JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 33, NO. 6, PP. 569-600 (1996)

Problems and Issues in the Use of Concept Maps in Science Assessment

Maria Araceli Ruiz-Primo and Richard J. Shavelson

School of Education, Stanford University, Stanford, California 94305-3096

Abstract

The search for new, authentic science assessments of what students know and can do is well under way. This has unearthed measures of students' hands-on performance in carrying out science investigations, and has been expanded to discover more or less direct measures of students' knowledge structures. One potential finding is concept mapping, the focus of this review. A concept map is a graph consisting of nodes representing concepts and labeled lines denoting the relation between a pair of nodes. A student's concept map is interpreted as representing important aspects of the organization of concepts in his or her memory (cognitive structure). In this article we characterize a concept map used as an assessment tool as: (a) a task that elicits evidence bearing on a student's knowledge structure in a domain, (b) a format for the student's response, and (c) a scoring system by which the student's concept map can be evaluated accurately and consistently. Based on this definition, multiple concept-mapping techniques were found from the myriad of task, response format, and scoring system variations identified in the literature. Moreover, little attention has been paid to the reliability and validity of these variations. The review led us to arrive at the following conclusions: (a) an integrative working cognitive theory is needed to begin to limit this variation in concept-mapping techniques for assessment purposes; (b) before concept maps are used for assessment and before map scores are reported to teachers, students, the public, and policy makers, research needs to provide reliability and validity information on the effect of different mapping techniques; and (c) research on students' facility in using concept maps, on training techniques, and on the effect on teaching is needed if concept map assessments are to be used in classrooms and in large-scale accountability systems.

Concept Maps as Potential Alternative Assessments in Science

Alternative assessments are intended to provide evidence about what students know and can do in a subject matter. Performance assessments in science, for example, yield evidence about what students can do when presented with a problem and provided a laboratory with which to carry out an investigation to solve it (e.g., Shavelson, Baxter, & Pine, 1991). Performance assessment scores reflect both the quality of the procedures used to solve the problem and the accuracy of the solution. Interpretations of performance assessment scores usually go beyond the immediate performance and 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, 1990; Wiggins, 1989). While research

Published by John Wiley & Sons, Inc. CCC 0022-4308/96/060569-32

has shown that multiple investigations are needed to draw inferences to a broader domain of investigations (e.g., Shavelson et al., 1993), little research is being conducted to see whether such inferential leaps from performance to cognition can be supported empirically.[1]

The insistence on interpretations that go beyond statements about the level of performance (no small accomplishment in itself) to provide insight into what students know, and how that knowledge is represented and used, has led to a search for techniques to measure more or less directly students' knowledge structures. One set of techniques is grounded on the assumption that understanding in a subject domain such as science is conceived as a rich set of relations among important concepts in that domain. Some assessment techniques indirectly probe students' perceptions of concept interrelations in a science domain 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 such as 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 article, in broad terms, was to examine the validity of claims that concept maps measure an important aspect of students' knowledge structures in a subject domain such as science. To this end, we: (a) provide a working definition of concept mapping and a brief theoretical background; (b) characterize concept maps as a potential assessment in science; and (c) review empirical evidence on the reliability and validity of various concept-mapping techniques, identifying areas for further research.

Concept Maps and Cognitive Theory

Most cognitive theories share the assumption that concept interrelatedness is an essential property of knowledge. Indeed, one aspect used in defining competence in a domain is that knowledge is well structured (e.g., Glaser & Bassok, 1989). As expertise in a domain is attained through learning, training, and/or experience, the elements of knowledge become increasingly interconnected. Likewise, as students acquire expertise in a subject domain their knowledge increasingly resembles the tightly integrated structures that characterize a subject-matter expert's representation of the knowledge (e.g., Glaser & Bassok, 1989; Royer, Cisero, & Carlo, 1993).

Assuming that knowledge within a content domain is organized around central concepts, to be knowledgeable in the domain thus includes having a highly integrated structure among these concepts. This organizational property of knowledge can be captured with structural representations (e.g., Goldsmith et al., 1991; Jonassen, Beissner, & Yacci, 1993; White & Gunstone, 1992).

A concept map is a structural representation 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 called a *proposition*. A proposition is the basic unit of meaning in a concept map and the smallest unit that can be used to judge the validity of the relation (line) drawn between two concepts (e.g., Dochy, 1994). Concept maps thus purport to represent some important aspect of a student's declarative knowledge in a content domain (e.g., chemistry).

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

Hierarchic Concept Maps

Based on Ausubel's theory, Novak and colleagues (e.g., Novak & Gowin, 1984) coined the term *concept map*. Concept maps 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, 1984, p. 40). Reliance on Ausubel's theory, which posited a hierarchical memory (cognitive) structure, and on his principles of progressive differentiation and integrative reconciliation inevitably led to a specific view of concept maps.

Ausubel's theory thus provided guidance as to what constitutes a legitimate concept map. Novak and Gowin (1984) argued that concept maps should be: (a) hierarchical with superordinate concepts at the apex; (b) labeled with appropriate linking words; and (c) crosslinked such that relations between sub-branches of the hierarchy are identified. The hierarchical structure arises because "new information often is related to and subsumable under more general, more inclusive concepts" (p. 97). Moreover, the hierarchy expands according to the 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 crosslinks—links between one segment of the concept hierarchy and another segment—represent the integrative connection among different subdomains of the structure. Figure 1 presents a hierarchical concept map (Novak & Gowin, 1984).

Novak and Gowin (1984) 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).

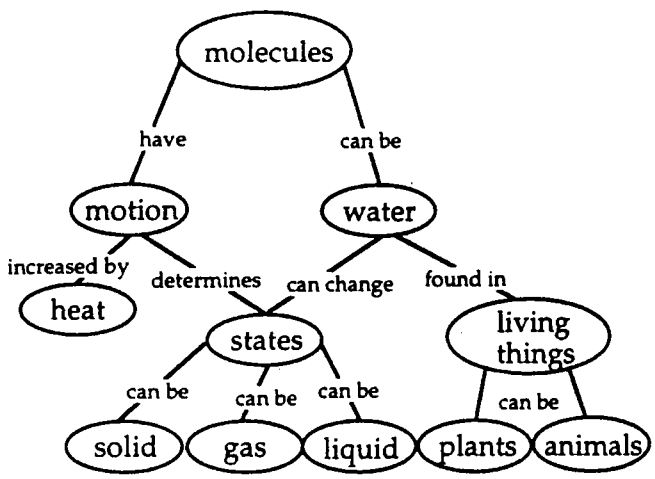

Figure 1. Hierarchical concept map (Novak & Gowin, 1984, p. 18).

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 with unlabeled lines. This theory became the basis for indirect approaches to eliciting representations of cognitive structure such as word associations, similarity judgments, card sorting, and tree building. Such methods produce networks or concept map 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). In short, a concept is defined by its relation to other concepts (cf. Shavelson, 1974). For example, the concept *plant* is partially defined by its property list: flower, nursery, rose, fragrance, and love (Figure 2). 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 links. The search is constrained by its context, such as a subject domain. For example, in Figure 2 the search would vary depending on whether one was thinking of poetry or botany.

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

Associationist network theory places requirements on concept mapping similar to those of Ausubel's theory with the important exception that maps do not have to be hierarchical. Based on this theoretical approach, then, (a) concept maps are networks with concept nodes linked directionally by labeled lines (arrows) to produce propositions, (b) the lines between the nodes

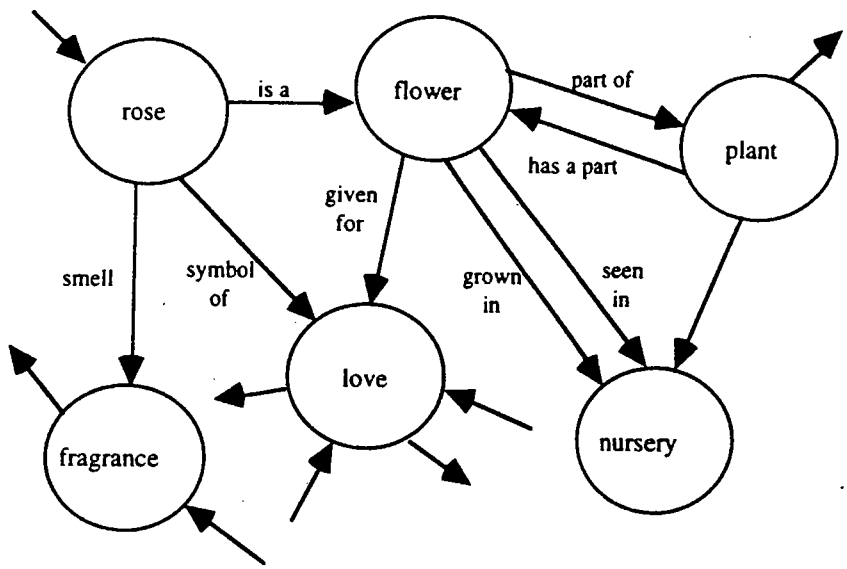

Figure 2. Fragment of a semantic network (after Fridja, 1972, p. 4).

represent various relations, (c) any number of lines may connect two nodes, and (d) the network may divide nodes into subsets and indicate the link (crosslink) between these subsets.

The implications of the theoretical conceptualizations in using concept maps as an assessment tool are straightforward. First, concept mapping is a means for eliciting students' knowledge structures in a content domain (e.g., Jonassen et al., 1993; White & Gunstone, 1992). Second, concept maps are more directly related to the knowledge of facts and concepts and how the concepts in a domain are related than how they are used or applied to solve a problem. Therefore, maps are limited in providing information about what students are able to do with the knowledge in a certain domain. Finally, there is a cognitive theoretic basis for concept maps, but the variation observed in the use of concepts maps seems to be unrelated to this body of theory. This issue of theory will be obvious when we next describe the variety of concept-mapping techniques used as an assessment tool.

Concept Maps as an Assessment Tool

Although the potential use of a concept map to assess students' knowledge structures has been recognized (e.g., Jonassen et al., 1993; White & Gunstone, 1992), maps are far more frequently used as instructional tools (e.g., Briscoe & LaMaster, 1991; Holley & Danserau, 1984; Pankratius, 1990; Schmid & Telaro, 1990; Stice & Alvarez, 1987; Willerman & Mac Harg, 1991) than as assessment tools (however, see, for example, Baxter, Glaser, & Raghavan, 1993; Beyerbach, 1988; Hoz, Tomer, & Tamir, 1990; Lomask, Baron, Greig, & Harrison, 1992).

As an assessment tool, concept maps can be thought of as a procedure to measure the structure of a student's declarative knowledge. We use the term *assessment* instead of *test* to reflect our belief that reaching a judgment about an individual's knowledge and skills requires the integration of several pieces of information; we consider concepts maps to be one of those pieces (see Cronbach, 1990). If, however, the purpose is to use concept maps alone as a procedure for describing a student's knowledge structure on a numeric scale, concept map tests would be a more appropriate term.

Here we propose a framework for conceptualizing concept maps as a potential assessment tool in science. We conceive of an assessment as a combination of a *task*, a *response format*, and a *scoring system*. Based on this framework, a concept map used as an assessment tool can be characterized as: (a) a task that invites students to provide evidence bearing on their knowledge structure in a domain, (b) a format for the students' response, and (c) a scoring system by which students' concept maps can be evaluated accurately and consistently. Without these three components, a concept map cannot be considered an assessment.

Table 1 identifies these three components in previous concept map assessment research. This categorization makes evident the wide variety of different assessment techniques that have been used under the name "concept map."

Concept Map Assessment Task

In the context of science, a concept map assessment task should elicit evidence of a student's conceptual knowledge in a content domain (e.g., biology, chemistry, physics). Here are some examples of tasks used in previous research:

> 1. Read every card out loud and make two piles: the words that you recognize and those that you do not recognize. . . . Define or explain each recognized concept. . . . Pro-

Table 1
Examples of Different Types of Tasks, Response Formats, and Scoring Systems Used in Research on Concept Maps

Authors	Task	Response	Scoring system
• Acton et al. (1994)	Rate the relatedness of pairs of 24 concepts on computer programming on a 7-point scale.	Computer response. Students entered their rating on a concept-relatedness rating program. A computer program later derived the structural representation.	Comparison with a criterion map. A theoretical measure called "closeness" (C)—the degree to which a concept has the same neighbors in two different networks—was used to compare the student's and the criterion map.
• Anderson & Huang (1989)	Fill in a map on types of muscles and their functions using the 15 concepts and 6 linkage terms provided.	Paper and pencil response. Students filled in a prestructured skeleton map.	Combination of scoring a student's map components and comparison with a criterion map. Students' map propositions were classified into 20 accuracy categories according to a criterion map and a score assigned.
• Baker et al. (1991)	Organize cards (nodes) containing an idea or piece of information in a multidimensional form using 11 link terms provided. Two content domains were evaluated: history and chemistry.	Computer response. Students wrote electronic node cards in a program and related them by using a pull-down menu with labeled links. Maps were constructed later by printing the stack of cards and links used by the student.	Score of map components: total number of nodes, number of links per node, total comment words for each completed stack of cards, and depth—number of successive links. Incorrect links were not scored. Two raters provided a content quality rating (1–5) based on the maps and the cards.
• Barenholz & Tamir (1992)	Select 20 to 30 concepts considered key concepts for a course in microbiology and use them to construct a map.	Paper and pencil response. Students drew the concept map in their notebooks.	Score of map components: number of concepts and propositions, the hierarchy and the branching, and quality of the map based on overall impression.
• Beyerbach (1988)	Construct a concept map on the topic teacher planning.	Paper and pencil response. Students drew the concept maps on a piece of paper.	Combination of scoring the students' map components and comparison with a criterion map. Score considered: number of concepts, degree of hierarchic organization, group consensus on concept used, similarity to instructor's map score,

			and item stream similarity to instructor's.
• Champagne et al. (1978)	Sort cards with concepts into two piles, known and unknown concepts on minerals and rocks. Arrange the known concepts into a structure. Sort the unknown words and make a final attempt to fit them into the structure. Justify arrangements (the ConSAT procedure).	Oral response. The interviewer drew and labeled the lines between concepts based on the students' responses.	Combination of scoring the students' map components and comparison with a criterion map. Score considered six dimensions of the complexity of the structure (e.g., size of the node that is structured).
• Fisher (1990)	Task 1. Enter concepts and relation names in the computer with as many links as desired. Task 2. Fill in the blank when a central concept is masked and the other nodes are provided.	Computer response in both tasks. Students construct their maps on a blank screen for task 1, and filled in the node(s) in a skeleton map for task 2.	The author only proposed the SemNet computer program as an assessment tool, but did not present any scoring system to evaluate the maps.
• Heine-Fry & Novak (1990)	Interview testing. Write down on a blank paper about four topics in biology (e.g., circulatory systems). Describe all they know about the topics verbally. Describe reasons for their answers.	Oral response. The researcher made a list of the concepts mentioned in the interview and constructed the map by using a template.	Score based on map components: number of linkages, number of hierarchy levels (multiplied by a factor of 5), and number of crosslinks (multiplied by a factor of 10). A final score was given by summing all the component scores.
• Hewson & Hamlyn (date unknown)	Semistructured interview about instances of life situations related to the physical phenomenon of heat.	Oral response. Subjects' conceptions were represented in a conceptual map template based on a conceptual profile inventory (CPI) compiled from all of the subjects.	Comparison with a criterion map. Subjects' maps were compared with the CPI template. The authors did not describe the procedure used to make the comparison.
• Hoz et al. (1990)	ConSAT individual interview. Lists used had 10 to 12 general and/or specific concepts. Last part of the interview asks subjects to indicate the strength of the links among concepts. Concept maps were used to measure science content and pedagogic content knowledge.	Oral response. Interviewer drew and labeled the lines between concepts based on the subjects' responses. Authors did not explain whether subjects or interviewer filled out the link matrix.	Combination of scoring the students' map components and comparison with a criterion map. Score considered: validity of the links, validity of the map as a whole, and the group of concepts. The validity of the links was based on experts' links.

(*continued*)

Table 1 (*Continued*)

Authors	Task	Response	Scoring system
• Lay-Dopyera & Beyerbach (1983)	Complete a map based on a node provided (e.g., classroom management, teaching).	Paper and pencil response. Students drew a map on a piece of paper.	Score based on map components: number of nodes, number of subordinate levels, and number of dis-
• Lomask et al. (1992)	Write an essay on two central topics on biology (i.e., growing plant and blood transfusion).	Paper and pencil response. Trained teachers construct a map from students' written essays. No effort was made to elicit any hierarchy.	Comparison with a criterion map. Two structural dimensions were identified for the comparison: the size and the strength of structure. The final scored was based on the combination of both dimensions.
• Mahler et al. (1991)	ConSAT interview with a list of 12 concepts. Students in medical school were asked to define verbally each familiar concept. Definitions were written by the interviewer.	Paper and pencil response. Students drew their maps on pieces of paper. Definitions of concepts were written by the researchers.	Combination of scoring the students' map components and comparison with a criterion map. Comparison considered: size of the semantic categories, conruence of the links, total number of links, correctness of definitions, and overlap of the groups.
• Mc Clure & Bell (1990)	Construct a concept map using 36 expressions on global climate. Linkages terms were provided.	Paper and pencil response. Students drew the concept map on a blank page.	Score based on map components: frequency and characteristics of the propositions.
• Markham et al. (1994)	Construct a hierarchical concept map from 10 given concepts on mammals.	Paper and pencil response. Students drew the concept map on a blank page.	Score based on map components: number of concepts, relations, branching, hierarchies, crosslinks, and examples. Number of concepts and relations were taken as indications of the extent of the students' knowledge.
• Nakhlch & Krajcik (1991)	Semistructured interview about acids and bases.	Oral response. The interviewer drew three concepts maps—one for acids, one for bases, and one for pH—based on statements that revealed the student's propositional knowledge.	Score based on map components: Propositions and examples, crosslinks, hierarchy. Experts' maps were used to identify critical nodes and relations.

• Novak et al. (1983)	Construct a hierarchical concept map from terms identified as key concepts in a short text.	Paper and pencil response. Students drew the concept map on a blank page.	Combination of scoring the students' map components and comparison with a criterion map. Score considered: 1 point for each correct linkage or relationship, 5 points for each level of hierarchy, and 5–10 points for crosslink showing a correct relation between two concepts in different sections of the hierarchy. Staff members constructed the criterion map and specified its criterion score.
• Roth & Roychoudhury (1993)	Construct a hierarchical concept map using 14 physics concepts printed in cards. Students were allowed to add other concepts and rearranged the concepts until their placement. Students constructed the map collectively and individually.	Paper and pencil response. Students drew the concept map on a blank 14 × 17-inch paper.	Score based on map components: number of links, levels of hierarchy, crosslinks, and examples. Scientifically correct crosslinks received more weight than a correct level, which in turn received more weight than a correct link between concepts.
• Schreiber & Abegg (1991)	Construct a map from 35 given concepts related to chemical reaction equations and chemical change.	Paper and pencil response. Students drew the concept map on a blank page.	Combination of scoring the students' map components and comparison with a criterion map. Propositional validity, hierarchical structure of students' maps were compared with the experts' map template.
• Wallace & Mintzes (1990)	Construct a hierarchical concept map from 10 given concepts on life zones.	Paper and pencil response. Students drew the concept map on a blank page.	Score based on map components: number of relations, levels of hierarchy, branchings, crosslinks, and general to specific examples.
• Wilson (1994)	Construct a hierarchical concept map from 24 chemical equilibrium concepts printed in cards. Subjects could rearranged the concepts until their placement.	Paper and pencil response. Subjects had to glue the cards to a sheet of paper and draw labeled links.	Score of map components: presence or absence of pairs of concepts, hierarchical levels and crosslinks. Nonmetric multidimensional scaling (MSD) was used to transform pairs of concepts into a proximity matrix, and Pathfinder to analyze structure of concepts.

duce a map on this large sheet of paper that shows how you think about these concepts and their interrelations . . . (Mahler, Hoz, Fischl, Tov-Ly, & Lernau, 1991, p. 49).

2. The following map (Figure 1) shows the relation among concepts related to genetic continuity. The concepts are arranged hierarchically and linked to each other. Please examine the map and supply, in the space provided, a word or two for labeling each link such that the association between concepts is made clear (Tamir, 1994, p. 103)
3. Following is a list of concepts related to genetic continuity. Please construct a hierarchic concept map including all the listed concepts. Each link should be accompanied by an appropriate descriptor. The concepts are not listed in the right order (Tamir, 1994, p. 103).

These examples make clear the wide variability in the way concept map tasks elicit students' knowledge structures (Table 1). We identified three ways in which concept-mapping tasks varied: (a) *task demands*, (b) *task constraints*, and (c) *task content structures*.

Task demands refers to the demands made on the students in generating their concept maps. For example, students can be asked to: (a) fill in a skeleton map (e.g., Anderson & Huang, 1989); (b) construct a concept map (e.g., Roth & Roychoudhury, 1993; Wallace & Mintzes, 1990); (c) organize cards (e.g., Baker, Niemi, Novak, & Herl, 1991); (d) rate relatedness of concept pairs (e.g., Acton, Johnson, & Goldsmith, 1994); (e) write an essay (e.g., Lomask et al., 1992); or (f) respond to an interview (e.g., Heinze-Fry & Novak, 1990; Nakhleh & Krajcik, 1991). Even with all of these task variations researchers have assumed that they were eliciting the same thing—students' knowledge structure.

Task constraints refers to the restrictiveness of the task. Constraints varied widely. Students may or may not be: (a) asked to construct a hierarchical map (e.g., Roth & Roychoudhury, 1993; Wilson, 1994); (b) provided with the terms to use in the task (e.g., Barenholz & Tamir, 1992); (c) provided with the labels for the links (e.g., McClure & Bell, 1990); (d) allowed to use more than one link between two nodes (e.g., Fisher, 1990); (e) allowed to move the concepts around physically until a satisfactory structure is arrived at (e.g., Wilson, 1994); (f) asked to define the terms used in the maps (e.g., Mahler et al., 1991); (g) required to justify their responses (e.g., Champagne, Klopfer, DeSena, & Squires, 1978); or (h) required to construct the map collectively (e.g., Roth & Roychoudhury, 1993).

Task content structures refers to the intersection of the task demands and constraints with the structure of the subject domain to be mapped. Methodologically and conceptually, there is no need to impose a hierarchical structure. If the content structure is hierarchical, a hierarchical map should be observed. Harnish, Sato, Zheng, Yamaji, and Connell (in press) proposed different map structures (e.g., spider maps, hierarchic maps, and chain maps) to represent different types of content structures. For example, they proposed the use of chain maps to represent procedural or sequential activities. Furthermore, they suggested that in many cases a combination of different types of concept maps has to be used to accurately represent the structure in a domain.

The usefulness of distinguishing among task demands, task constraints, and content structure is to capture the wide variation among concept mapping tasks concisely with just a few parameters (Table 1). For example, Barenholz and Tamir (1992) asked students to choose 20 to 25 concepts that they considered to be key in microbiology and to construct a map that could summarize what they considered to be the most important topics studied in the course. Anderson & Huang (1989) asked students to fill in a skeleton map consisting of a title, 15 concepts, and six linkages. 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 in Baxter et al., 1993, p. 49). No terms were provided. Nakhleh and Krajcik (1991) conducted a semistructured interview—specific questions about examples and demonstra-

tions—with students about chemical reactions to obtain verbal data to construct students' concept maps.

When all possible combinations of demands, constraints, and content structures are considered, an even wider variety of tasks than those used to date can be produced. We suspect that concept map task variations will tap different aspects of cognitive structure and lead students to produce different concept maps. Preference for one or another type of task should rest on some cognitive theory and the characteristics of the subject-matter domain. That is, both cognitive and subject-matter theories should play an explicit role in guiding the design of concept map assessment tasks by helping the assessment developer decide which combinations are to be preferred over others. How we assess a knowledge structure should be consistent with how we assume knowledge is organized. Research is needed to analyze and compare the effects of different types of tasks that are aimed at measuring the same construct—a student's cognitive structure.

Concept Map Assessment Response Format

Response format refers to the response the student makes, whether drawing a map, writing on computer-generated cards, or giving an oral explanation. As expected, the response is closely related to the characteristics of the task. Three types of response variation were identified in concept mapping: (a) the *response mode*, (b) the *characteristics of the response format*, and (c) the *mapper*.

Type of response mode refers to whether the student's response is paper and pencil, oral, or computer-generated. Students may be asked to: (a) draw (e.g., a concept map) or write (e.g., an essay), (b) say their response (e.g., interview), or (c) enter concept and relation names on a computer.

Format characteristics vary according to the task, usually fitting the specifics of the task. For example, if the task asks the student to fill in a skeleton map and provide the terms for doing so, the response format may look like the one presented in Figure 3. Typically the fill-in-the-blanks (nodes) variant assumes a hierarchical structure as shown in the figure. A set of terms that

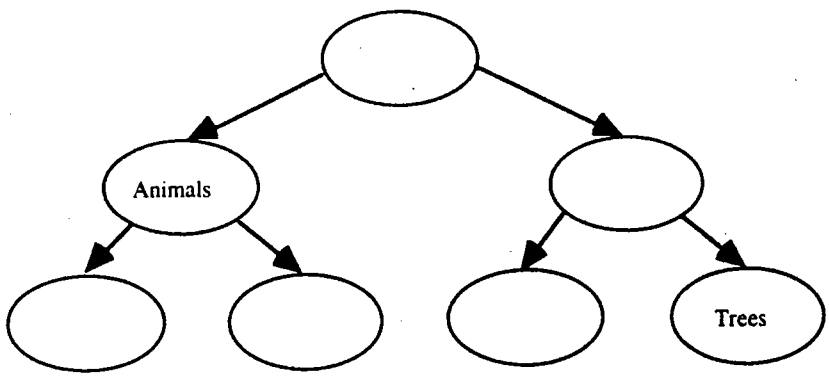

Figure 3. Fill-in-the-blank concept map response format.

includes correct concepts and distractors may be provided, and the student writes those terms in the empty nodes.

If students are asked to draw a map without terms provided, the format will only include a piece of paper with the instructions to construct the map in a specific content domain, whereas if the task asks students to draw the map but provides the terms, the format includes the list of terms. Finally, terms for links between concept pairs may be provided. For example, Figure 4 presents a student's network map in which the concepts and labeled links were provided.

The mapper refers to who draws the map. Most frequently students draw the map. However, when concept maps are developed from students' essays or interviews, researchers have trained teachers (e.g., Lomask et al., 1992) or themselves (e.g., Champagne et al., 1978; Hewson & Hamlyn, date unknown; Nakhleh and Krajcik, 1991) to identify key terms along with the phrases used to link them, and to draw the concept maps for the students. Figure 5 provides

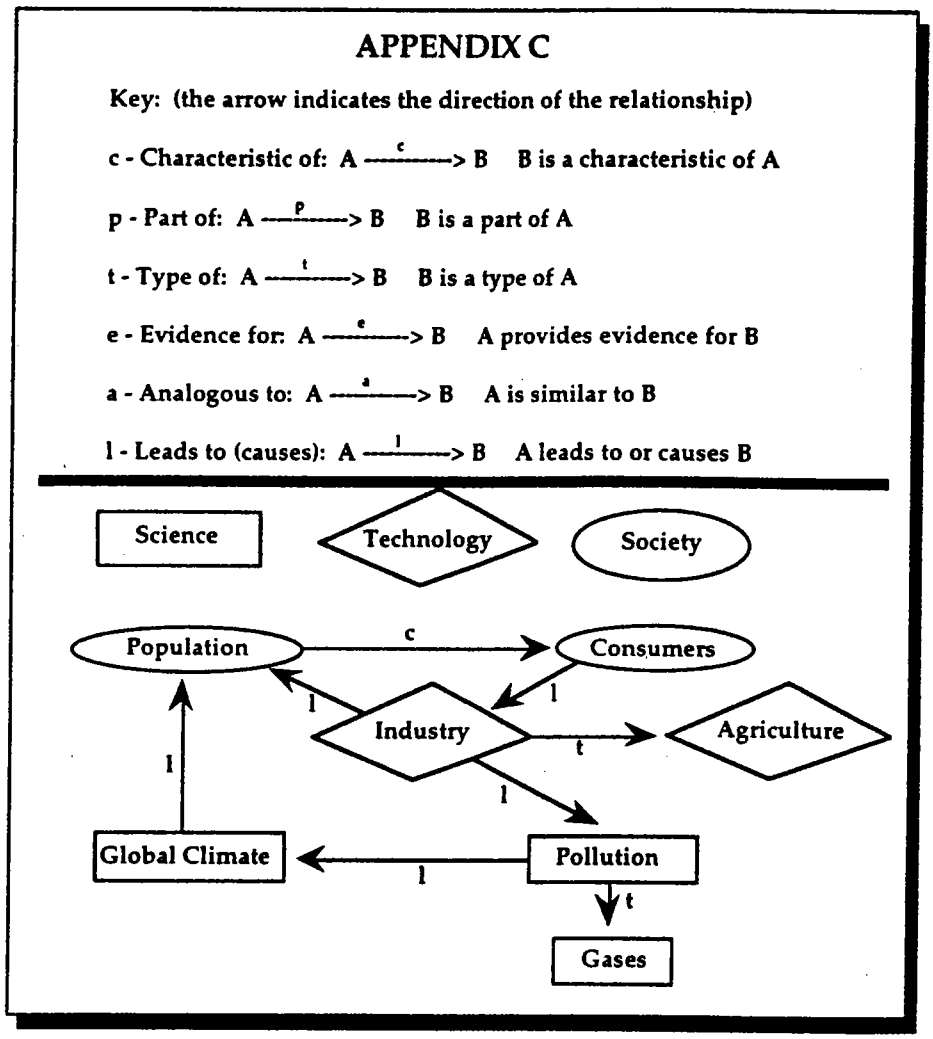

Figure 4. Concept map with semantic labels provided (McClure & Bell, 1990, p. 9).

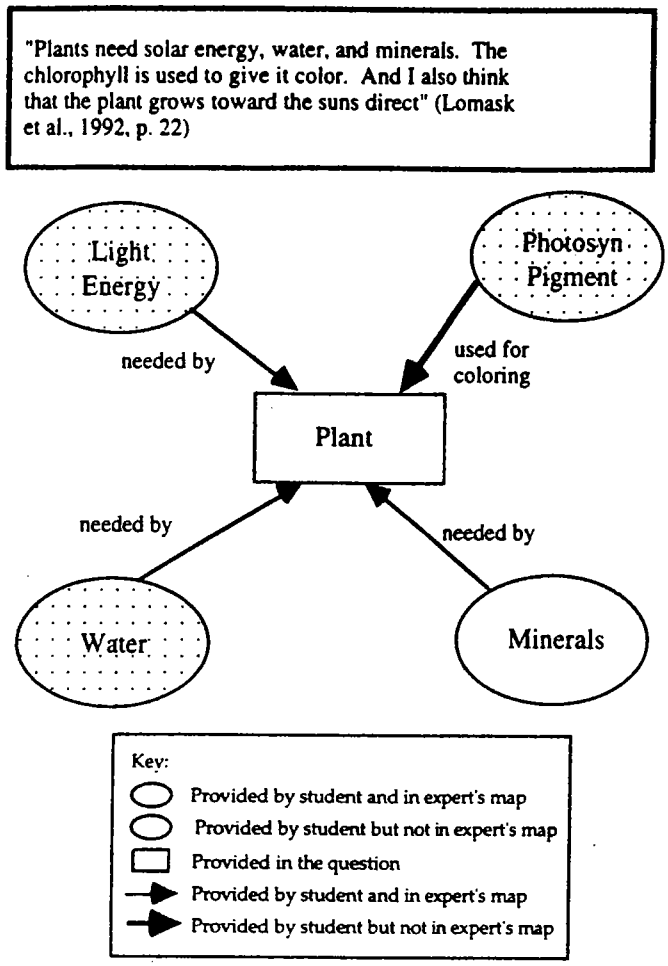

Figure 5. Concept map derived from student essay (Lomask et al., 1992, p. 22).

an example of a concept map constructed from a student's essay written in response to the task of describing possible forms of energies and types of materials in growing plants and how they are related (Lomask et al., 1992).

By taking all possible combinations of mode, characteristic, and mapper (if all exist), a wide variety of response formats can be generated. However, do all produce exchangeable representations of students' cognitive structures, or is one format to be preferred over another? Again, cognitive and subject-matter theory is needed to narrow these alternatives, and empirical research is needed to validate theoretical claims and evaluate reliability.

Concept Map Scoring System

A scoring system is a systematic method with which students' concept maps can be evaluated accurately and consistently. As expected, a myriad of alternative scoring systems can be found. However, they can be classified into three general scoring strategies: (a) score the

components of the student's map, (b) compare the student's map with a criterion map, and (c) use a combination of both strategies.

The first strategy, scoring a student's map components, focuses on three components: propositions (i.e., number, accuracy, crosslinks), hierarchy levels, and examples. Note that not all scoring systems take into consideration all three components. Multiple combinations can be identified from scoring systems that consider all components (Novak & Gowin, 1984) to systems that only consider propositions (McClure & Bell, 1990).

To our knowledge, Novak & Gowin (1984, Table 2.4, pp. 36–37) provided the most comprehensive system to score a student's map components. Table 2 presents the scoring system proposed by these authors. Note that this system is limited to hierarchical maps.

Another scoring system focuses specifically on propositions in the concept map, when two concepts are linked and labeled together via a directional arrow. With this method three parts of the proposition are scored: (a) the existence of a relation between the concepts, (b) the accuracy of the label, and (c) the direction of the arrow indicating either a hierarchical or causal relation between concepts.

McClure and Bell (1990) used propositional scoring to address the question, "How does STS [Science, Technology, and Society] instruction affect cognitive structure?" (p. 2). Figure 6 presents their scoring rules.

The second strategy, the use of a criterion map, compares a student's map with that of an expert and scores the overlap between them. This strategy assumes that there is some ideal organization that best reflects the structure in a domain. Again, different methods have been used to define the expert structure and make the comparison. For example, the criterion map can be defined using the course instructor, individual experts other than the instructor, an average of experts, or an average of top students taking the course (e.g., Acton et al., 1994).

As expected, different methods have also been used to compare the criterion map and the student's map. Lomask et al. (1992) 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

Table 2
Novak and Gowin's (1984) Scoring System

Component	Description	Score
Propositions	Is the meaning relation between two concepts indicated by the connecting line and linking word(s)? Is the relation valid?	1 point for each meaningful, valid proposition shown.
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)?	5 points for each valid level of the hierarchy.
Crosslinks	Does the map show meaningful connections between one segment of the concept hierarchy and another segment?	10 points for each valid and significant crosslink. 2 points for each crosslink that is valid but does not illustrate a synthesis between concepts or propositions.
Examples	Specific events or objects that are valid instances of those designated by the concept level.	1 point for each example.

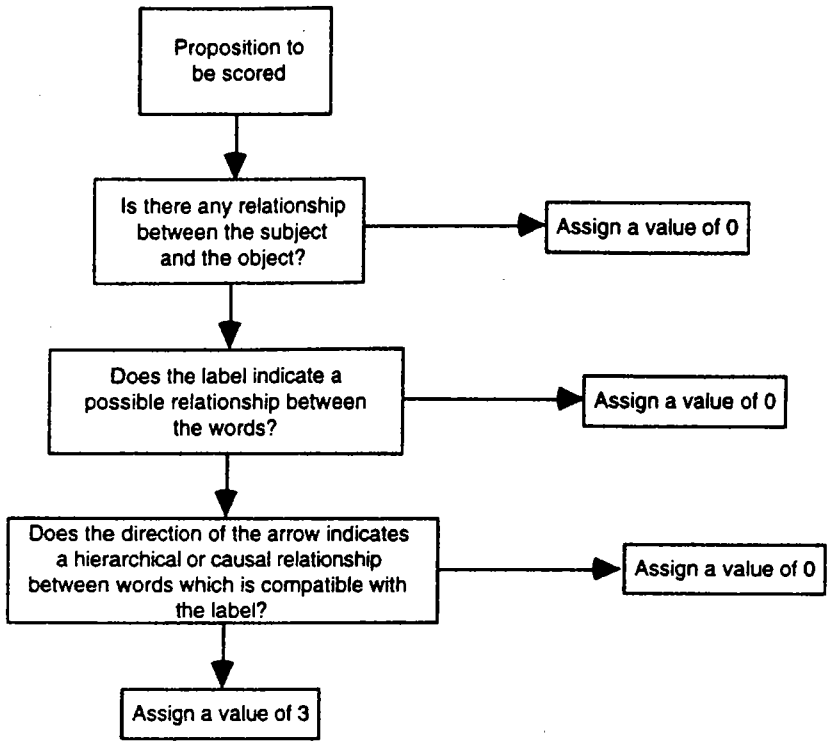

Figure 6. Scoring system for a concept network (McClure & Bell, 1990, p. 10, Appendix D).

concept map mentioned by the student. This proportion was scaled from complete (100%) to substantial (67% to 99%), to partial (33% to 66%), to small (0% to 32%), to none (no terms mentioned or irrelevant terms mentioned). Likewise, they characterized the strength of the links (propositions) between concepts as a proportion of necessary, accurate connections with respect to the expert map. Strength ranged from strong (100%) to medium (50% to 99%), to weak (1% to 49%), to none (0%). Then they provided a table that produced scores taking into account both the size of terms and the strength of links (Table 3).

Table 3
Scores Based on Combinations of Size and Strength of Students' Concept Maps

		Strength		
Size	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

A third strategy is a combination of map components and a criterion map. Following this strategy, Novak and Gowin (1984) added a fifth rule to their scoring system: comparison. They proposed scoring the criterion map according to their four rules, then divide the student's map score 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.

Hoz et al. (1990) also combined the two strategies to score teachers' concept maps on conceptual, disciplinary and pedagogical knowledge. They scored the biconcept links, the map as a whole, and the concept groups on a teacher's map. However, they also used an expert's link-strength matrix to validate teachers' map links and knowledge structure. The validity of the links was determined by the experts' link-strength matrix: a score of 2 was assigned for a mandatory link, 1 for a possible link, and 0 for a forbidden link. Table 4 presents how each component was scored. A biconcept link was considered to be invalid when the score was 1 and 0. On this basis they defined the validity of a knowledge structure as the percentage of the sum total of the validity scores out of the maximal validity score.

White and Gunstone (1992) proposed another combination: Count the number of linked concept pairs. The links can be hierarchic, multiple, or crosslinked. Points are given for the number of links that are the same as those on a criterion map (i.e., expert'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 number of links by the number of semantic categories (Mahler et al., 1991).

The strategy most frequently used in research is to score the components of the student's map using the system developed by Novak and Gowin (1984) or a modification of the Novak–Gowin approach (e.g., Anderson & Huang, 1989; Beyerbach, 1988; Lay-Dopyera & Beyerbach, 1983; Schreiber & Abegg, 1991). However, the possible scoring strategies require research to evaluate their ability to provide useful and valid information about students' cognitive structures in a content domain. For example, which criterion map provides a more valid reference—the course instructor's map, an expert's map, or the average of a group of experts' maps? Acton et al. (1994) evaluated the use of different criterion maps for predicting students' performance on standard classroom measures in two college-level computer-programming courses. The study compared different criterion maps: those produced by the course instructor, individual experts, an average of experts, and an average of the best students in the class. They

Table 4
Example of a Combined Scoring System (Hoz et al., 1990)

Component	Dimension	Score
Biconcept links	Validity of the link	Four-level ordinal scale
Map as a whole	Validity of the knowledge structure	
	• Congruence	Number of valid links in the map divided by the number of legitimate links presented in the criterion map
	• Salience	Number of valid links in the map divided by the number of links in the map.
Concept group	Quality of a group	Proportion of the sum total of the 3 partial scores (i.e., homogeneity—three-level ordinal scale; structure—four-level ordinal scale; title fit—four-level ordinal scale) out of 7, the maximal score.

found that using individual experts' maps as the criterion was highly variable in predicting students' performance. However, when the average rating of experts was used as a reference, the coefficient increased. Equally important was the finding that some experts were better referents than others. We wonder whether using different criterion measures, such as performance-based assessments instead of multiple-choice tests, would lead to different conclusions. Still, these findings indicate that selecting a criterion map to score students is problematic. Different experts' maps may lead to different conclusions about a student's knowledge structure.

In summary, an assessment can be defined as the combination of a task, response, and scoring system. Based on this definition, multiple concept-mapping techniques can emerge from variations in tasks, response formats, and scoring systems that we have identified in the literature (Table 5). Unfortunately we cannot look to cognitive theory to decide which technique to prefer, because many of the techniques reviewed had no direct connection with such a theory. We believe a connection should be made between how we assess knowledge structure and our conception of how knowledge is organized. With such a connection, data obtained from the application of a mapping technique would provide an empirical evaluation of the theory. This form of representational validation has been largely lacking in the cognitive literature (e.g., Johnson & Goldsmith, 1992).

Nevertheless, current practice holds that all variations in mapping techniques are interpreted in the same way, as representing a student's cognitive structure—the relations of concepts in a student's memory (Shavelson, 1972). How can the variations produce somewhat different representations and scores of goodness of cognitive structure, yet all be interpreted as a measure of the same thing? What evidence is there that concept maps provide a reasonable representation of a student's cognitive structure? What evidence is there that concept map scores are reliable? We next review the psychometric research evaluating the technical qualities of concept maps as an assessment tool.

Technical Qualities of Concept Maps as Assessments

As noted earlier, concept maps have been used more for instruction than for formal assessment. As a consequence authors have rarely addressed issues of the reliability and validity of concepts maps. Even more difficult to find are studies whose main focus was the systematic examination of the technical qualities of concept maps as a form of assessment. In this section we review past research and forecast new research needed to evaluate concept maps as an assessment tool in science. Table 6 summarizes the information on reliability and validity available from the studies reviewed.

Reliability

Reliability refers to the consistency (or "generalizability"; cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972) of scores assigned to students' concept maps. Few studies provided this information, and when they did they focused on interrater reliability or agreement (Table 6). Barenholz and Tamir (1992), for example, reported that the interrater agreement estimated in their study was always above 80%. However, they did not specify the procedure used to establish agreement or the number of raters involved.

Lay-Dopyera and Beyerbach (1983) reported a reliability coefficient for each map component evaluated. For the number of concepts in maps, they reported an interrater (n_r = 2) reliability coefficient of .99. For the number of levels found in the maps, the coefficient was .80, and for the number of item streams—each line drawn out from the central concept that led to

Table 5
Concept Map Components and Variations Identified

Map assessment components	Variations	Instances
Task	• Task demands	Students can be asked to: • fill in a map • construct a map from scratch • organize cards • rate relatedness of concept pairs • write an essay • respond to an interview
	• Task constraints	Students may or may not be: • asked to construct a hierarchical map • provided with the concepts used in the task • provided with the concept links used in the task • allowed to use more than one link between nodes • allowed to physically move the concepts around until a satisfactory structure is arrived at • asked to define the terms used in the map • required to justify their responses • required to construct the map collectively
	• Content structure	The intersection of the task demands and constraints with the structure of the subject domain to be mapped.
Response	• Response mode	Whether the student response is: • paper and pencil • oral • computerized
	• Format characteristics	Format should fit the specifics of the task
	• Mapper	Whether the map is drawn by a: • student • teacher or researcher
Scoring system	• Score components of the map	Focus is on three components or variations of them: • propositions • hierarchy levels • examples
	Use of a criterion map	Compare a student's map with an expert's map. Criterion maps can be obtained from: • one or more experts in the field • one or more teachers • one or more top students
	• Combination of map components and a criterion map	The two previous strategies are combined to score the student's map.

Table 6
Information on the Psychometric Characteristics of Concept Maps

Authors	Reliability	Validity
• Acton et al. (1994)	No information provided.	*Predictive validity.* The study evaluated the ability of eight different criterion maps to predict students' performance on an exam. Results indicated great variability among individual-based criterion maps. *Known group differences.* The study also evaluated the structural similarity of maps constructed by subjects at different levels of domain expertise. Substantial variability among experts was found. In general, experts' maps were more similar to each other than students' maps.
• Anderson & Huang (1989)	No information provided.	*Concurrent validity.* Correlation between six school achievement measures (e.g., Otis-Lennon School Ability, Stanford Vocabulary Achievement Test) and scores on a mapping test were obtained. Correlations ranged from .49 (with the Science Grade School) to .74 (with the Otis-Lennon). *Instructional sensitivity.* The study compared students who only read the material with those who received specific instruction. Differences between groups—reading/instruction—were significant when mapping scores were used, but not when short answers scores were used. It was concluded that concept maps were sensitive to the gain in knowledge as a result of the instruction.
• Baker et al. (1991)	Reliability coefficients or percentage of agreement not provided. Authors said that raters independently judged the quality of the representations.	*Content validity.* Three independent experts judged the passage and concepts as appropriate of the content area. Two other experts developed the criterion map. *Convergent validity.* Concept map scores were correlated with two other assessments: an essay and a short-answer test of prior knowledge. Correlation coefficients with the essay were .51 for the history topic and .30 for the science topic. Correlations with the prior knowledge measure were .38 for the history and .42 for the science topic.

(*continued*)

Table 6 (*Continued*)

Authors	Reliability	Validity
• Barenholz & Tamir (1992)	*Interrater agreement.* Percentage of agreement exceeded 80%. No information about the procedure used to estimate the percentage of agreement (e.g., agreement based on total score or partial scores; correction for chance agreement).	*Content validation.* Validity of the instrument was established by a panel of 4 science experts. No mention of the procedure used.
• Beyerbach (1988)	No information provided.	*Instructional sensitivity.* The study compared students' concept map scores before and after instruction. Posttest scores indicated conceptual growth after instruction. Significant differences with the pretest were found. *Similarity with a criterion map.* Students' maps became more similar to instructors' scores after the instruction.
• Champagne et al. (1978)	No information provided.	*Instructional sensitivity.* The study compared students' maps before and after instruction. From qualitative analysis of the maps, it was concluded that students' representations changed (improved) as the result of instruction.
• Fisher (1991)	No information provided	No information provided
• Heinze-Fry & Novak (1990)	No information provided	No information provided
• Hewson & Hamlyn (date unknown)	*Interrater agreement.* The scores of 5 independent judges on two transcripts were compared with the original analyses. Agreement between the judges was 97.5% for metaphorical heat and 82.1% for the physical heat.	No information provided
• Hoz et al. (1990)	No information provided.	No information provided.
• Lay-Dopyera & Beyerbach (1983)	*Interrater reliability.* Reliability was calculated for each aspect of the map that was scored. Results indicated coefficients of .99 for number of items, .80 for number of levels, and nearly perfect coefficient for number of item streams. *Stability.* Reliability coefficients were obtained for each map component: .73 for items, .21 for level, .75 for streams. It	*Instructional sensitivity.* Comparison of concept maps before and after instruction revealed increases in the number of items, levels, and streams on the map. *Known group differences.* Differences in the number of nodes and subordinate levels on the concept maps constructed by subjects with different educational levels and professional experience were found. When the focus was only on professional experience, years of teaching were not related to map scores.

	was concluded that concept mapping was not a stable measure across time.	
• Lomask et al. (1992)	*Interrater reliability.* Four trained teachers scored students' responses by constructing the maps from essays. Reliability coefficients, averaged across the two topics, were .87 for number of concepts, .81 for correct connections, and .72 for expected connections of the map.	*Content validity.* It was established by teachers' review of the essay questions based on the content of the units in most Connecticut schools and the characteristics of the questions (e.g., the knowledge required to answer the questions).
• Mahler et al. (1991)	No information provided.	*Content validity.* Validity of map components was established by discipline experts who continuously reviewed and revised the measures when necessary.
• Mc Clure & Bell (1990)	No information provided.	*Instructional sensitivity.* Comparison of concept maps before and after instruction revealed significant mean increase in the number of propositions identified. *Validity.* Correlations between a concept map score and a final examination score was .50. Authors interpreted this correlation to mean that concept map scores represented the course content.
• Markham et al. (1994)	No information provided.	*Known group differences.* The study examined the differences between advanced biology majors and beginning nonmajors. Results indicated that biology majors' concept maps were structurally more complex. *Concurrent validity.* The study examined how concept maps reflect predictable patterns of knowledge usage in another task (card sorting). Results indicated that organizational patterns depicted in concept maps were also reflected in the ways students sorted the cards.
• Nakhleh & Krajcik (1991)	*Interrater reliability.* Three concept maps were constructed and three others were scored by a second rater. All of them were compared with the three maps constructed and scored previously by the researchers. Percent agreement between second rater and the researchers was .82 for the maps constructed and .83 for the maps scored.	No information provided

(continued)

Table 6 (*Continued*)

Authors	Reliability	Validity
• Novak et al. (1983)	No information provided.	*Concurrent validity.* Correlation between six school achievement measures (e.g., SAT Reading, SAT Math, SCAT Verbal, SCAT quantitative, final examination grade) and scores on the concept maps were obtained. Correlations ranged from −.02 (with SAT Reading) to .34 (with SCAT verbal).
• Roth & Roychoudhury (1993)	No information provided.	*Known group differences.* The study examined the differences between good and poor performers. Results indicated that good performers' concept maps had more valid levels and concepts.
• Schreiber & Abegg (1991)	No information provided.	*Concurrent validity.* A significant positive correlation (coefficient was not provided) was found between concept map scores and the group assessment of logical thinking—the higher the student's map scores the more formal the student's reasoning.
• Wallace & Mintzes (1990)	No information provided.	*Instructional sensitivity.* Brief instructional interventions yielded substantial differences in concept mapping scores in favor of the group receiving instruction.
• Wilson (1994)	No information provided.	*Known group differences.* Comparison of concepts maps structure among chemistry teachers and students. In general, teachers and high-achievement students revealed more crosslinkages and levels of abstraction than low-achievement students. It was concluded that multilevel knowledge structures with many connections within and between levels characterize expertise in a domain.

one or more words—they reported nearly perfect agreement. Note that the reliability coefficients were based on counts of nodes and levels. We suspect that interrater reliability would not have been so high if different scoring criteria had been used, such as the number of valid links. In addition, this was the only study to examine retest reliability of concept map scores (reliability across time). Reliability coefficients were obtained for each map component: .73 for items, .21 for level, and .75 for streams. The authors concluded that the study failed to establish concept map scores as stable measures.

Nakhleh and Krajcik (1991) trained a doctoral student in science education to construct and score maps from students' interviews. Three concept maps constructed and another three scored by this second rater were compared with maps constructed and scored previously by the researchers. A total of 82% agreement was obtained (by dividing the number of agreed-upon nodes by the total number of nodes on the original map) for the maps constructed, and .83% agreement for the scored maps (by dividing the total number of agreed-upon relations by the total number of relations on the original maps). We wonder what the agreement coefficient would have been had more than three maps been used in each case.

Only Lomask et al. (1992) extensively studied the reliability with which concept maps collected in a statewide assessment could be scored—the only study in which concept maps were used in large-scale statewide assessments. (Recall that the 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 that of an expert. Three features of the concept maps were scored: (a) number of concepts in the expert map used by the student; (b) number of correct connections among concepts; and (c) expected number of connections among all of the concepts mentioned in an essay.

The findings of Lomask et al. (1992) are summarized in Table 7, two of which stand out. The first is that raters introduced negligible error in estimating the level of a student's performance (rater effect 0 or close to 0) (Table 7). This finding, however, must be interpreted cautiously, because Lomask et al. eliminated from the analysis data from one teacher whose scores deviated substantially from those of the other teachers. Nevertheless, this finding is consistent with Hewson and Hamlyn's (date unknown) finding that agreement among five independent judges evaluating two interview transcripts was 97.5% for metaphoric heat concep-

Table 7
Person × Rater Generalizability Studies (Adapted from Lomask et al., 1992, Table 4): Percentage 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 × R, *e*	16	23	18	10	15	29
Reliability[b]	0.84	0.77	0.81	0.89	0.84	0.62

[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" (Lomask et al., 1992, Table 4).
[b] Generalizability coefficient for absolute decisions (f); 1 rater (see Shavelson & Webb, 1991).

tions and 82.1% for physical heat conceptions. Moreover, the literature on performance assessment has established that raters can be trained to score complex performance reliably (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).

Even though seldom studied, reliability of concept map scores is an important issue that must be addressed before the scores are reported to teachers, students, the public, or 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 or learning has intervened?
- How large a sample of concept maps might be needed to assess students' knowledge in a science domain reliably?
- If terms are given, 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.

Content Validity. Cognitive theory posits that the interrelations among concepts is an essential property of knowledge. Moreover, research has accumulated evidence that the structural properties of domain knowledge are closely related to competence in the domain. One important criteria for evaluating content validity, therefore, is the use of experts in a domain to judge the representativeness of concepts and the accuracy of maps within that subject domain. Only a few concept map studies reported that on logical grounds, experts judged the terms and maps as consistent with the subject domain (e.g., Anderson & Huang, 1989; Barenholz & Tamir, 1992; Lomask et al., 1992; Nakhleh & Krajcik, 1991).

Concurrent Validity. Empirically several studies showed consistent correlations between concept map scores and other measures of student achievement or accomplishment, whereas other studies suggested that scores derived from concept maps seem to measure a different aspect of achievement than that measured by multiple-choice tests. For example, Anderson and Huang (1989) reported substantial correlations (range .49 to .74) between concept map scores (e.g., total number of propositions in a map) and measures of aptitude and science achievement (e.g., Stanford Science Achievement Test, school science grades, and Ottis Lennon School Ability Test). McClure and Bell (1990) found correlations above .50 between concept map scores and other measures of student achievement (e.g., final examination score). However, Novak, Gowin, and Johansen (1983) reported different correlations. Correlations between concept maps scores, final course grades, and the conventional measures of learning such as scholastic aptitude tests were close to zero; for example, the correlation between concept map scores and (a) final examination grade was .02, (b) SAT reading score was −.02, and (c) SAT math scores was .02. They interpreted these correlations as evidence that multiple-choice tests measure a different type of learning (viz., "rote learning") than the learning measured by concept maps (viz., "meaningful learning").

Instructional Sensitivity. Other studies focused on the instructional sensitivity of concept maps. In these studies, concept maps were typically given in a pre- and posttreatment design. In general, despite large differences in the quality of methods used, findings were consistent: The structure and organization of concept maps differed (improved) after students received instruction on a specific topic (e.g., Anderson & Huang, 1989; Beyerbach, 1988; Champagne et al., 1978; Lay-Dopyera & Beyerbach, 1983; Wallace and Mintzes, 1990). Some studies focused on counting nodes (e.g., Lay-Dopyera & Beyerbach, 1983); others focused on structural complexity and organizational patterns (e.g., Wallace and Mintzes, 1990).

Known Group Differences. Still other studies examined the ability of concept maps to differentiate among groups varying in expertise in the content domain evaluated. Maps from experts, instructors, and advanced students have been compared with those of beginning students. Lay-Dopyera and Beyerbach (1983) found differences in the number of nodes and levels used in maps constructed by subjects differing in educational level and professional experience on the topic of teaching. However, when the focus was only on professional experience, years of teaching were not related to map scores. Markham, Mintzes, and Jones (1994) compared the concept map scores of advanced biology majors and beginning nonmajors and concluded that the structure and organization of the concept maps were more complex for biology majors. Similar results were found by Wilson (1994), who compared the concept map structure among chemistry teachers and high- and low-achieving students. She found that experts (i.e., teachers and high-achieving students) constructed multilevel knowledge structures with many connections within and between levels, whereas low-achieving students constructed simpler maps (see also Roth & Roychoudhury, 1993).

Acton et al. (1994) compared the maps constructed by three groups ranging in expertise in physics: course instructors, individual experts other than the instructors, and top graduate students. Map mean scores were higher for experts (i.e., instructor and experts) than top students, and experts' maps varied substantially among themselves. However, when similarity among experts' maps was compared with that among students' maps, the experts' maps were more similar among themselves than the group of students' maps.

These studies indicate that concept maps can distinguish between experts and novices in a subject domain. However, results also suggest that experts' concept maps are not as similar as we would want and expect (Acton et al., 1994). One implication of these results is directly related to the common scoring practice of using criterion maps to score students' maps. (See discussion in Concept Map Scoring System.) The findings of Acton et al. (1994) show that selecting a criterion map to score students' map is problematic, as different experts' maps lead to different conclusions about students' knowledge structures.

Process Tracing. Finally, few studies have focused on a systematic evaluation of the validity of the cognitive-structure interpretations of concept map scores. The study conducted by Baxter et al. (1993) is an exception. After a detailed protocol analysis derived from interviews with the students who participated in the study of Lomask et al. (1992), these authors concluded that concept map scores overestimated what students understand. They discussed the possibility that the characteristics of the scoring system used was associated with this discrepancy.

Construct validation studies of concept map techniques need to be carried out before scores

from such assessments are reported to teachers, students, the public, and policy makers. Among the questions that need to be addressed are:

- 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 map 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, such as cognitive structure?
- Do concept map scores unfairly distinguish among diverse student groups varying in socioeconomic status, race or 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?

In summary, as concept maps are increasingly used as assessments in science education, we need to provide evidence that they tap some important aspect of students' knowledge structures and provide reliable and valid scores. The process of collecting this evidence should start during the development of the concept map assessment. Decisions about the characteristics of the task, response format, and scoring system should be based on a theoretical framework. Then research should focus on obtaining various types of evidence on reliability and validity until this evidence establishes the value of the concept map as a structural representation of what a student knows in a content domain.

Conclusions

A concept map is an structural representation that purports to tap some important aspect of the structure of a student's knowledge in a subject domain. The use of concept maps to supplement traditional multiple-choice tests has attracted attention among educators and researchers. We characterized concept map–based assessments as: (a) a task that invites students to provide evidence bearing on their knowledge structure in a domain; (b) a format for the students' response, and (c) a scoring system by which concept maps can be evaluated accurately and consistently. Without these three components, a concept map cannot be considered to be an assessment.

By taking into account all possible tasks, response formats, and scoring options reported in the literature, our characterization has made evident the enormous variation in concept-mapping techniques, which in turn produces different representations and scores. Nevertheless, all are assumed to measure the same construct: a student's cognitive structure.

Before concept maps are used for either classroom or large-scale assessment, and before concept map scores are reported to teachers, students, the public, and policy makers, research needs to provide reliability and validity information on the effect of different techniques on assessing a student's cognitive structure. The following issues need to be examined empirically:

Concept Map Assessment Techniques

Multiple techniques in concept mapping can emerge from the variations of tasks, response formats, and scoring systems (Table 5). For example, if each of the six task demands identified (e.g., fill in a map) is combined with each of the eight types of task constraints (e.g., hierarchi-

cal vs. nonhierarchical), there are no fewer than 1,530 (i.e., $6 \times 2^8 - 1$) different ways to produce a concept-mapping task! Of course, not all combinations are realistic.

From this example, it is clear that concept map techniques can vary widely in the way they elicit a student's knowledge structure. Preference for one or another technique should lie at the intersection of some cognitive theory and structural notions of the subject domain to be mapped. There is thus the need for a theory to drive the development of concept map assessments. Such theory, along with empirical research, should provide guidelines that would eventually narrow the number of possible techniques to a manageable set.

As this cognitive theory develops, research on concept maps should proceed by applying reasonable criteria that can help discard some techniques. Criteria such as differences in the cognitive demands required by the task, appropriateness of a structural representation in a content domain, appropriateness of the scoring system for evaluating accuracy of the representation, and practicality of the technique deserve to be explored. For example, we believe that the fill-the-blank task should be regarded as inappropriate for measuring a student's knowledge structure, because the task itself too severely restricts the representation. We also think that imposing a hierarchical structure, regardless of content domain, is also inadequate, because an accurate concept map representation of a hierarchical domain will be hierarchical itself. Furthermore, not all subject-matter domains have a hierarchical structure (see, for example, Shavelson, 1972). In addition, we favor scoring criteria that focus more on the adequacy of the propositions over those that focus simply on counting the number of map components (i.e., nodes and links). Finally, if concept maps are to be used in large-scale assessment, mapping techniques that require one-on-one interaction between student and tester should be also discarded on practical grounds.

We also believe that the participation of science educators, experts, and instructors in designing a concept map assessment is essential. Their expertise in a subject-matter domain can be used not only to determine the appropriateness of certain map structures (e.g., Can the content domain of mechanics be structured hierarchically?), but also to design a criterion map (e.g., How do experts represent their knowledge on a domain? Are their representations similar? How can we select the best criterion map to score students?), and judge the concepts and the maps as consistent with a subject domain.

Reliability of Map Scores

One psychometric issue that needs to be addressed with concept maps is whether they can provide reliable scores and representations. How reliable are map scores across raters? Although results reported in the literature suggested high coefficients, these findings should be interpreted cautiously because of the scoring criteria used (e.g., counting map nodes and levels instead of focusing on the validity of the map propositions). How large a sample of concept map tasks is needed to measure a student's knowledge structure reliably in a content domain? Research on science performance assessments (e.g., Shavelson et al., 1991) has found task sampling to be a major source of unreliability. How stable are map scores across time? With science performance assessments, we found that occasion sampling was also a major source of unreliability (Ruiz-Primo, Shavelson, & Baxter, 1993).

Validity of Map Inferences

Beyond reliability it is essential to justify proposed interpretations of concept maps as measures of a student's knowledge structure in a given domain. Do concept maps provide a

sensible representation of knowledge in a domain as judged by subject-matter experts? Do process-tracing studies converge on the same knowledge represented in a map? Do different mapping techniques provide the same information about a student's knowledge structure? Do alternative techniques provide information about science understanding beyond what is already known from traditional multiple-choice tests? Do different assessment techniques correlate differently with traditional multiple-choice tests?

Practical Issues

Research should also focus on the use of concept map assessments in classrooms and in large-scale contexts. Three important related issues can be identified immediately. The first issue has to do with students' facility in using the technique (Carey & Shavelson, 1989), the second with how to train students to create concept maps in a brief period of time, and the third with the consequences of teachers teaching to the test (cf. Shavelson, Carey, & Webb, 1990). Concept maps can potentially increase teachers' repertoire of instructional and assessment techniques. However, teachers may teach to the test. They may, for example, present an expert/criterion map to students and require them to memorize the map. 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.

The research agenda is long but necessary if we want to test the potential of concept maps as an alternative assessment in science. One fourth-grade teacher wrote to us about the potential of concept maps as an assessment tool:

> 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 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.

This study was supported in part by the Center for Research on Evaluation, Standards, and Student Testing (Grant R117G10027). However, the opinions expressed here represent those of the authors and not necessarily those of the funding agency. The authors are deeply grateful to Heather Lange and Bridget Lewin for their valuable comments on previous versions of the article.

Notes

[1] Notable exceptions are Baxter et al. (1993), and Magone, Cai, Silver, and Wang (1994).

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

[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, we will follow this convention.

References

Acton, W.H., Johnson, P.J., & Goldsmith, T.E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86*, 303–311.

Anderson, T.H., & Huang, S.-C.C. (1989). *On using concept maps to assess the comprehension effects of reading expository text* (Technical Report No. 483). Urbana-Champaign: Center for the Studying of Reading, University of Illinois at Urbana-Champaign. (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. (1991, July). *Hypertext as a strategy for teaching and assessing knowledge representation*. Paper presented at NATO Advanced Research Workshop on Instructional Design Models for Computer-Based Learning Environments, Enschede, The Netherlands.

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

Baxter, G.P., Glaser, R., & Raghavan, K. (1993). *Analysis of cognitive demand in selected alternative science assessments*. Report for the Center for Research on Evaluation, Standards, and Student Testing. Westwood, CA: UCLA Graduate School of Education.

Beyerbach, B.A. (1988). Developing a technical vocabulary on teacher planning: Preservice teachers' concept maps. *Teaching & Teacher Education, 4*, 339–347.

Briscoe, C., & LaMaster, S.U. (1991). Meaningful learning in college biology through concept mapping. *The American Biology Teacher, 53*, 214–219.

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 Shavelson, R.J., McDonnell, L.M., & Oakes, J. (Eds.), *Indicators for monitoring mathematics and science education: A sourcebook* (pp. 147–191). Santa Monica, CA: RAND Corporation.

Champagne, A.B., Klopfer, L.E., DeSena, A.T., & Squires, D.A. (1978). *Content structure in science instructional materials and knowledge structure in students' memories* (Report No. LRD-1978/22). Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh. (ERIC Document Reproduction Service No. ED 182 143).

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

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: Johns Hopkins Press.

Dochy, F.J.R.C. (1994). Assessment of domain-specific and domain-transcending prior knowledge: Entry assessment and the use of profile analysis. In M. Birenbaum & F.J.R.C. Dochy (Eds.), *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 93–129). Boston: Kluwer Academic.

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 in Science Teaching, 27*, 1001–1018.

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

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology, 40*, 631–666.

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

Harnisch, D.L., Sato, T., Zheng, P., Yamaji, S., & Connell, M. (in press). Concept mapping approach and its implications in instruction and assessment. *Computers in the Schools*.

Heinze-Fry, J.A., & Novak, J.D. (1990). Concept mapping brings long-term movement toward meaningful learning. *Science Education, 74*, 461–472.

Hewson, M.G., & Hamlyn, D. (date unknown). *The influence of intellectual environment on conceptions of heat*. Johannesburg: National Institute of Personnel Research. (ERIC Document Reproduction Service No. ED 231 655).

Holley, C.D., & Danserau, D.F. (1984). The development of spatial learning strategies. In C.D. Holley & D.F. Danserau (Eds.), *Spatial learning strategies. Techniques, applications, and related issues* (pp. 3–19). Orlando: Academic Press.

Hoz, R., Tomer, Y., & Tamir, P. (1990). The relations between disciplinary and pedagogical knowledge and the length of teaching experience of biology and geography teachers. *Journal of Research in Science Teaching, 27*, 973–985.

Johnson, P.J., & Goldsmith, T.E. (1992). *Structural assessment of knowledge and skill* (Technical Report). Office of Navy Research, Washington, DC.

Jonassen, D.H., Beissner, K., & Yacci, M. (1993). *Structural knowledge. Techniques for representing, conveying, and acquiring structural knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lay-Dopyera, M., & Beyerbach, B. (1983). *Concept mapping for individual assessment*. Syracuse, NY: School of Education, Syracuse University. (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 Science Teaching, Cambridge, MA.

Magone, M., Cai, J., Silver, E.A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research, 21*, 317–340.

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

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

Markham, K.M., Mintzes, J.J., & Jones, M.G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching, 31*, 91–101.

Nakhleh, M.B., & Krajcik, J.S. (1991). *The effect of level of information as presented by different technology on students' understanding of acid, base, and pH concepts*. Paper presented

at the annual meeting of the National Association for the Research in Science Teaching, Lake Geneva, WI. (ERIC Document Reproduction Service No. ED 347 062).

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

Novak, J.D., Gowin, D.B., & Johansen, G.T. (1983). The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education, 67*, 625–645.

Pankratius, W. (1990). Building an organized knowledge base: Concept mapping and achievement in secondary school physics. *Journal of Research in Science Teaching, 27*, 315–333.

Resnick, L.B., & Resnick, D.P. (1990). 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*. Boston: Kluwer Academic.

Roth, W.-M., & Roychoudhury, A. (1993). The concept map as a tool for the collaborative construction of knowledge: A microanalysis of high school physics students. *Journal of Research in Science Teaching, 30*, 503–534.

Royer, J., Cisero., C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201–224.

Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*, 41–53.

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 the 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., & 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. (1990). 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.

Schmid, R.F., & Telaro, G. (1990). Concept mapping as an instructional strategy for high school biology. *Journal of Educational Research, 84*, 78–85.

Stice, C.F., & Alvarez, M.C. (1987). Hierarchical concept mapping in the early grades. *Childhood Education, 64*, 86–96.

Tamir, P. (1994). Science assessment. In M. Birenbaum & F.J.R.C. Dochy (Eds.), *Alternatives in assessment of achievements, learning process and prior knowledge* (pp. 93–129). Boston, MA: Kluwer Academic.

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*, 1033–1052.

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

White, R.T. & 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.

Willerman, M., & Mac Harg, R.A. (1991). The concept map as an advance organizer. *Journal of Research in Science Teaching, 28*, 705–711.

Wilson, J.M. (1994). Network representations of knowledge about chemical equilibrium: Variations with Achievement. *Journal of Research in Science Teaching, 31*, 1133–1147.

Received March 24, 1995
Revised January 22, 1996
Accepted January 26, 1996