know that various concepts are related, but not how; Hay & Kinchin, 2006) or as rote learning (new knowledge is added, but not meaningfully integrated; Hay et al., 2008).

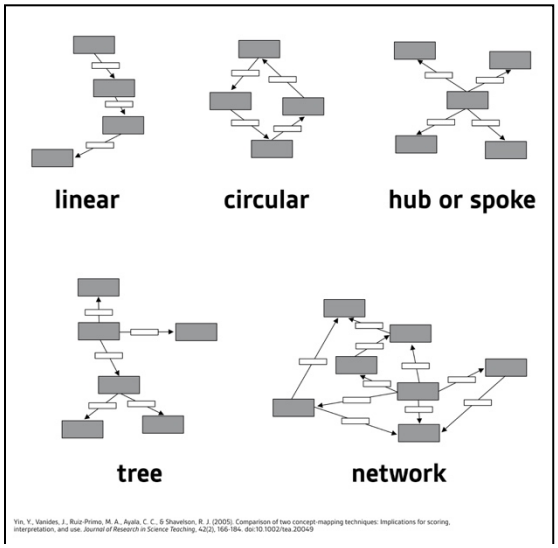

Figure 3: Structural categories (Yin et al., 2005)

- *Chain structures* show sequentially linked concepts. They are typically interpreted as goal-directed learning (Hay & Kinchin, 2006) or non-learning (Hay et al., 2008), because the structure of knowledge largely stays identical.
- *Net structures* show a large variety of links at various hierarchical levels. They are typically interpreted as rich expert knowledge (Hay & Kinchin, 2006) or meaningful learning (Hay et al., 2008).
- Yin et al. (2005) added two categories of concept map structure: *circular structures*, where various concepts are linked together as a loop, and *tree structures*, where linear chains have additional branches.

An important consideration when scoring the structure of concept maps is to decide whether the concept maps should be standardized. Concept maps are typically very individualistic in nature which might make it hard to identify the underlying structures. Thus, Buhmann and Kingsbury (2015) suggested specific rules to standardize concept maps:

- First, removing of all contents to only keep the concepts and links.

- Second, if concepts have several contents (“hunger & poverty”), split them in several concepts (with one content each, “hunger” and “poverty”).
- Third, ordering of the branches following the rules:
    - branch length from longest to shortest
    - number of concepts in branch from largest to smallest (if branches have equal lengths),
    - number of sub-branches from most to fewest (if branches have equal number of concepts)
    - number of cross-links from most to fewest (if branches have equal number of sub-branches)

Standardizing concept maps facilitates comparison. However, the individual aspects communicated in concept maps could themselves be interesting to consider.

Besides scoring of structural categories of concept maps, it is also possible to base structural scoring of entire concept map on the path lengths or distances. In graph theory, *distance* is defined as the number of links on the shortest path between two concepts (Benjamin et al., 2015). Thus, distance can be calculated for any two concepts as a criterion of how closely related they are in a concept map. When using distance for investigating a concept map holistically, instructors and researchers need to define which of the different distances in a concept map should be considered. Our systematic literature review found several approaches to address this issue:

- *diameter* = longest of the shortest paths between the most distant concepts (Ifenthaler, 2010b)
- *average* of shortest paths (Ley et al., 2011)
- the *root of the mean sum* (RMS) of squared distances between all concept pairs (Luckie et al., 2011)
- *radius* = minimal distance between center and farthest terminal concept (Buhmann & Kingsbury, 2015)
- *Structural Complexity Index* (Arneson, 2005) = (average number of propositions in the independent paths through concept map [chain] x number of proposition) + (number of branches x number of independent paths through concept map [chains])

3) Combinations of criteria

Concept maps rely on complex thinking processes. Therefore, instructors and researchers typically use a combination of different criteria to adequately interpret a concept map

(Jonassen et al., 1997). An example of such a combination of different criteria is the scoring approach proposed by Novak and Gowin (1984). They suggest scoring examples as a specific category of concepts, quality of propositions (using a distinction of valid and invalid propositions), hierarchical levels, and cross-links. Furthermore, our systematic literature review found several approaches that combine different criteria in a single formula.

Relating number of concepts to number of links

A frequent approach is creating a *ratio of concepts to links*, often interpreted as a criterion of connectedness (Brakoniecki & Shah, 2017; Hao et al., 2010; Mavers et al., 2002). Related to these are various criteria called "density". They all relate concepts to links, but in slightly different ways:

- Greene et al. (2013) calculated density as

$$density = \frac{\text{total number of links}}{\text{total number of concepts per node}}\ x\ number\ of\ nodes\ (without\ primary)$$

- Richmond et al. (2014) defined density as the total number of links in a student concept map divided by the total number of possible connections:

$$density = \frac{\text{number of links by students}}{\text{number of concepts x (number of concepts} - 1)}$$

- Schwendimann (2014) described two types of density: relative and standardized density. Relative density is the total number of links in a learner concept map divided by the total number of possible connections. Standardized density is the total number of links in a student concept map divided by the total number of connections in an expert map, meaning that this measurement is only possible in an expert frame of reference.
- Ifenthaler (2010a) suggested that the ratio of concepts to links should ideally be in a medium range. He argued that an indication of weak concept map might be that the maps either have too many (connecting everything to everything without meaningfully distinguishing relevant connections) or too few connections (connecting pairs of concepts only).

Nixon et al. (2017)'s version of a "connectedness score" includes clusters ("chunks") of propositions, crosslinks, and the ratio of links to concepts. They define it as (number of clusters + 1) x (number of crosslinks + 1) x (correct links / concepts). The addition of 1 to the number of clusters and crosslinks is to avoid that a concept map without clusters or crosslinks would result in zeroing out the connectedness score.

Relating number of concepts to number of cross-links

Similar to relating the number of concepts to the number of links, it is possible to calculate a *ratio of concepts and cross-links*. This criterion is referred to as "interconnectedness" (Martin et al., 2000), defined as (cross-links / concepts x 100). Interconnectedness is interpreted as a measurement of the cohesiveness of a concept map.

Combining large numbers of criteria into a single formula

Finally, a small number of papers introduce scores combining a large number of criteria into one formula. For example, Roehler et al. (1990) introduced the "extensiveness" criterion that converts the number of concepts, the number of groups of concepts linked to superordinate concepts, the average number of concepts per chunk, and the hierarchical structure into an overall extensiveness score. Hao et al. (2010) suggested a criterion called "EntropyAvg" that is made up of several other criteria like the number of concepts (nodes), the number of branches, and the number of terminal concepts (nodes). They used it to predict students' problem-solving abilities.

Visual criteria

Visual criteria refer to features like color, shape, font, or line thicknesses. They are often not included in concept map scoring, although they can be used meaningfully and communicate important aspects of a concept map (Preston, 2009). Two criteria were found in our systematic literature review: additional resources added to concept maps (e.g., photos; 4 papers) and meaningful use of visual features in concept maps (e.g., colors; 7 papers).

First, additional resources added to concept maps are usually scored individually based on counting (Oliver, 2008) or using pre-defined scales (Schacter et al., 1997). Figueiredo et al. (2004) replaced words with pictures in a concept mapping study with preschool children and assessed them qualitatively.

Second, the meaningful use of visual features in concept maps is rarely scored, with four papers adopting an individual scoring approach to include design features like colors (D'Antoni et al., 2009), two papers proposing a holistic quality parameter covering multimodal features like font, color, or shape (Calafate et al., 2009), and one paper (Preston, 2009) interpreting design features through the lens of semiotic analysis.

Criteria of creation processes

Concept maps are often interpreted as artefacts that learners use to re-represent cognitive structures and thinking processes (Ifenthaler, 2010b). However, digital concept mapping tools also allow to record the creation processes of concept maps, for example using log files (Miller et al., 2008). There is comparatively little research on using these criteria of creation processes for scoring concept maps. Besides basic descriptive criteria like the time spent on a concept mapping task or how often learners used provided help functionalities (Anohina-Naumeca, 2015), researchers have explored using sequential pattern mining (Chiu & Lin, 2011), thinking-aloud data (Ghani et al., 2017), discourse analysis (Roth & Roychoudhury, 1994; Schwendimann & Linn, 2016), proposition generation rate (speed of construction of propositions; Yin et al., 2005), or identifying proposition generation strategies (Yin et al., 2005) to investigate creation processes. Dias et al. (2019) proposed a promising approach of analyzing the construction processes of concept maps using fuzzy inference. Finally, a concept mapping assessment tool described by Anohina-Naumeca et al. (2011) differentiated item types according to their difficulty (with more free item types being more difficult) and used these difficulty degrees in calculating scores (3 papers).

Frames of reference

A frame of reference describes the specific type of comparison that instructors use (Fischbach et al., 2015):

- *descriptive*: not comparing concept maps, but describing them
- *expert*: comparing learner-created concept maps to expert-created concept maps (assumed to represent an idealized representation of the topic, although this assumption is questionable; Ruiz-Primo & Shavelson, 1996; Jonassen et al., 1997; Acton et al., 1994)
- *social*: comparing concept maps created by different learners
- *individual*: comparing concept maps created by the same learner at different points in time
- *group*: taking average values (e.g., means) to evaluate concept maps created by a group (= average out individual differences)

Comparing the content of concept maps

The majority of criteria used to score the content of concept maps is useful for different frames of reference. For example, instructors and educators can score the quality of

propositions to describe a concept map created by a learner or to compare it to another concept map, for example created by an expert, another learner, or by the same learner at another point in time. Thus, criteria from the different concept maps are typically matched against each other. However, our systematic literature review also discovered some noteworthy content criteria that apply to comparative frames of reference. For example, Kornilakis et al. (2004) proposed an approach that defines synonyms and allows for multiple correct answers, enabling digital concept mapping tools to better handle scoring of content criteria. Yao et al. (2006) described a scoring algorithm based on proposition chains. Kao et al. (2008) used a holistic scoring rubric to have both experts and students create scores for concept maps. Afterward, these two scores are compared as a criterion of self-awareness of students.

Comparing the structure of concept maps

Research on methods to structurally compare concept maps is manifold. Our systematic literature review found four families of methods: approaches based on union and intersection of concept maps, correspondence analysis, distances between pairs of concepts, and modal maps.

Closeness Index and other approaches using union and intersection

Regarding the approaches using union and intersection, the most frequent criterion is the *Closeness Index C* by Goldsmith et al. (1991). It compares a student concept map with an expert concept map by "the degree to which a concept has the same neighbors in two different networks" (Acton et al., 1994, p. 306). Three steps are necessary to calculate C.

- The first step is to determine the neighborhood of the first concept in the two compared graphs (= linked concepts). This step is repeated for every concept.
- The second step is to calculate the intersection (concepts linked in both concept maps) and union (the sum of all concepts that are linked in any of the two concept maps) for the first concept. Again, this step is repeated for every concept.
- Finally, the third step consists of calculating the quotient of the sizes of intersection and union for the first concept. Again, this step is repeated for every concept.

C is defined as the mean of the sum of all the quotients derived from step 3. Thus, C takes values between 0 (no similarity) and 1 (identical concept maps).

The Closeness Index C has been highly influential in structural scoring of concept maps and inspired a range of related criteria (Chang et al., 2005; Chen et al., 2001). Furthermore,

alternative criteria are based on the Galanter metric (Fürstenau & Trojahner, 2005), the Jaccard coefficient (Sørmo, 2005) and the Tversky Similarity (Ifenthaler, 2010b).

Correspondence analysis

However, the Closeness Index C and similar criteria based on intersection and union between concepts are not the only approaches to calculate structural similarities between concept maps. Another important approach is *correspondence analysis* (Eckert, 1998). It compares learner and expert concept map with four categories of scores:

- Hits: sum of all concepts, which are connected in both networks;
- Correct rejections: sum of all concepts, which are not connected in both networks;
- False alarms: sum of all concepts, which are connected in the network of interest, but not in the reference network;
- Misses: sum of all concepts, which are connected in the reference network, but not in the network of interest.

Building on these scores, Eckert (1998) defined the most basic correspondence coefficient, likewise called C, as follows:

C = hits + correct rejections – (misses + false alarms) / sum of all possible propositions

Thus, the correspondence coefficient C can take values from -1 (one concept map is the complete opposite of the other) and 1 (identical concept map). Furthermore, Eckert (1998) distinguished different grades of strictness regarding what is considered an error and described an additional weighted correspondence coefficient.

Similarity based on distances

The third group of similarity criteria is based on distances between pairs of concepts, for example using Pathfinder analyses (Schvaneveldt et al., 1989) or multidimensional scaling (Wilson, 1996). Pathfinder is a network built on proximity data that can be derived from having students judge the relatedness of pairs of concepts (Goldsmith et al., 1991) or from counting the number of links between pairs of concepts in a group of concept maps. Thus, Pathfinder networks are very similar to concept maps without link labels (Kim & Clariana, 2015). Wilson (1996) transformed concept maps into a matrix indicating presence (1) or absence (0) of a link between pairs of concepts (without consideration of the link label). In a second step, he used non-metric multidimensional scaling (MDS) to visualize their similarities.

Modal maps

Finally, a modal map is a combined concept map that aggregates the most frequent propositions across all individual concept maps (Fürstenau & Trojahner, 2005). One approach is to specify a threshold value. For example, Ley et al. (2012) used a threshold of 10 %, meaning that propositions were included in the modal map when they were part of at least 10 % of all maps. Alternatively, all propositions from the individual maps are retained in the modal map with their frequencies indicated by the line type of the link (Wellbrock & Klein, 2014). Chen et al. (2001) proposed a method to aggregate concept maps that extends the aforementioned Closeness Index.

However, a problem with modal maps is their artificial character (Fürstenau & Trojahner, 2005): they represent the most common propositions, but have not been created by any participant as such. An important criterion is to calculate the percentage of propositions from a given participant concept map that is present in the modal map ("Abbildungsleistung"; Fürstenau & Trojahner, 2005; Ley et al., 2012).

Scoring concept maps created collaboratively by several learners

The social frame of reference typically compares concept maps created by different learners, for example from the same class. However, concept mapping is increasingly done collaboratively. In these collaborative settings, comparing the similarity of concept maps at different stages can reveal the influence that the collaboration had on the learning outcomes (cf. Fig. 4). Stoyanova and Kommers (2002) and Nomura et al. (2014) described such a collaborative setting. Learners first created an individual concept map ("pre-map"). Then, they collaborated with other learners in a group to create a collaborative concept map. Finally, they again created an individual concept map ("post-map"). In such an approach, instructors and researchers can use the scores from the different concept maps to evaluate the role that the collaborative process has played in learning.

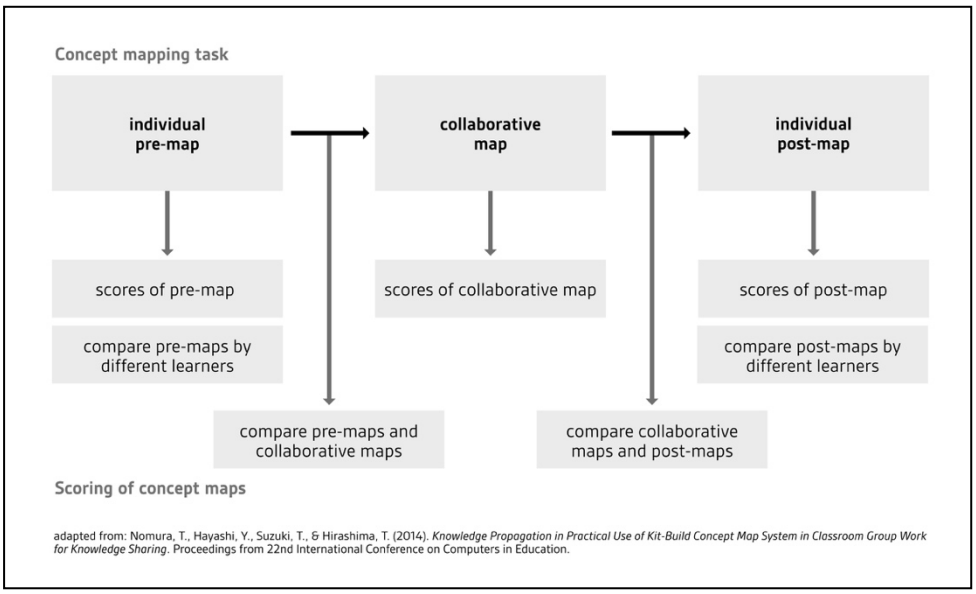

Figure 4: Combination of individual and collaborative concept mapping (Nomura et al., 2014)

Scoring changes in multiple concept maps over time

The individual frame of reference compares concept maps created by the same learner at different points in time, for example for the purpose of formative assessment. Typically, instructors and researchers ask learners to create concept maps before and after an instruction to compare which elements in a concept map have remained (Bernd et al., 2000; França et al., 2004). An alternative is to holistically score categories of changes in two maps from the same student, but at different points in time. For example, Martin et al. (2000) distinguished between three types of changes: restructuring (adding or deleting concepts from first level), accretion (adding ten or more concepts to a concept that already existed = elaboration of existing knowledge), and tuning (changing the meaning of a concept by adding or deleting). In a second step, they counted the frequencies of each of these changes across a group of learners, thus combining individual and social frames of reference. Deshpande and Ahmed (2019) used a scoring rubric to assess the cognitive progression visible in concept maps from different points in time. McGowen and Davis (2019) proposed a method that focuses on the changes between concept maps created by the same person at different points in time. They converted the individual concept maps into schematic diagrams that point out which propositions were moved, added, removed or remained in the same position. In a second step, they added a social frame of reference by qualitatively comparing the changes visible in concept maps from high- and low-gain learners.

References

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural Knowledge Assessment: Comparison of Referent Structures. *Journal of Educational Psychology*, *86*(2), 303-311.

Andrews, K. E., Tressler, K. D., & Mintzes, J. J. (2008). Assessing environmental understanding: an application of the concept mapping strategy. *Environmental Education Research*, *14*(5), 519-536. https://doi.org/10.1080/13504620802278829

Anohina-Naumeca, A. (2015). Justifying the usage of concept mapping as a tool for the formative assessment of the structural knowledge of engineering students. *Knowledge Management & E-Learning*, *7*(1), 56-72.

Anohina-Naumeca, A., Grundspenkis, J., & Strautmane, M. (2011). The concept map-based assessment system: Functional capabilities, evolution, and experimental results. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21*(4), 308-327.

Arneson, B. T. (2005). *On the Role of Concept Mapping Assessments in Today's Constructivist Classroom*. University of Texas].

Austin, L. B., & Shore, B. M. (1995). Using concept mapping for assessment in physics. *Physics Education*, *30*(1), 41-45.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart & Winston.

Benjamin, A., Chartrand, G., & Zhang, P. (2015). *The Fascinating World of Graph Theory*. Princeton University Press.

Bernd, H., Hippchen, T., Jüngst, K.-L., & Strittmatter, P. (2000). Durcharbeiten von Begriffsstrukturdarstellungen in unterrichtlichen und computergestützten Lernumgebungen. In H. Mandl & F. Fischer (Eds.), *Wissen sichtbar machen. Wissensmanagement mit Mapping-Techniken* (pp. 15-36). Hogrefe.

Besterfield-Sacre, M., Gerchak, J., Lyons, M., Shuman, L. J., & Wolfe, H. (2004). Scoring Concept Maps: An Integrated Rubric for Assessing Engineering Education. *Journal of Engineering Education*, 105-115.

Bezemer, J., & Kress, G. (2008). Writing in Multimodal Texts. A Social Semiotic Account of Designs for Learning. *Written Communication*, *25*(2), 166-195.

Bielefeldt, A. R. (2016). *First-Year Students' Conceptions of Sustainability as Revealed through Concept Maps*. ASEE's 123rd Annual, New Orleans, USA: American Society for Engineering Education.

Brakoniecki, A., & Shah, F. (2017). The Use of Concept Maps to Assess Preservice Teacher Understanding: A Formative Approach in Mathematics Education. *Journal of Education*, *197*(1), 23-32.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Buhmann, S. Y., & Kingsbury, M. (2015). A standardised, holistic framework for concept-map analysis combining topological attributes and global morphologies. *Knowledge Management & E-Learning*, *7*(1), 20-35.

Calafate, C. T., Cano, J.-C., & Manzoni, P. (2009). *Improving the Evaluation of Concept Maps: a Step-by-step Analysis*. 20th EAEEIE Annual Conference, Valencia, Spain: IEEE.

Cañas, A. J. (2009). *What are Propositions? …from a Concept Mapping Perspective*. https://cmap.ihmc.us/docs/proposition.php

Chang, K.-E., Sung, Y.-T., Chang, R.-B., & Lin, S.-C. (2005). A New Assessment for Computer-based Concept Mapping. *Educational Technology & Society*, *8*(3), 138-148.

Chen, S.-W., Lin, S. C., & Chang, K. E. (2001). Attributed concept maps: Fuzzy integration and fuzzy matching. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *31*(5), 842-852.

Chen, W., & Allen, C. (2017). Concept Mapping: Providing Assessment of, for, and as Learning. *Medical Science Educator*, *27*(2), 149-153. https://doi.org/10.1007/s40670-016-0365-1

Chiu, C.-H., & Lin, C.-L. (2011). Sequential pattern analysis: Method and application in exploring how students develop concept maps. *Turkish Online Journal of Educational Technology*, *11*(1), 145-153.

Clariana, R. B., Engelmann, T., & Yu, W. (2013). Using centrality of concept maps as a measure of problem space states in computer-supported collaborative problem solving. *Educational Technology Research and Development*, *61*(3), 423-442. https://doi.org/10.1007/s11423-013-9293-6

Conradty, C., & Bogner, F. X. (2008). *Faults in Concept Mapping: A Matter of Technique Or Subject?* A. J. Cañas, P. Reiska, M. Åhlberg, & J. D. Novak (Eds.), Concept Mapping: Connecting Educators. Third Int. Conference on Concept Mapping (pp. 399-405), Tallinn, Estonia & Helsinki, Finland.

Corrêa, R. R., Nascimento, T. S., Ballega, R., & Correia, P. R. M. (2018). *Concept Maps with Errors as an Assessment Task in Elementary School*. Eighth Int. Conference on Concept Mapping (pp. 70-78), Medellín, Colombia.

Correia, P., Cabral, G., & Aguiar, J. (2016). Cmaps with Errors: Why not? Comparing Two Cmap-Based Assessment Tasks to Evaluate Conceptual Understanding. In A. J. Cañas & E. al. (Eds.), *CMC 2016, CCIS 635* (pp. 1-15). Springer. https://doi.org/10.1007/978-3-319-45501-3_1

D'Antoni, A. V., Zipp, G. P., & Olson, V. G. (2009). Interrater reliability of the mind map assessment rubric in a cohort of medical students. *BMC Medical Education*, *9*(1). https://doi.org/10.1186/1472-6920-9-19

de Sousa, L. O., Hay, E. A., & Liebenberg, D. (2019). Teachers' understanding of the interconnectedness of soil and climate change when developing a systems thinking concept map for teaching and learning. *International Research in Geographical and Environmental Education*, *28*(4), 324-342. https://doi.org/10.1080/10382046.2019.1657684

DeFranco, J. F., Jablokow, K. W., Bilen, S. G., & Gordon, A. (2012). *The Impact of Cognitive Style on Concept Mapping: Visualizing Variations in the Structure of Ideas*. ASEE Annual Conference and Exposition, American Society for Engineering Education.

Deshpande, P., & Ahmed, I. (2019). *Topological Scoring of Concept Maps for Cybersecurity Education*. SIGCSE 2019, Minneapolis, MN, USA.

Dias, S. B., Dolianiti, F. S., Hadjileontiadou, S. J., Diniz, J. A., & Hadjileontiadis, L. J. (2019). On modeling the quality of concept mapping toward more intelligent online learning feedback: a fuzzy logic-based approach. *Universal Access in the Information Society*. https://doi.org/10.1007/s10209-019-00656-z

Eckert, A. (1998). *Kognition und Wissensdiagnose. Die Entwicklung und empirische Überprüfung des computerunterstützten wissensdiagnostischen Instrumentariums Netzwerk-Elaborierungs-Technik (NET)*. Pabst.

Figueiredo, M., Sofia Lopes, A., & de Sousa, S. (2004). *"Things We Know About The Cow": Concept Mapping in a Preschool Setting*. First Int. Conference on Concept Mapping, Pamplona, Spain.

Fischbach, A., Brunner, M., Krauss, S., & Baumert, J. (2015). Die Bezugsnormorientierung von Mathematiklehrkräften am Ende der Sekundarstufe I: Konvergenz verschiedener Messverfahren und Wirkung auf motivational- affektive Aspekte des Mathematiklernens und Leistung. *Journal for Educational Research Online*, *7*(3), 3-27.

França, S., D'Ivernois, J. F., Marchand, C., Haenni, C., Ybarra, J., & Golay, A. (2004). Evaluation of nutritional education using concept mapping. *Patient Education and Counseling*, *52*, 183-192.

Freeman, L. A., & Urbaczewski, A. (2002). *Concept Maps as an Alternative Technique for Assessing Students' Understanding of Telecommunications.* International Academy for Information Management (IAIM) Annual Conference: International Conference on Informatics Education Research (ICIER) (pp. 135-145), Barcelona, Spain.

Fürstenau, B., & Trojahner, I. (2005). Prototypische Netzwerke als Ergebnis struktureller Inhaltsanalysen. In P. Gonon, F. Klauser, R. Nickolaus, & R. Huisinga (Eds.), *Kompetenz, Kognition und Neue Konzepte der beruflichen Bildung* (pp. 191-202). VS Verlag für Sozialwissenschaften.

Ghani, I. B. A., Ibrahim, N. H., Yahaya, N. A., & Surif, J. (2017). Enhancing students' HOTS in laboratory educational activity by using concept map as an alternative assessment tool. *Chemistry Education Research and Practice*, *18*, 849-874. https://doi.org/10.1039/c7rp00120g

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing Structural Knowledge. *Journal of Educational Psychology*, *83*(1), 88-96.

Greene, B. A., Lubin, I. A., Slater, J. L., & Walden, S. E. (2013). Mapping Changes in Science Teachers' Content Knowledge: Concept Maps and Authentic Professional Development. *Journal of Science Education and Technology*, *22*(3), 287-299. https://doi.org/10.1007/s10956-012-9393-9

Habib, S. A., & Freiheit, T. I. (2007). *Shop floor modeling with concept maps*. Transactions of the North American Manufacturing Research Institution of SME (pp. 449-456).

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: an empirical study of student peer-group rating. *International Journal of Science Education*, *25*(12), 1509-1528. https://doi.org/10.1080/0950069022000038268

Hao, J.-X., Kwok, R. C.-W., Lau, R. Y.-K., & Yu, A. Y. (2010). Predicting problem-solving performance with concept maps: An information-theoretic approach. *Decision Support Systems*, *48*(4), 613-621. https://doi.org/10.1016/j.dss.2009.12.001

Hay, D., Kinchin, I., & Lygo-Baker, S. (2008). Making learning visible: the role of concept mapping in higher education. *Studies in Higher Education*, *33*(3), 295-311. https://doi.org/10.1080/03075070802049251

Hay, D. B., & Kinchin, I. M. (2006). Using concept maps to reveal conceptual typologies. *Education + Training*, *48*(2/3), 127-142. https://doi.org/10.1108/00400910610651764

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of u.s. history knowledge. *Journal of Educational Research*, *89*(4), 206-218.

Himangshu-Pennybacker, S. (2016). Transforming Science Pedagogy: Using Concept Mapping to Design an Interdisciplinary Approach to Teaching Middle School Science. In *Innovating with Concept Mapping: Communications in Computer and Information Science* (pp. 265-274). Springer International Publishing. https://doi.org/10.1007/978-3-319-45501-3_21

Ifenthaler, D. (2010a). Bridging the Gap between Expert-Novice Differences: The Model-Based Feedback Approach. *Journal of Research on Technology in Education*, *43*(2), 103-117.

Ifenthaler, D. (2010b). Relational, structural, and semantic analysis of graphical representations and concept maps. *Education Tech Research Dev*, *58*(1), 81-97. https://doi.org/10.1007/s11423-008-9087-4

Jacobs-Lawson, J. M., & Hershey, D. A. (2002). Concept Maps As an Assessment Tool in Psychology Courses. *Teaching of Psychology*, *29*(1).

Jonassen, D. H., Reeves, T. C., Hong, N., Harvey, D., & Peters, K. (1997). Concept Mapping as Cognitive Learning and Assessment Tools. *Journal of Interactive Learning Research*, *8*(3), 289-309.

Kao, G. Y.-M., Lin, S. S. J., & Sun, C.-T. (2008). Breaking concept boundaries to enhance creative potential: Using integrated concept maps for conceptual self-awareness. *Computers & Education*, *51*(4), 1718-1728. https://doi.org/10.1016/j.compedu.2008.05.003

Kim, K., & Clariana, R. B. (2015). Knowledge Structure Measures of Reader's Situation Models Across Languages: Translation Engenders Richer Structure. *Technology, Knowledge and Learning*, *20*(2), 249-268. https://doi.org/10.1007/s10758-015-9246-8

Kinchin, I. M., Hay, D. B., & Adams, A. (2000). How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development. *Educational Research*, *42*(1), 43-57. https://doi.org/10.1080/001318800363908

Kinchin, I. M., Möllits, A., & Reiska, P. (2019). Uncovering Types of Knowledge in Concept Maps. *Education Sciences*, *9*(2), 131. https://doi.org/10.3390/educsci9020131

Kornilakis, H., Grigoriadou, M., Papanikolaou, K. A., & Gouli, E. (2004). *Using wordnet to support interactive concept map construction*. IEEE International Conference on Advanced Learning Technologies, 2004. Proceedings., IEEE. http://dx.doi.org/10.1109/icalt.2004.1357485

Krabbe, H. (2014). Digital concept mapping for formative assessment. In D. Ifenthaler & R. Hanewald (Eds.), *Digital Knowledge Maps in Education: Technology-Enhanced Support for Teachers and Learners* (pp. 275-297). Springer.

Ley, S. L., Krabbe, H., & Fischer, H. E. (2012). *Convergent validity: concept maps and competence test for students' diagnosis in Physics*. Fifth Int. Conference on Concept Mapping. Concept Maps: Theory, Methodology, Technology (pp. 149-155).

Ley, T., Schweiger, S., & Seitlinger, P. (2011). *Implicit and Explicit Memory in Learning from Social Software: A dual-process account*. C. Delgado Kloos, D. Gillet, R. M. C. García, F. Wild, & M. Wolpers (Eds.), European Conference on Technology Enhanced Learning (EC-TEL). LNCS 6964. (pp. 449-454), Berlin & Heidelberg, Germany: Springer.

Luckie, D., Harrison, S. H., & Ebert-May, D. (2011). Model-based reasoning: using visual tools to reveal student learning. *Adv Physiol Educ*, *35*(1), 59-67. https://doi.org/10.1152/advan.00016.2010

Martin, B. L., Mintzes, J. J., & Clavijo, I. E. (2000). Restructuring knowledge in Biology: cognitive processes and metacognitive reflections. *International Journal of Science Education*, *22*(3), 303-323. https://doi.org/10.1080/095006900289895

Mavers, D., Somekh, B., & Restorick, J. (2002). Interpreting the externalised images of pupils' conceptions of ICT: Methods for the analysis of concept maps. *Computers and Education*, *38*, 187-207.

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept Map Assessment of Classroom Learning: Reliability, Validity, and Logistical Practicality. *Journal of Research in Science Teaching*, *36*(4), 475-492.

McGowen, M. A., & Davis, G. E. (2019). Spectral analysis of concept maps of high and low gain undergraduate mathematics students. *The Journal of Mathematical Behavior*, *55*, 100686. https://doi.org/10.1016/j.jmathb.2019.01.002

Mendia, E. P., & García, F. M. G. (2008). *Concept maps as a teaching/learning tool in secondary school mathematics. Analysis of an experience*. A. J. Cañas, P. Reiska, M. Åhlberg, & J. D. Novak (Eds.), Concept Mapping: Connecting Educators. Third Int. Conference on Concept Mapping (pp. 268-275), Tallinn, Estonia & Helsinki, Finland.

Miller, N. L., & Cañas, A. J. (2008). *A Semantic Scoring Rubric For Concept Maps: Design and Reliability*. Concept Mapping: Connecting Educators. Third Int. Conference on Concept Mapping, Tallinn, Estonia & Helsinki, Finland.

Miller, N. L., Cañas, A. J., & Novak, J. D. (2008). *Use of Cmaptools Recorder to Explore Acquisition of Skill in Concept Mapping*. A. J. Cañas, P. Reiska, M. Åhlberg, & J. D. Novak (Eds.), Concept Mapping: Connecting Educators. Third Int. Conference on Concept Mapping (pp. 674-681), Talinn, Estonia & Helsinki, Finnland.

Morine-Dershimer, G. (1993). Tracing conceptual change in preservice teachers. *Teaching and Teacher Education*, *9*(1), 15-26. https://doi.org/10.1016/0742-051x(93)90012-6

Nixon, R. S., Hill, K. M., & Luft, J. A. (2017). Secondary science teachers' subject matter knowledge development across the first 5 years. *Journal of Science Teacher Education*, *28*(7), 574-589.

Nomura, T., Hayashi, Y., Suzuki, T., & Hirashima, T. (2014). *Knowledge Propagation in Practical Use of Kit-Build Concept Map System in Classroom Group Work for Knowledge Sharing*. 22nd International Conference on Computers in Education.

Novak, J. D. (2010). *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Routledge.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press. https://doi.org/10.1017/cbo9781139173469

Oliver, K. (2008). A Comparison of Web-Based Concept Mapping Tasks for Alternative Assessment in Distance Teacher Education. *Journal of Computing in Teacher Education*, *24*(3).

Plomer, M., Jessen, K., Rangelov, G., & Meyer, M. (2010). Teaching physics in a physiologically meaningful manner. *Physical Review Special Topics - Physics Education Research*, *6*(2). https://doi.org/10.1103/physrevstper.6.020116

Preston, C. J. (2009). Exploring Semiotic Approaches to Analysing Multidimensional Concept Maps Using Methods that Value Collaboration. In *Handbook of Research on Collaborative Learning Using Concept Mapping* (pp. 256-282). IGI Global. https://doi.org/10.4018/978-1-59904-992-2.ch013

Reiska, P., Möllits, A., & Rannikmäe, M. (2016). Enhancing the Value of Active Learning Programs for Students' Knowledge Acquisition by Using the Concept Mapping Method. In *Innovating with Concept Mapping: Communications in Computer and Information Science* (pp. 83-97). Springer International Publishing. https://doi.org/10.1007/978-3-319-45501-3_7

Reiska, P., Soika, K., & Cañas, A. J. (2018). Using concept mapping to measure changes in interdisciplinary learning during high school. *Knowledge Management & E-Learning*, *10*(1), 1-24.

Reiss, M., & Haussmann, K. (1990). Deklarative Wissensdiagnostik im Bereich rekursiven Denkens. In K. Haussmann & M. Reiss (Eds.), *Mathematische Lehr-Lern-Denkprozesse* (pp. 131-151). Hogrefe.

Richmond, S. S., DeFranco, J. F., & Jablokow, K. W. (2014). A Set of Guidelines for the Consistent Assessment of Concept Maps. *International Journal of Engineering Education*, *30*(5), 1072-1082.

Rivard, L. P., & Straw, S. B. (2000). The Effect of Talk and Writing on Learning Science: An Exploratory Study. *Science Education*, *84*, 566-593.

Roehler, L. R., Duffy, G. G., Conley, M., Herrmann, B. A., Johnson, J., & Michelsen, S. (1990). Teachers' Knowledge Structures: Documenting Their Development and Their Relationship to Instruction. *Research Series No. 192*.

Romero, C., Cazorla, M., & Buzón, O. (2017). Meaningful learning using concept maps as a learning strategy. *Journal of Technology and Science Education*, *7*(3), 313. https://doi.org/10.3926/jotse.276

Roth, W., & Roychoudhury, A. (1994). Science discourse through collaborative concept mapping: new perspectives for the teacher. *International Journal of Science Education*, *16*(4), 437-455.

Ruben Pierre-Antoine, S. D. S., & Mark, S. (2014). *Utilizing Concept Maps to Improve Engineering Course Curriculum in Teaching Mechanics*. ASEE Annual Conference & Exposition, Indianapolis, Indiana: ASEE Conferences.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001a). Comparison of the Reliability and Validity of Scores from Two Concept-Mapping Techniques. *Journal of Research in Science Teaching*, *38*(2), 260-278.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and Issues in the Use of Concept Maps in Science Assessment. *Journal of Research in Science Teaching*, *33*(6), 569-600.

Ruiz-Primo, M. A., Shavelson, R. J., & Schultz, S. E. (2001b). On the Validity of Cognitive Interpretations of Scores From Alternative Concept-Mapping Techniques. *Educational Assessment*, *7*(2), 99-141.

Safayeni, F., Derbentseva, N., & Cañas, A. J. (2005). A theoretical note on concepts and the need for Cyclic Concept Maps. *Journal of Research in Science Teaching*, *42*(7), 741-766. https://doi.org/10.1002/tea.20074

Schacter, J., Herl, H. E., Chung, G. K. W. K., O'Neil, H. F. O., Dennis, R. A., & Lee, J. J. (1997). Feasibility of a Web-Based Assessment of Problem Solving. *ERIC Document Reproduction Service No. ED 410 255*.

Schreiber, D. A., & Abegg, G. L. (1991). Scoring Student-Generated Concept Maps in Introductory College Chemistry. *ED 347 055*.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 249-284). Academic Press.

Schwendimann, B. A. (2014). Making Sense of Knowledge Integration Maps. In D. Ifenthaler & R. Hanewald (Eds.), *Digital Knowledge Maps in Education: Technology-Enhanced Support for Teachers and Learners* (pp. 17-40).

Schwendimann, B. A., & Linn, M. C. (2016). Comparing two forms of concept map critique activities to facilitate knowledge integration processes in evolution education. *Journal of Research in Science Teaching*, *53*(1), 70-94. https://doi.org/10.1002/tea.21244

Segalàs, J., Ferrer-Balas, D., & Mulder, K. F. (2010). What do engineering students learn in sustainability courses? The effect of the pedagogical approach. *Journal of Cleaner Production*, *18*(3), 275-284. https://doi.org/10.1016/j.jclepro.2009.09.012

Shallcross, D. C. (2016). Concept Maps for Evaluating Learning of Sustainable Development. *Journal of Education for Sustainable Development*, *10*(1), 160-177. https://doi.org/10.1177/0973408215625551

Shui-Cheng, L., Kuo-En, C., Yao-Ting, S., & Gwo-Dong, C. (2002). *A new structural knowledge assessment based on weighted concept maps*. International Conference on Computers in Education, 2002. Proceedings., IEEE Comput. Soc. http://dx.doi.org/10.1109/cie.2002.1186041

Siew, C. S. Q. (2018). Using network science to analyze concept maps of psychology undergraduates. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3484

Soika, K., & Reiska, P. (2014). *Assessing Student's Interdisciplinary Approach with Concept Mapping*. Concept Mapping to Learn and Innovate. Sixth Int. Conference on Concept Mapping (pp. 71-79), Santos, Brazil.

Sørmo, F. (2005). *Case-Based Student Modeling Using Concept Maps*. H. Muñoz-Avila & F. Ricci (Eds.), ICCBR 2005, LNAI 3620 (pp. 492-506), Heidelberg, Germany: Springer.

Stoyanova, N., & Kommers, P. (2002). Concept Mapping as a Medium of Shared Cognition in Computer-Supported Collaborative Problem Solving. *Journal of Interactive Learning Research*, *13*(1/2), 111-133.

Strautmane, M. (2014). *Increasing the flexibility of automated concept map based knowledge assessment*. 15th International Conference on Computer Systems and Technologies - CompSysTech '14, New York, New York, USA: ACM Press. http://dx.doi.org/10.1145/2659532.2659621

Taricani, E. M., & Clariana, R. B. (2006). A Technique for Automatically Scoring Open-Ended Concept Maps. *ETR&D*, *54*(1), 65-82.

Terrio, K., & Auld, G. W. (2002). Osteoporosis knowledge, calcium intake, and weight-bearing physical activity in three age groups of women. *Journal of Community Health*, *27*(5), 307-320.

Wallace, J. D., & Mintzes, J. J. (1990). The concept map as a research tool: Exploring conceptual change in biology. *Journal of Research in Science Teaching*, *27*(10), 1033-1052.

Wei, W., & Yue, K.-B. (2017). Integrating Concept Mapping into Information Systems Education for Meaningful Learning and Assessment. *Information Systems Education Journal (ISEDJ)*, *15*(6), 4-13.

Wellbrock, C.-M., & Klein, K. (2014). Journalistische Qualität – eine empirische Untersuchung des Konstrukts mithilfe der Concept Map Methode. *Publizistik*, *59*(4), 387-410. https://doi.org/10.1007/s11616-014-0212-6

Wilson, J. (1996). Concept Maps about Chemical Equilibrium and Students' Achievement Scores. *Research in Science Education, 26*(2), 169-185.

Wu, P.-H., Hwang, G.-J., Milrad, M., Ke, H.-R., & Huang, Y.-M. (2012). An innovative concept map approach for improving students' learning performance with an instant feedback mechanism. *British Journal of Educational Technology*, *43*(2), 217-232. https://doi.org/10.1111/j.1467-8535.2010.01167.x

Yao, Q., Yang, K., Zhao, G., & Huang, R. (2006). *A Concept Mapping Scoring Algorithm Based on Proposition Chains*. J. D. Novak & A. J. Cañas (Eds.), Concept Maps: Theory, Methodology, Technology. Second Int. Conference on Concept Mapping (pp. 8-15), San José, Costa Rica.

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J. (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*, *42*(2), 166-184. https://doi.org/10.1002/tea.20049