C-R-E-S-S-T

National Center for Research on Evaluation, Standards, and Student Testing

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

On Concept Maps as Potential "Authentic" Assessments in Science

CSE Technical Report 388

Richard J. Shavelson, Heather Lang, and Bridget Lewin
CRESST/University of Santa Barbara

August 1994

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

ON CONCEPT MAPS AS POTENTIAL "AUTHENTIC" ASSESSMENTS IN SCIENCE

Richard J. Shavelson, Heather Lang, and Bridget Lewin
CRESST/University of California, Santa Barbara

Abstract

The search for new, "authentic" science assessments of what students know and can do is well underway. This search has unearthed measures of students' "hands-on" *performance* in carrying out science investigations. And now it has been expanded to discover more or less direct measures of students' *knowledge structures*. One potential "find" is "concept mapping." A concept map, constructed by a student, is a graph consisting of nodes representing concepts and labeled lines denoting the relation between a pair of nodes (concepts). The external concept map constructed by a student is interpreted as representing important aspects of the organization of concepts in that student's memory ("cognitive structure"). In this review, concept mapping techniques were found to vary widely. No less than 128 possible variations were identified; for example, hierarchical versus nonhierarchical maps, student-generated or map-provided concept terms, student-generated or map-generated spatial structure, single- or multiple-concept links. Methods for scoring maps varied almost as widely, from the admonition "don't score maps" to a detailed scoring system for hierarchical maps. The review led to the conclusion that an integrative, "working" cognitive theory is needed to begin to limit this variation for assessment purposes. Such a theory would also serve as a basis for much-needed psychometric studies of the reliability (generalizability) and construct validity of concept maps since such studies are almost nonexistent in the literature. Issues of reliability include generalizability of scores over raters, samples of concept terms, and test occasions. Issues of validity include the representativeness of the sample of concept terms in the map of the subject domain as a whole, the correspondence between the concept map representation and representations based on other measures of cognitive structure, the exchangeability of one mapping technique with another, and the impact of the technique on the performance of diverse groups of students. Finally, this review sets forth issues arising from large-scale use of mapping techniques, including the importance of students' familiarity with and skills in using concept maps, and the possible negative impact of teachers teaching to the assessment if students have to *memorize* concept maps provided by textbooks or themselves.

ON CONCEPT MAPS AS POTENTIAL "AUTHENTIC" ASSESSMENTS IN SCIENCE

Richard J. Shavelson, Heather Lang, and Bridget Lewin
CRESST/University of California, Santa Barbara

"Authentic" assessments are intended to provide evidence about what students *know* and *can do* in a subject matter. Performance assessments in science (e.g., Shavelson, Baxter, & Pine, 1991), for example, yield evidence about what students can do when provided a problem and a "laboratory" to carry out an investigation to solve the problem (e.g., Shavelson et al., 1991). Interpretations of performance assessment scores, however, usually make large inferential leaps. One such leap goes from the investigation at hand to a broad domain of possible investigations that could have been used in the assessment (Shavelson, Baxter, & Gao, 1993). A far bigger inferential leap goes from observed performance to cognitive processes or "higher order thinking" used by the student in carrying out the investigation (e.g., Resnick & Resnick, 1992; Wiggins, 1989). While research has shown that multiple investigations are needed to draw inferences to a broader domain of investigations (e.g., Shavelson et al., 1993), unfortunately precious little research is being conducted to see if such inferential leaps from performance to cognition can be supported empirically.[1]

The insistence on interpretations that go beyond statements about *levels of performance* (no small accomplishment in itself) to provide insights into what students know, and how that knowledge is represented and used, has led to a search for more cognitively oriented assessments than performance assessments. This search has identified one set of techniques that provide more or less direct measures of students' *knowledge structures*. These techniques are grounded on the assumption that understanding in a subject domain such as science is to conceive a rich set of relationships among important concepts in that domain. Some assessment techniques probe students' perceptions of the concept

---

[1] Notable exceptions are Baxter, Glaser, and Raghavan (1994) and Magone, Cai, Silver, and Wang (in press).

interrelations in a science domain *indirectly* by eliciting their (a) word associations to concepts (e.g., Shavelson, 1972, 1974), (b) judgments of similarity between pairs of concepts (e.g., Goldsmith, Johnson, & Acton, 1991), or (c) categorizations of concepts into groups based on similarity (cf. Shavelson & Stanton, 1975). Other techniques, namely, concept maps, probe perceived concept relatedness more *directly* by having students build graphs or trees and make explicit the nature of the links between concept pairs.[2]

The virtue of both techniques is that, unlike other assessments of student cognition such as talk-aloud interviews (e.g., Ericsson & Simon, 1984) or dynamic assessment (Brown & Ferrara, 1985), once students understand the process of the task, maps can be used with large numbers of students in short periods of time without intensive adult involvement (White, 1987).

The purpose of this paper, in broad terms, is to examine the validity of claims that concept maps measure an important aspect of students' knowledge structure in a subject domain, namely, science. To this end, we (a) provide a working definition of concept mapping and characterize the variety of measurement techniques that are commonly used to gather and score concept map data; (b) evaluate the cognitive theory that underlies the use of concept maps; (c) review empirical evidence on reliability and validity of various concept mapping techniques, identifying areas for psychometric research; and (d) set forth possible consequences of using the techniques in high-stakes, large-scale assessment.

Before turning to this examination of concept maps, a caveat is in order. Concept maps have frequently been used extensively as instructional tools. They have been used less frequently as pre- and post-instruction measures to evaluate learning and instruction. They have been used very infrequently as formal assessments. Not surprisingly, then, the literature linking cognitive theory and psychometric characteristics of concepts maps is sparse.

---

[2] There is some controversy as to whether concept maps can be interpreted as measures of cognitive structure, the notion being that subjects are not aware of their cognitive structures and so more indirect methods such as word association or similarity judgment need to be used (e.g., Goldsmith et al., 1991).

Concept Maps

A concept map is a graph consisting of nodes and labeled lines. The nodes correspond to important terms (standing for concepts) in the domain.[3] The lines denote a relation between a pair of concepts (nodes). And the label on the line tells how the two concepts are related. The combination of two nodes and a labeled line is often called a *proposition*. Concept maps, then, purport to represent some important aspect of a student's propositional knowledge in a subject domain.

White and Gunstone (1992, pp. 17-18) recommended a series of steps in concept mapping. For the teacher (or assessment developer), the steps are:

1. **Select a set of concept terms.**

2. **Provide students with concept terms on 3 x 5 inch index cards and a sheet of paper.**

3. **Give the following instructions to students, one step at a time.[4]**

    a. **Sort through the cards, and put to one side any that have a term you don't know or which you think is not related to any other term.**

    b. **Put the remaining cards on the sheet of paper, and arrange them in a way that makes sense to you. Terms you see as related should be kept fairly close together, but leave space between even the closest cards.**

    c. **When you are satisfied with your arrangement of the cards, stick them to the sheet. [So, for example, a student at this stage might have a sheet that looks like Figure 1.]:**

---

[3] Formally, terms or words used in concept mapping are not concepts. They stand for concepts. Nevertheless, the terms used in concept mapping are called "concepts" and from here on out, we will follow this convention.

[4] These can be oral, or, better yet, displayed on a board or screen.

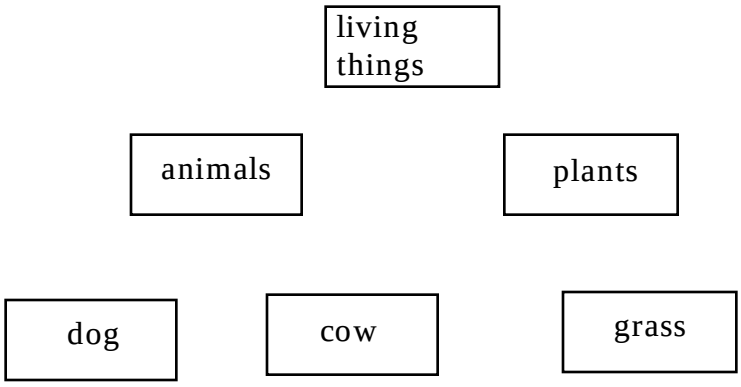


***Figure 1*. Concept map at the stage of arranging terms (White & Gunstone, 1992, p. 17).**

d. Draw lines between the terms you see to be related.

e. Write on each line the nature of the relation between the terms. It can help to put an arrowhead on the line to show how to read the relation. [Now the student might have a sheet like Figure 2.]:

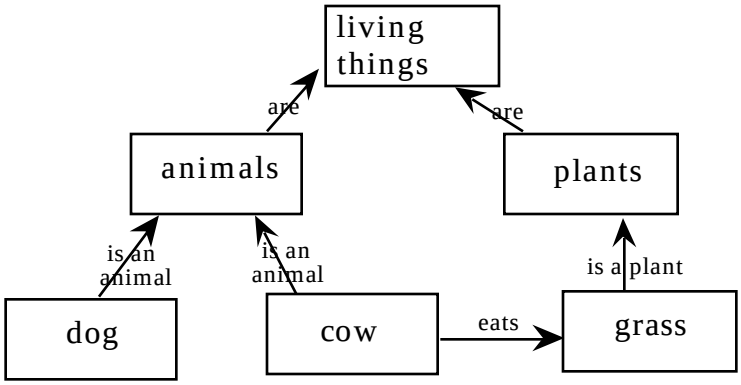


***Figure 2*. The completed map (White & Gunstone, 1992, p. 18).**

f. If you put any cards to one side in the first step [a], go back to these and see if you now want to add any to the map. If you do add any make sure you write the nature of the links between them and the other terms.

An example of a concept map (White & Gunstone, 1992, p. 16), constructed by an 11-year-old with concept mapping experience, is shown in Figure 3. The student was given the seven concepts shown in the map on 3 x 5 inch cards and asked to sort out those cards with terms she didn't know. Next, the student was asked to arrange the remaining cards on a piece of paper in any way that made sense to her. Then she was asked to draw lines between the terms that she perceived to be related and write on each line the nature of the relationship. Finally, she was asked to incorporate any of the cards placed aside into the map, drawing and labeling the line connecting the term to other terms in the map.

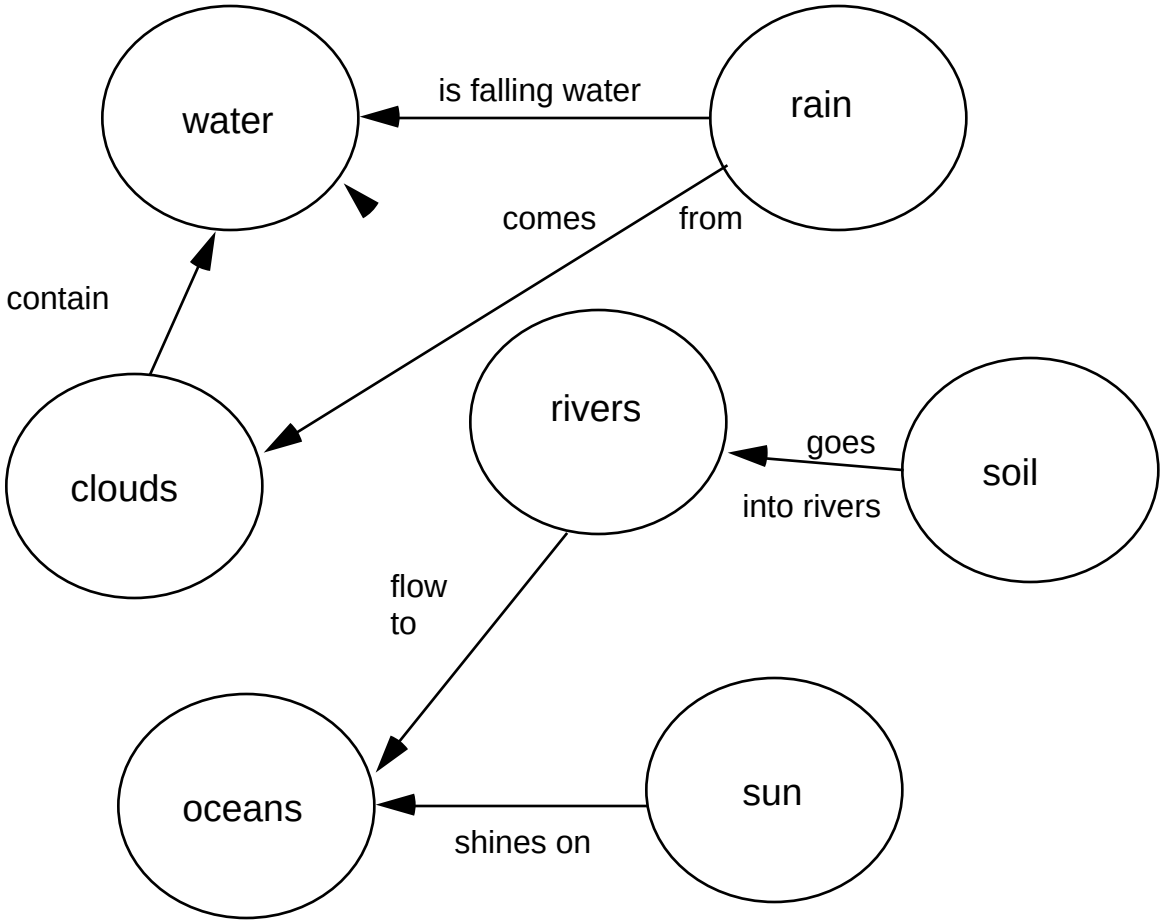


***Figure 3*. Eleven-year-old's concept map (White & Gunstone, p. 16).**

Variations in Concept Mapping Tasks

Concept mapping tasks vary widely along a number of dimensions, including whether or not: (a) a hierarchical map is to be produced, (b) concepts (terms) are provided to the student for mapping, (c) a structure for the map is provided, (d) students can physically move terms around before settling on a map, (e) the student draws the map, (f) more than a single line connects two concepts, and (g) students use preselected labels for lines.

Perhaps the most salient variation in concept maps is whether or not the student is asked to produce a hierarchical map. For example, some concept mappers require hierarchical maps (e.g., Novak & Gowin, 1984) while others do not (e.g., White & Gunstone, 1992). Figure 4 presents a hierarchical concept map (Novak & Gowin, 1984). Contrast it with Figure 1, a non-hierarchical map.

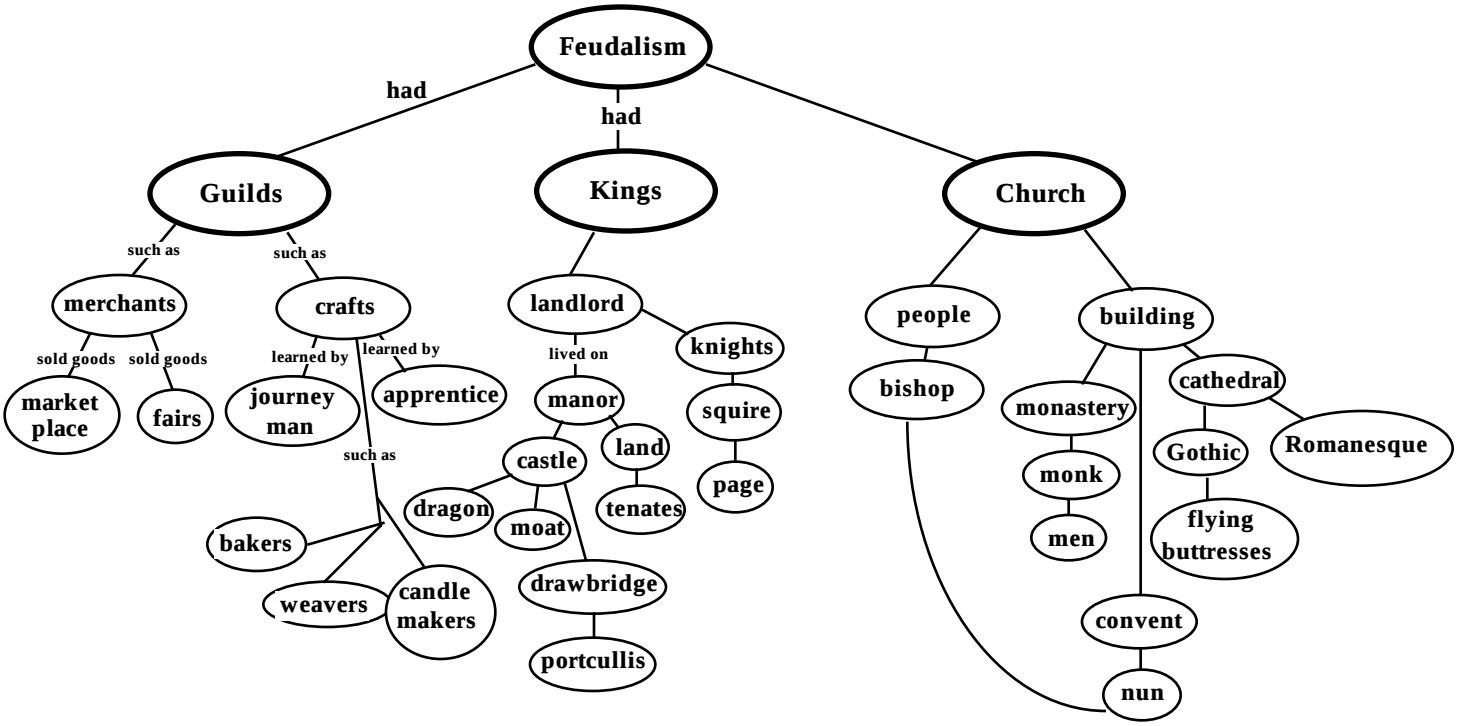


***Figure 4*. Concept map for history prepared by a previously low-achieving student in sixth grade (Novak & Gowin, 1984, p. 41).**

Preference for one or another type of map must lie at the intersection of some cognitive theory and structural notions of the subject domain to be mapped. Methodologically and conceptually, there is no need to impose hierarchical structure. If the subject matter is not hierarchical, there is no reason to impose such a structure on it (White, 1987). If a map is hierarchical, this structure can be retrieved methodologically (e.g., with graph theory; cf. Goldsmith & Davenport, 1990). This issue of theory will arise many times again in describing concept mapping techniques and is taken up in the next section.

Concept maps also vary as to whether individual concept terms are given to students for mapping or students generate the terms (Barenholz & Tamir, 1992; Beyerbach, 1988; Lay-Dopyera & Beyerbach, 1983).[5] When concept maps are developed from students' essays (Lomask, Baron, Greig, & Harrison, 1992) or interview protocol (Hewson & Hamlyn, nd), key terms are identified in the essay along with the phrases used to link them. For example, Lomask et al. (1992) asked students to "describe the possible forms of energies and types of materials involved in growing a plant and explain fully how they are related" (as reported by Baxter, Glaser, & Raghavan, 1994, p. 39). One student responded:

> **Plants need solar energy, water, and minerals. The chlorophyll is used to give it color. And I also think that the plant grows toward the suns direct. (Lomask et al., 1992, p. 22)**

The cognitive map of this "essay" is shown in Figure 5.[6] Once again, the question arises: "What cognitive theory would lead one to prefer having students generate terms over providing terms, or vice versa?"

Concept mapping techniques also vary as to whether a structure is given (cf. Anderson & Huang, 1989) and the student fills in the nodes, or the student builds the structure from scratch. Typically the "fill-in-the-nodes" variant assumes a hierarchical structure as shown in Figure 6. A set of terms that includes correct concepts and "distractors" is provided, and the student writes

---

[5] White (1987) recommended providing terms to younger students and asking older students to generate terms.

[6] Note that certain terms differ from the student's essay. This is because one term out of several more or less equivalent terms was used to stand for a particular concept. Moreover, the box apparently denotes the term given to the student in the essay prompt. Distinctions between shaded and unshaded figures and solid and broken lines were not explained by Lomask et al. (1992).

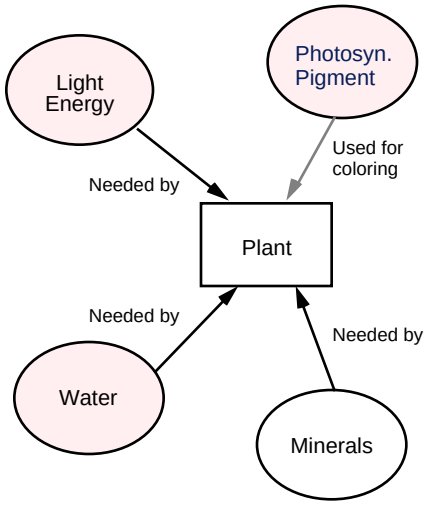
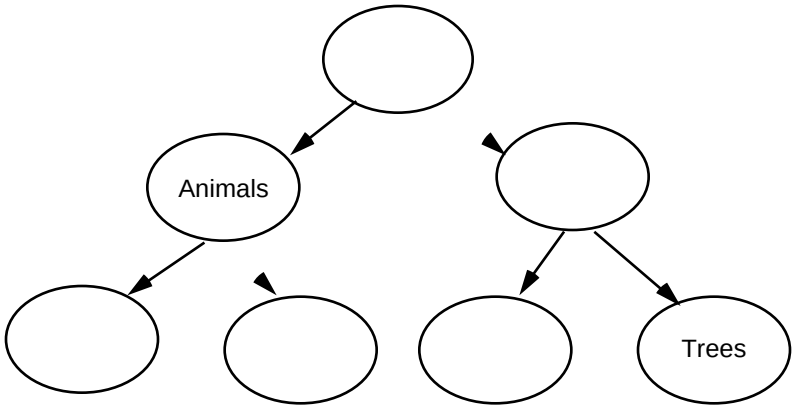


***Figure 5*. Concept map derived from student essay (Lomask et al., 1992, p. 22).**



- Living things
- Cars
- Glass
- Plastic bag
- Cat
- Picture
- Plants
- Dog
- Flowers
- Videotape

***Figure 6*. "Fill-in-the-blanks" concept map.**

those terms in the empty nodes. One wonders what cognitive theory leads to preferring the fill-in version over the student-generated concept map.

Still another variation in mapping is whether the student can physically move the concepts around until a satisfactory structure is arrived at (Figure 1; e.g., Fisher, 1990; White & Gunstone, 1992), or students draw maps on paper. To permit students to move concepts around spatially, White and Gunstone used 3" x 5" cards with concept labels on them; Fisher (1990) used a computer. Once again, preference for one or another technique for spatially arranging concepts assumes a cognitive theory that needs to be made explicit.

Concept maps also vary as to who draws the map. Most frequently students draw their own maps. However, to draw one's own map requires experience in mapping; the older the student, the less experience needed to map (Novak & Gowin, 1984; White & Gunstone, 1992). Some researchers such as Lomask et al. (1992) have trained teachers to draw concept maps from students' essays (see Figure 5). The practical consequence of this tack was that, in the context of a statewide assessment, students did not have to be taught to draw maps. Nevertheless, this translation from essay to map adds a layer onto the structural representation and raises both cognitive-theoretical and methodological issues.

Maps may also vary as to whether students are asked to draw a single line between a pair of concepts and label it, or draw as many lines as possible between concept pairs and label each (e.g., White & Gunstone, 1992; see Figure 7). Apparently there are cognitive-theoretical reasons to believe that terms can be linked in more than one way in a subject domain, although the theory driving this alternative has not been explicated.

Finally, maps vary as to whether the student provides the label for the line linking two concepts or selects a label from a set of semantic terms (e.g., "characteristic of," "prior to," "like," "part of," "type of,"; Baker, Niemi, Novak, & Herl, 1992; Holly & Dansereau, 1984a, 1984b; McClure & Bell, 1990). According to Holly and Dansereau (1984a, p. 10):

> **The student identifies important concepts or ideas in the material and represents their interrelationships and structure in the form of a network map. To assist the student in this endeavor she or he is taught a set of named links that can be used to code the relationships between ideas. The networking process emphasizes the identification and representation of (1) hierarchies (type-part), (2) chains (lines of reasoning-temporal orderings-causal sequences), and (3) clusters (characteristics-definitions-analogies).**

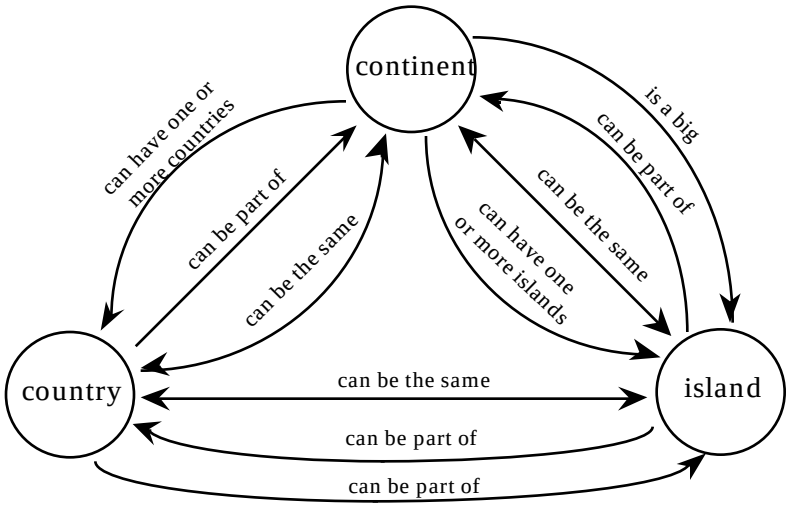


***Figure 7*. Concept map with multiple labeled links between concepts (White & Gunstone, 1992, p. 31).**

Figure 8 provides an example of a "networking" map. Again, preference for one or another mapping technique must lie on the intersection of cognitive theory and the subject domain.

There are more variations in concept mapping tasks than described here. Nevertheless, the variation documented is considerable. Indeed, there are no less than 128 (2^7) different ways to produce concept maps based on what we presented(!). Needless to say, cognitive and subject matter theories are needed to decide on which combinations are to be preferred over others.

Variations in Scoring Methods

As might be expected, scoring methods vary widely, as well. At one extreme is the recommendation not to score maps but to use them to gain a clinical impression of a student's conceptual understanding (e.g., White & Gunstone, 1992). At the other extreme is a detailed scoring system that assumes hierarchical knowledge representation. Novak and Gowin (1984, Table 2.4, pp. 36-37) have provided the most comprehensive scoring system to our knowledge:

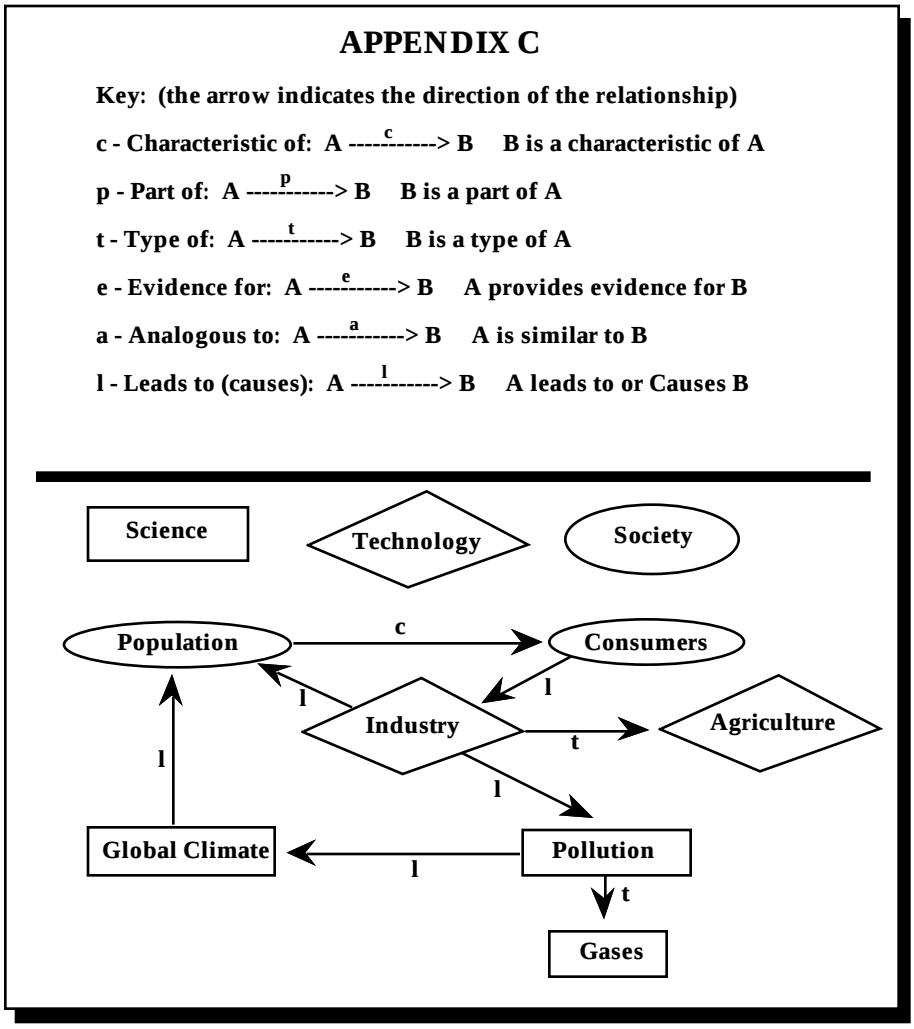


***Figure 8*. Concept map with semantic labels (McClure & Bell, 1990, p. 9).**

1. ***Propositions*** **Is the meaning relationship between two concepts indicated by the connecting line and linking word(s)? Is the relationship valid? For each meaningful, valid proposition shown, score 1 point.**

2. ***Hierarchy*** **Does the map show [sic] hierarchy? Is each subordinate concept more specific and less general than the concept drawn above it (in the context of the material being mapped)? Score 5 points for each valid level of the hierarchy.**

3. *Cross links*[7] **Does the map show meaningful connections between one segment of the concept hierarchy and another segment? Is the relationship shown significant and valid? Score 10 points for each cross link that is both valid and significant and 2 points for each cross link that is valid but does not illustrate a synthesis between sets of related concepts or propositions.**

4. *Examples* **Specific events or objects that are valid instances of those designated by the concept label can be scored 1 point each.**

A myriad of alternative scoring methods lie in between these extremes. One widely used method matches a student's map against a standard and scores the overlap. For example, Novak and Gowin (1984, p. 36) add the following fifth rule to their scoring system:

5. *Comparison* **In addition, a criterion concept map may be constructed, and scored, for the material to be mapped, and the student scores divided by the criterion map score to give a percentage for comparison. (Some students may do better than the criterion and receive more than 100% on this basis.)**

This scoring criterion is apt for many mapping techniques but not others such as the "fill-in-the-nodes" map (see Figure 6).

Some scoring systems count the number of linked concept pairs (e.g., White & Gunstone, 1992). The links can be hierarchical, multiple, or cross links. Points are given for the number of links that are the same as those on a target map (e.g., the instructor's map). Additional points are given for "insightful" links, and points are deducted for incorrect links. Alternatively, the links can be classified into semantic categories and a score formed by dividing the total

---

[7] "Cross links that show valid relationships between two distinct segments of the concept hierarchy signal possibly important integrative reconciliations, and may therefore be better indicators of meaningful learning than are hierarchical levels. . . . Since it is possible to construct some kind of cross link between almost any two concepts on a map, you must use judgment to decide if a given cross link represents substantial integrative reconciliation of two sets of concepts" (Novak & Gowan, 1984, p. 107).

number of links[8] by the number of semantic categories (Mahler, Hoz, Fischl, Tov-ly, & Lernau, 1991).

Another method focuses specifically on propositions in the concept map; a proposition is formed when two terms or concepts are linked together via a directional arrow and the link is labeled. With this method, three parts of the proposition are scored:

1. **The relation between the concepts,**
2. **The label, and**
3. **The direction of the arrow indicating either a hierarchical or causal relationship between concepts.**

For example, McClure and Bell (1990) used concept maps ("networks") to address the question "How does STS [Science, Technology and Society] instruction affect cognitive structure?" (p. 2). They focused on propositions created by students using such a set of scoring rules (see Figure 9).

Still another scoring method focuses on students' definitions of the terms provided for mapping. Mahler et al. (1991) scored each definition on a 5-point scale ranging from correct (4 points) to partially correct (3 to 1 points) to incorrect (0).

For mapping tasks in which students generate their own terms, the number of "allowable" terms—terms constrained by the subject domain—are counted. Hence, Lomask et al. (1992) counted the number of key terms in concept map representations of students' essays.

Finally, some scoring systems combine a count of terms with a count of the number of correct links between the terms to arrive at an overall score (e.g., Lomask et al., 1992; for a critique, see Baxter et al., 1994). To begin, Lomask et al. scaled both the count of terms and the count of links as follows. The "size" of the count of terms was expressed as a proportion of terms in an expert concept map mentioned by a student. This proportion was scaled from complete (100%) to substantial (67%–99%) to partial (33%–66%) to small (0%–32%) to none (no terms mentioned or irrelevant terms mentioned). Likewise, they characterized the "strength" of the links (propositions) between terms as a proportion of

---

[8] Mahler et al. do not say whether the total number of links or the total number of correct links are counted.

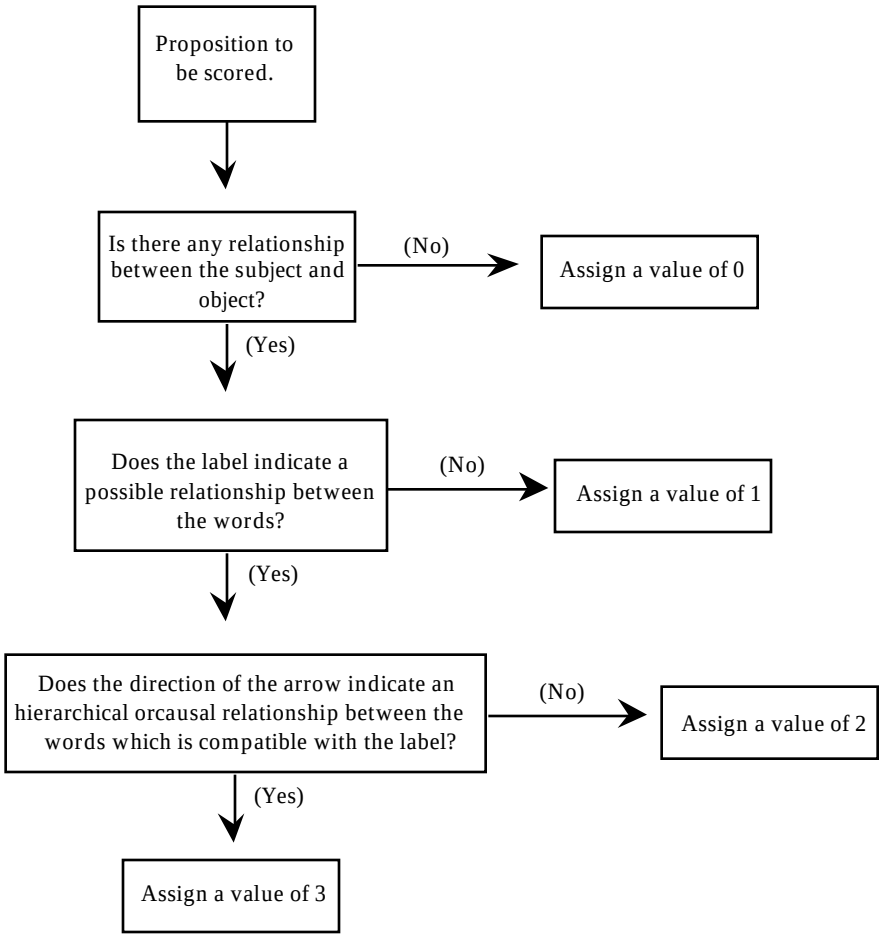


***Figure 9*. Scoring system for a concept network (McClure & Bell, 1990, p. 10, Appendix D).**

necessary, accurate connections with respect to the expert map. Strength ranged from strong (100%) to medium (50%–99%) to weak (1–49%) to none (0%). Then they provided a table that produced scores taking into account both "size" of terms and "strength" of links (Table 1).

Conclusions

An assessment can be defined as a combination of a *task*, a *response format*, and a *scoring system*. Without all three, the assessment is not completely known. Based on this definition, the variation in concept map assessments that we have described is enormous. Moreover, these assessments are unlikely to produce equivalent scores (Baxter & Shavelson, in press). Nevertheless, all

Table 1

Scores Based on Combinations of "Size" and "Strength" of Students' Concept Maps

Size	Strength			
	Strong	Medium	Weak	None
Complete	5	4	3	2
Substantial	4	3	2	1
Partial	3	2	1	1
Small	2	1	1	1
None/Irrelevant	1	1	1	1

variations have been interpreted as representing a student's "cognitive structure"—the relationships of concepts in a student's memory (Shavelson, 1972). How can it be that the variations produce somewhat different representations and scores of "goodness of cognitive structure," yet all are interpreted as a measure of the same thing?

A cognitive theory is needed to evaluate the variations that arise in tasks, response formats, and scoring systems. This theory would help reduce the number of tasks, formats, and scoring systems, perhaps to a manageable set. This set could then be reduced even further by taking into consideration practical constraints imposed by large-scale testing conditions. Finally, this small, practical set of concept map assessments could then be evaluated psychometrically with the goal of recommending one or a few potentially equivalent techniques for research and possibly large-scale assessment.

We now turn to a review of cognitive theories that have been used in conjunction with concept mapping with the hope of uncovering a theory that accounts for the variation described above, and guides us in reducing the number of possible assessments. This review will show, unfortunately, that, in many studies, far too little theory has driven concept mapping. This paucity of theory may account for the proliferation in concept mapping techniques.

Cognitive Theory and Concept Mapping

Cognitive theory underlying concept mapping in science grew out of two related traditions: Ausubel's (1968) hierarchical memory theory and Deese's (1965) associationist memory theory. Both theories eventually arrived at the same place—a concept or cognitive map from which a student's cognitive structure was inferred. The former theory posited a hierarchical memory structure while the latter posited a network of concepts that did not assume but could accommodate hierarchies. We sketch each in turn, drawing implications for concept map assessments.

Hierarchical Concept Maps

Working from Ausubel's (1968) theory, Novak and his colleagues (e.g., Novak & Gowin, 1984) coined the term "concept map." According to Novak (1990, p. 29):

> **Concept maps . . . are a representation of *meaning* . . . specific to a domain of knowledge, for a given context. We define *concept* as a perceived regularity in events or objects . . . designated by a label. . . . Two or more concepts can be linked together with words to form *propositions* and we see propositions as the *units* of psychological meaning. The *meaning* of any concept for a person would be represented by all the propositional linkages that the person could construct that include that concept.**

Concept maps, then, were intended to "tap into a learner's cognitive structure and to externalize, for both the learner and the teacher to see, what the learner already knows" (Novak & Gowin, p. 40). Novak and Gowin recognized that any one representation would be incomplete—not all concepts or propositions would be represented. Nevertheless, such maps would provide a "workable representation" (p. 40).

Reliance on Ausubel's (1968) theory, which posited a hierarchical memory (cognitive) structure, inevitably led to the view that concept maps must be hierarchical. This hierarchical structure arises because "new information often is relatable to and subsumable under more general, more inclusive concepts" (Novak & Gowin, 1984, p. 97). Moreover, the hierarchy expands according to Ausubel's principle of progressive differentiation: New concepts and new links are added to the hierarchy, either by creating new branches, or by differentiating existing ones even further. Finally, meaning increases for students as they

recognize new links between sets of concepts or propositions at the same level in the hierarchy. These links are "cross links"; they cut horizontally across the hierarchy to link sub-branches.

Ausubel's theory, then, provided guidance as to what does and does not constitute a legitimate concept map. Based on this theory, Novak and Gowin (1984) argued that all concept maps should be:

- **Hierarchical with superordinate concepts at the apex;**
- **Labeled with appropriate linking words;**
- **Cross-linked such that relations between subbranches of the hierarchy are identified.**

In addition, several other constraints emerge from both the theory and Novak and Gowin's explication of concept maps. Concept maps should be:

- **Structural representations generated by students freely and not constrained by a given structure;**
- **Labeled by students in their own words;**
- **Based on a few (say 10 or fewer) important concept terms in the subject domain; and**
- **Provided by the assessment.**

Network Concept Maps

Associationist theory (e.g., Deese, 1965) provided a beginning for characterizing cognitive structure as a set of concepts and their interrelations (e.g., Shavelson, 1972). Concepts were represented as nodes in a network. The nodes were linked by the associative overlap of two concepts. This theory lay as the basis for *indirect approaches* to eliciting representations of cognitive structure such as word associations, similarity judgments, and tree building. Such methods produce networks or concept maps with *unlabeled lines*.

This network characterization led naturally to the current view of propositional knowledge as a "semantic network" with concept nodes linked directionally by labeled lines (arrows) to produce propositions. The meaning of a concept is determined by a list of its properties which are other concepts (nodes). For example, the concept *plant* is partially defined by its property list: flower, nursery, rose, fragrance, and love (Figure 10). In short, a concept is defined by

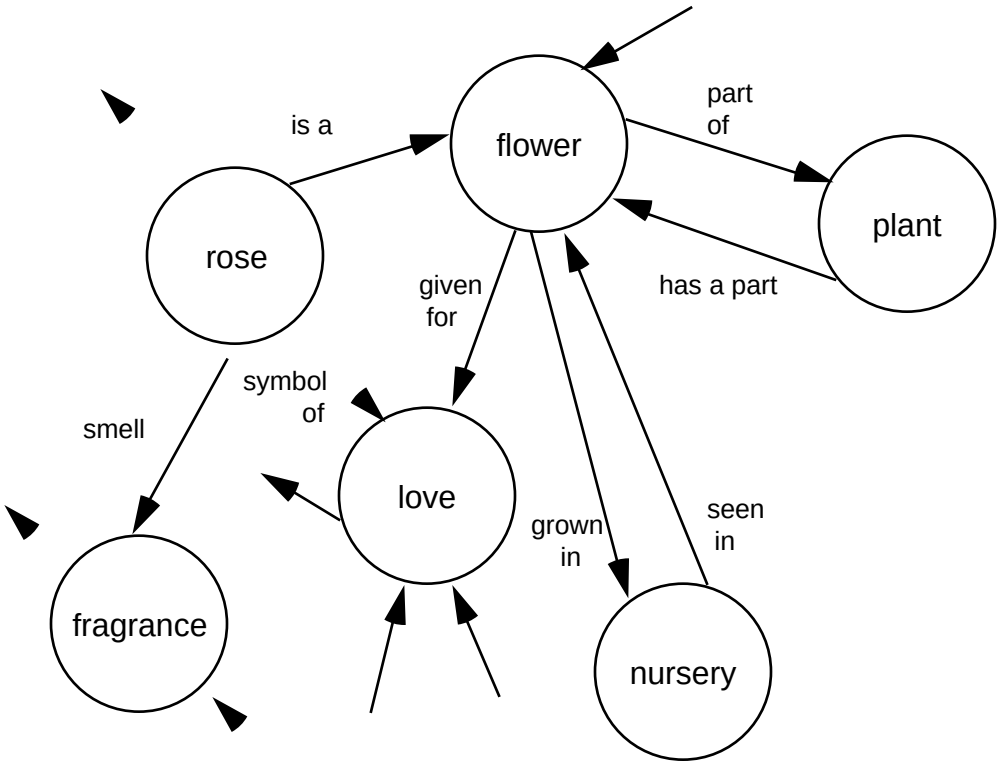


***Figure 10*. Fragment of a semantic network (after Fridja, 1972, p. 4).**

its relation to other concepts (cf. Shavelson, 1974). The lines between the nodes represent various relationships; any number of lines may connect two nodes. Some common relationships are: (a) superset—flowers may be roses; (b) subset—a rose is a flower; (c) attribute—fragrance is an attribute of a rose; and (d) part-whole—a flower is part of a plant.

Networks become increasingly elaborated as the individual learns, linking new concepts to prior existing ones. Moreover, the network may divide nodes into subsets and indicate the link ("cross link") between these subsets.

The basic unit of meaning in the network is a concept pair and its link: a proposition (cf. Gagné, Yekovich, & Yekovich, 1993). To search the network for the meaning of a concept, the concept node is activated and a search radiates out across adjacent and more distinct nodes via the links (arrows). The search is

constrained by its context, such as a subject domain. For example, the search of Figure 8 would vary depending on whether one was thinking of poetry or of botany(!).

Champagne, Klopfer, DeSena, and Squires (1981), working within the tradition of indirect methods of eliciting representations of cognitive structure, reasoned that such representations were incomplete and difficult to interpret without meaningful labels linking the concepts. Consequently, they asked students to arrange geology terms on a large piece of paper in a way that showed how they went together and to link them with labeled lines. Champagne et al. (1981) had placed themselves squarely in the knowledge-network camp: The students, then, produced concept maps akin to Novak and Gowin's (1984) with one very important exception—the concept maps did not have to be hierarchical (see White, 1987).

Network theory, then, places requirements on concept mapping very similar to those in Ausubel's theory with the important exception that maps do not have to be hierarchical. Nevertheless, several additional requirements for maps can be identified from network theory. These are combined here with those arising out of Ausubel's theory and Novak and Gowin's work on concept maps.

Concept maps should:

- **Be networks with nodes representing concept terms and lines representing directional relations between concept pairs.**
- **Be hierarchical with superordinate concepts at the apex when the subject domain is clearly hierarchical.**
- **Contain labeled links with appropriate linking words.**
- **Contain cross links such that relations between subbranches of the network are identified.**
- **Be structural representations generated by students freely and not constrained by a given structure.**
- **Be labeled by students in their own words.**
- **Be based on a few (say 10 or fewer) important concepts in the subject domain.**
- **Either permit students to provide their own terms in a subject domain, or provide concept terms in the assessment.**

- **Contain sufficiently clear, unambiguous instructions to permit students to search memory in the desired manner and to establish appropriate criteria against which to test alternative responses.**

Technical Quality of Concept Maps as Assessments

Concept maps are far more frequently used as instructional tools than as formal assessments. When maps are used as assessments, they are typically found in pre and posttreatment evaluations. In only one study were concept maps used in a large-scale, statewide assessment (Lomask et al., 1992). As a consequence, few studies report reliability and validity information, and none that we have reviewed has systematically studied the technical qualities of concept maps.

Reliability

Reliability refers to the consistency (or "generalizability"; cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972) of scores assigned to students' concept maps. Only Lomask et al. (1992) provided information on the reliability with which concept maps can be scored. (Recall that, in this study, concept maps were generated from students' essays.) Four teachers scored 39 students' concept maps in the domain of growing plants, and 42 students' concept maps in the domain of blood transfusion. Each teacher first produced a concept map from a student's essay. Next, the teacher compared the student's map with an expert's map of the domain. Three features of the concept maps were scored:

- **Number of concepts in the expert map used by the student;**
- **Number of correct connections between concepts;**
- **Expected number of connections among all of the concepts mentioned in an essay.**

Lomask et al.'s (1992) findings are summarized in Table 2. Two findings stand out. The first is that raters introduced negligible error in estimating the level of student performance (rater effect is 0 or close to zero, Table 2). This finding, however, must be interpreted cautiously because Lomask et al. eliminated data from one teacher when his/her scores deviated substantially from those of the other teachers. Nevertheless, this finding is consistent with Hewson and Hamlyn's (nd) finding that the agreement among 5 independent

Table 2

Person x Rater Generalizability Studies (adapted from Lomask et al. 1992, Table 4): Percent of Total Variation

Source of variation	Growing plants			Blood transfusion		
	$N_{concepts}$	$N_{connect}$	$E_{connect}$	$N_{concepts}$	$N_{connect}$	$E_{connect}$
Student (S)	84	77	81[a]	89	84	62
Rater (R)	0	0	1	1	1	9
S x R, e	16	23	18	10	15	29
Reliability[b]	.84	.77	.81	.89	.84	.62

[a] One rater's scores were eliminated from these analyses on the basis of their frequent inconsistency with the scores of the other three raters (Table 4).

[b] Generalizability coefficient for absolute decisions (φ); one rater (see Shavelson & Webb, 1991).

judges evaluating two interview transcripts was 97.5% for metaphorical heat conceptions and 82.1% for physical heat conceptions. Moreover, the literature on performance assessment has established that raters can be trained to reliably score complex performance (e.g., Shavelson et al., 1993).

The second finding is that concept maps drawn from students' essays can be scored reliably (mean = .81 and .78 for plants and blood transfusion, respectively). This was no simple feat. Recall that the concept maps had to be drawn from students' essays; the teachers might very well have drawn substantially different maps. Apparently they did not. We wonder if the small unreliability in the data was due to differences in the maps teachers drew, or in their scoring of the maps once drawn.

Even though seldom studied, reliability of concept map scores is an issue that must be addressed before they are reported to the public and policy makers. A number of reliability studies can be conceived; each must be adapted to the particular task and scoring procedures used. The following questions should be raised:

- **Can raters reliably score concept maps?**
- **Do students produce the same or highly similar maps from one occasion to the next when no instruction/learning has intervened?**
- **Are map scores sensitive to the sampling of concept terms?**

Validity

Validity refers to the extent to which inferences to students' "cognitive structures" on the basis of their concept map scores can be supported logically and empirically. On logical grounds, a few concept map studies report that "experts" and/or teachers judged the terms and maps as consistent with the subject domain (e.g., Anderson & Huang, 1989; Barenholz & Tamir, 1992). On empirical grounds, several studies show a consistent correlation between concept map scores and other measures of student achievement or accomplishment. For example, Anderson and Huang (1989) reported substantial correlations (above 0.50) between concept map scores (e.g., total number of propositions in map) and measures of aptitude and science achievement (see Table 3). In general, we found correlations at or above .50 between concept map scores and measures of student achievement (e.g., McClure & Bell, 1990), and in some studies we found significant gains in concept map scores from pretest to posttest (e.g., Beyerbach, 1988). None of the studies, however, focused on a systematic evaluation of the validity of the cognitive-structure interpretations of concept map scores.

Construct validation studies of concept map techniques need to be carried out before scores from such assessments are reported to the public and policy makers. Among the questions that need to be addressed are:

Table 3

Correlations Between Concept Map Scores and Measures of Achievement and Ability (see Anderson & Huang, 1989)

Achievement/Ability	Correlation
Essay test on unit	.69
Stanford Science Achievement Test	.66
School science grades	.49
Otis Lennon School Ability Test	.74

- **Are the concept terms used in the assessment representative of the subject domain?**
- **Is the concept map "interpretable" within the subject domain?**
- **Does the concept-representation of cognitive structure correspond to cognitive-structure representations generated from other techniques?**
- **How exchangeable are concept maps developed from different concept mapping techniques?**
- **What types of scoring systems capture the underlying construct being measured by concept maps, namely, cognitive structure?**
- **Do concept map scores correlate with other measures of achievement in the subject domain?**
- **Do concept map scores "unfairly" distinguish among diverse student groups varying in socioeconomic status, race/ethnicity, gender, or proficiency in English?**
- **Do concept map assessments lead to teaching science concepts and procedures in a manner consistent with science education reform?**

Considerations in Using Concept Maps in Large-Scale Assessments

Assuming concept maps tap some important aspect of students' knowledge structures, their use in a large-scale, high-stakes assessment context needs to be considered. Two important, related issues can be identified immediately. The first issue has to do with students' facility in using the technique (Carey & Shavelson, 1989), and the second with the consequences of teachers teaching to the test (cf. Shavelson, Carey, & Webb, 1989).

Students' Facility With Concept Mapping

The literature on concept mapping (e.g., Baker et al., 1992; Novak & Gowin, 1984; White & Gunstone, 1992) is quite consistent: Students must be trained in using concept mapping. Attempts to provide brief training before the assessment is given are usually inadequate. Indeed, one explanation for students' poor performance on concept maps is often that more and/or better training in the task was needed. Perhaps research on how to train students, in a brief period of time, to create concept maps is needed for large-scale use of the technique. We think not (cf. White & Gunstone, 1992). Training on a generic mapping skill, outside the domain to be assessed, probably will not transfer to

the particular subject domain, especially for students from culturally and economically different backgrounds.

Concept map training, then, must be well contextualized in the domain. But how can this be done without letting the proverbial cat out of the bag and teaching students what they are to be assessed on? The alternative is to embed concept mapping within the curriculum. Concept maps would be both instructional tools and assessments—creating a symmetry between teaching and testing (see Shavelson & Baxter, 1992). Note that this tack assumes that concept mapping is an effective learning and assessment tool (the jury is still out), important enough to be used regularly in classrooms. With this scenario, "training" students to use concept maps specifically for large-scale assessment would be unnecessary.

Indeed, the use of concept maps as part of high-stakes, large-scale assessment would be tantamount to mandating their classroom use. Hence, to use concept mapping in large-scale assessment means that the education community must be prepared to say that maps are important teaching and assessment tools that should be used regularly in schools. Moreover, the benefits must outweigh the drawbacks of the time it takes to teach teachers[9] to use the technique, otherwise teachers may very well view this as an intrusion on their already over-mandated teaching time. Not only might they view this as an intrusion, they may very well warp the method in unintended ways. This concern leads to a second issue in using concept maps in large-scale assessment.

Teaching to the Concept Map Test

If used appropriately (see Novak & Gowin, 1984; White & Gunstone, 1992), concept maps would increase teachers' repertoire of instructional techniques and their repertoire of assessment techniques. The focus of teaching would be on students arriving at a broad conceptual understanding of the domain of study. One fourth-grade teacher wrote to us about the potential of concept maps as instructional tools:

> **My belief about this** [the value of concept maps] **is based, first, on the idea that understanding the "big picture" of any topic (math and science in this case) is very**

---

[9] The staff-development implications are substantial. Large-scale staff development would be needed not only for the purpose of learning how to teach the new mapping skills, but also to become aware of the philosophical and pedagogical reasons for using this assessment.

> **important and necessary not only to explore, and create new things, but to understand the existing world around us. Using concept maps as an assessment tool will urge educators to teach students more than simple facts and concepts, but how different concepts relate to each other. An evaluational tool such as concept mapping urges the individual to think on a deeper cognitive level than a "fill in the blank" test would require. There is value in both assessment tools—neither should be ignored.**

In spite of the potential salutary impact of concept maps on classroom teaching and assessment, there is also reason for alarm. Suppose that, instead of the practice advocated by this teacher, another teacher decides to present an "expert" map to students and require them to memorize the map. Indeed, the teacher might even test and grade students for the accuracy of their recall of the expert map, in anticipation of its use, for example, in a statewide assessment. This use of concept mapping runs contrary to intentions of assessors and curricular reform (see also White, 1987). Nevertheless the possibility of inappropriately teaching to the test should be considered real and monitored.

References

Anderson, T.H., & Huang, S-C.C. (1989). ***On using concept maps to assess the comprehension effects of reading expository text.*** **(ERIC Document Reproduction Service No. ED 310 368)**

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

Baker, E.L., Niemi, D., Novak, J., & Herl, H. (1992). Hypertext as a strategy for teaching and assessing knowledge representation. In S. Dijkstra, H.P.M. Krammer, & J.J.G. van Merriënboer (Eds.), ***Instructional models in computer-based learning environments*** **(pp. 365-384). New York: Springer-Verlag.**

Barenholz, H., & Tamir, P. (1992). A comprehensive use of concept mapping in design instruction and assessment. ***Research in Science and Technological Education, 10*****(1), 37-52.**

Baxter, G.P., Glaser, R., & Raghavan, K. (1994). ***Analysis of cognitive demand in selected alternative science assessments*** **(CSE Tech. Rep. No. 382). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).**

Baxter, G.P., & Shavelson, R.J. (in press). Science performance assessments: Benchmarks and surrogates. ***International Journal of Educational Research.***

Beyerbach, B. (1988). Developing a technical vocabulary on teacher planning: Preservice teachers' concept maps. ***Teaching and Teacher Education 4*****(4), 339-47.**

Brown, A.L., & Ferrara, R.A. (1985). Diagnosing zones of proximal development: An alternative to standardized testing? In J. Wertsch (Ed.), ***Culture, communication and cognition: Vygotskian perspectives*** **(pp. 273-305). New York: Cambridge University Press.**

Carey, N., & Shavelson, R. (1989). Outcomes, achievement, participation, and attitudes. In R.J. Shavelson, L.M. McDonnell, & J. Oakes (Eds.), ***Indicators for monitoring mathematics and science education: A sourcebook*** **(R-3742-NSF/RC; pp. 147-191). Santa Monica, CA: The RAND Corporation.**

Champagne, A.B., Klopfer, L.E., DeSena, A.T., & Squires, D.A. (1981). Structural representations of students' knowledge before and after science instruction. ***Journal of Research in Science Technology, 8,*** **97-111.**

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). ***The dependability of behavioral measurements: Theory of generalizability of scores and profiles.*** **New York: John Wiley.**

Deese, J. (1965). ***The structure of associations in language and thought.*** **Baltimore: The Johns Hopkins Press.**

Ericsson, A.K., & Simon, H.A. (1984). ***Protocol analysis. Verbal reports as data.*** **Cambridge, MA: MIT Press.**

Fisher, K.M. (1990). Semantic networking: The new kid on the block. ***Journal of Research on Science Teaching, 27,*** **1001-1018.**

Fridja, N.H. (1972). Simulation of human long-term memory. ***Psychological Bulletin, 77,*** **1-31.**

Gagné, E.D., Yekovich, C.W., & Yekovich, F.R. (1993). ***The cognitive psychology of school learning.*** **New York: Harper Collins.**

Goldsmith, T.E., & Davenport, D.M. (1990). Assessing structural similarity of graphs. In R.W. Schvaneveldt (Ed.), ***Pathfinder associative networks: Studies in knowledge organization*** **(pp. 75-87). Norwood, NJ: Ablex.**

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

Hewson, M.G., & Hamlyn, D. (nd). ***The influence of intellectual environment on conceptions of heat.*** **(ERIC Document Reproduction Service No. ED 231 655)**

Holly, C.D., & Dansereau, D.F. (1984a). The development of spatial learning strategies. In C.D. Holly & D.F. Dansereau (Eds.), ***Spatial learning strategies: Techniques, applications and related issues*** **(pp. 3-19). New York: Academic Press.**

Holly, C.D., & Dansereau, D.F. (1984b). Networking: The technique and the empirical evidence. In C.D. Holly & D.F. Dansereau (Eds.), ***Spatial learning strategies: Techniques, applications and related issues*** **(pp. 81-108). New York: Academic Press.**

Lay-Dopyera, M., & Beyerbach, B. (1983). ***Concept mapping for individual assessment.*** **(ERIC Document Reproduction Service No. ED 229 399)**

Lomask, M., Baron, J.B., Greig, J., & Harrison, C. (1992, March). ***ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science.*** **Paper presented at the annual meeting of the National Association of Research in Teaching, Cambridge, MA.**

Magone, M., Cai, J., Silver, E.A., & Wang, N. (in press). Validating the cognitive complexity and content quality of a mathematics performance assessment. ***International Journal of Educational Research.***

Mahler, S., Hoz, R., Fischl, D., Tov-ly, E., & Lernau, O.Z., (1991). Didactic use of concept mapping in higher education: Applications in medical education. ***Instructional Science, 20,*** **25-47.**

McClure, J.R., & Bell, P.E. (1990). ***Effects of an environmental education-related STS approach instruction on cognitive structures of preservice science teachers.*** **(ERIC Document Reproduction Service No. ED 341 582)**

Novak, J.D. (1990). Concept maps and Vee diagrams: Two metacognitive tools to facilitate meaningful learning. ***Instructional Science, 19,*** **29-52.**

Novak, J.D., & Gowin, D.R. (1984). ***Learning how to learn.*** **New York: Cambridge University Press.**

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), ***Future assessments: Changing views of aptitude, achievement, and instruction*** **(pp. 37-75). Boston: Kluwer.**

Schreiber, D.A., & Abegg, G.L. (1991). ***Scoring student-generated concept maps in introductory college chemistry.*** **Paper presented at the annual meeting of the National Association for Research in Science Teaching, Lake Geneva, WI. (ERIC Document Reproduction Service No. ED 347 055)**

Shavelson R.J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. ***Journal of Educational Psychology, 63,*** **225-234.**

Shavelson, R.J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. ***Journal of Research in Science Teaching, 11,*** **231-249.**

Shavelson, R.J., & Baxter, G.P. (1992). Linking assessment with instruction. In F.K. Oser, A. Dick, & J-L. Patry (Eds.), ***Effective and responsible teaching: The new synthesis.*** **San Francisco: Jossey-Bass.**

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. ***Journal of Educational Measurement, 30,*** **215-232.**

Shavelson, R.J., Baxter, G.P., & Pine, J. (1991). Performance assessment in science. ***Applied Measurement in Education, 4,*** **347-362.**

Shavelson, R.J., Carey, N.B., & Webb, N.M. (1989). Indicators of science achievement: Options for a powerful policy instrument. ***Phi Delta Kappan, 71,*** **692-697.**

Shavelson, R.J. & Stanton, G.C. (1975). Construct validation: Methodology and application to three measures of cognitive structure. ***Journal of Educational Measurement, 12,*** **67-85.**

Shavelson, R.J., & Webb, N.M. (1991). ***Generalizability theory: A primer.*** **Newbury Park, CA: Sage.**

White, R.T. (1987). Learning how to learn. *Journal of Curriculum Studies, 19,* 275-276.

White, R., & Gunstone, R. (1992). *Probing understanding*. New York: Falmer Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70,* 703-713.