

Using Concept Maps to Assess Student Learning in the Science Classroom: Must Different Methods Compete?

Diana C. Rice,¹ Joseph M. Ryan,² Sara M. Samson³

¹*School of Education, University of South Carolina Aiken, 171 University Parkway,
Aiken, South Carolina 29801*

²*Research Consulting Center, 4701 W. Thunderbird Road, Arizona State University–West,
Phoenix, Arizona 85069*

³*Instructional Services, Midlands Technical College, Columbia, South Carolina 29205*

Received 15 July 1997; revised 14 November 1997; accepted 20 March 1998

Abstract: This yearlong study was implemented in seventh-grade life science classes with the students' regular teacher serving as teacher/researcher. In the study, a method of scoring concept maps was developed to assess knowledge and comprehension levels of science achievement. By linking scoring of concept maps to instructional objectives, scores were based upon the correctness of propositions. High correlations between the concept map scores and unit multiple choice tests provided strong evidence of the content validity of the map scores. Similarly, correlations between map scores and state criterion-referenced and national norm-referenced standardized tests were indicators of high concurrent validity. The approach to concept map scoring in the study represents a distinct departure from traditional methods that focus on characteristics such as hierarchy and branching. A large body of research has demonstrated the utility of such methods in the assessment of higher-level learning outcomes. The results of the study suggest that a concept map might be used in assessing declarative and procedural knowledge, both of which have a place in the science classroom. One important implication of these results is that science curriculum and its corresponding assessment need not be dichotomized into knowledge/comprehension versus higher-order outcomes. © 1998 John Wiley & Sons, Inc. *J Res Sci Teach* 35: 1103–1127, 1998.

The development of concept mapping can be traced back to the well-known work of Ausubel, Novak, and Gowin in the early 1970s. Since its introduction, the concept map has become a useful and sometimes invaluable tool in science education research as well as in science teaching. The prominent role of concept mapping in science education was reflected in the publication in 1990 of a special issue of the *Journal of Research in Science Teaching* on the topic of concept mapping, which included an article listing 100 references related to the use of concept mapping (Al-Kunified & Wandersee, 1990).

A review of the literature reveals that concept maps have been used to assess or describe a variety of constructs and outcomes (Anderson & Huang, 1989; Lay-Dopyera & Beyerback, 1983; Lomask, Baron, Greig, & Harrison, 1992; Malone & Dekkers, 1984; Markham, Mintzes,

Correspondence to: D.C. Rice

& Jones, 1994; Novak & Gowin, 1984; Novak & Musonda, 1991; Powers & Wright, 1992; Roth & Roychoudhury, 1993; Starr & Krajcik, 1990; Wallace & Mintzes, 1990) as well as in instruction (Jegede, Alaiyemola, & Okebukola, 1990; Roth & Roychoudhury, 1992; Willerman & Mac Harg, 1991). An examination of a small sampling of teaching methods and precollege science texts provides evidence of the increasingly important emphasis on concept maps in science teaching, particularly at the K–12 levels (see, for example, Aldridge et al., 1995; Carin, 1997; Collette and Chiappetta, 1994; Martin, 1997; Martin, Sexton, Wagner, & Gerlovich, 1997).

During the early stages of the development of concept mapping, a “variety of scoring keys” were tried, the “common form” providing points for correct linkages, number of levels of hierarchy, and number of crosslinks (Novak, Gowin, & Johansen, 1983, p. 627). Educators were encouraged to “experiment with their own scoring keys and refinements of scoring criteria” (Novak & Gowin, 1984, p. 108). Recent reports (Ruiz-Primo & Shavelson, 1996; Shavelson, Lang, & Lewin, 1993) indicate that a wide range of methods for constructing and scoring maps exists.

Questions about scoring concept maps—that is, the meaning of the scores—were raised at the time concept maps were first introduced, and many questions persist today (Lay-Dopyera & Beyerback, 1983; Novak et al., 1983; Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993; Wallace & Mintzes, 1990; White, 1987). The relatively low correlation of concept map scores with other measures of achievement (Novak et al., 1983) has been of particular concern. Such weak relationships have been interpreted as an indication that concept map scores obtained using the traditional scoring method measure abilities other than those characterized by “standardized achievement tests or conventional course performance measures” (Novak et al., p. 638). If concept maps are to be used with confidence as valid measures of student achievement in support of classroom instruction, it is essential that scoring methods be developed and validated that result in scores that reflect a stronger relationship between concept maps and student learning in science, scores that are reliable measures of intended learning outcomes (Ruiz-Primo, 1996; Shavelson et al., 1993).

More recently, Ruiz-Primo and Shavelson (1996) provided a comprehensive analysis of concept maps as assessment devices, raising a variety of very important theoretical and practical issues focusing on the reliability and validity of concept maps as assessments. Ruiz-Primo and Shavelson proposed a “framework for conceptualizing concept maps as a potential assessment tool in science” based upon the definition of a concept map as a “combination of a task, a response format and a scoring system” (p. 573). The important implication of this definition for the current study is that changing the scoring system or applying multiple scoring systems for a single concept map task and response format may result in “multiple” assessments examining different aspects of students’ science knowledge.

Ruiz-Primo and Shavelson (1996) referred to concept maps as “alternative assessments” (p. 569), which raises an important issue. With recent concerns about science achievement and the growing interest in authentic or alternative assessment, there seems to be an increasing tendency to consider traditional and alternative assessment as competing strategies for collecting information about what students know and can do, and that the latter is somehow superior to the former. As a result, the increasingly common practice of referring to concept maps as “alternative” assessments (Collins, 1993; Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993) perpetuates the perception that concept maps must necessarily measure or reflect more complex levels of thinking in much the same way that portfolios, performance or other hands-on assessment methods are designed to do. Identifying concept maps with alternative assessment methods reflects the assumption that concept maps cannot be used to assess knowledge and understanding (Bloom, 1964) of facts, terms, and concepts (Roid & Haladyna, 1982) or that using concept maps for such assessments would be inappropriate.

There are several reasons why it is more useful to consider traditional assessment and alternative assessment as complementary rather than competing strategies for collecting information about what students know and are able to do. First, the two types of assessments explore different domains of student knowledge, both of which are important. Knowledge of facts, terms, and concepts is an important part of students' science knowledge in its own right and is also the basis for the development of procedural knowledge (Marzano, 1992). While declarative knowledge (Marzano, 1992) may tend to be less valued, such knowledge is necessary for solving authentic science problems, and therefore is worthy of assessment. Furthermore, while the psychometric characteristics of traditional assessments of declarative level knowledge are well established, this is not the case for many of the alternative assessment methods. This fact suggests that some caution might be advised in the use and interpretation of alternative assessments, particularly given that many political and fiscal decisions are based upon the results of assessments of students in today's classrooms and that the use of concept maps in assessing knowledge of facts, terms, and concepts should not be devalued out of hand.

The purpose of the research reported in this study was to explore the use of concept maps in assessing students' declarative knowledge in the context of the science classroom. It is the contention of the authors that this level of knowledge traditionally assessed by multiple choice, matching, short answer, and other objective measures can be reliably and validly assessed using concept maps. Furthermore, doing so should serve to complement or supplement the use of concept maps to explore procedural aspects of students' science knowledge and skills as advanced by many researchers.

The establishment of concept maps as a vehicle for assessing both declarative and procedural aspects of science knowledge is significant for a number of reasons. First, it reifies the legitimacy of different domains of science learning and mitigates against the emphasis upon one aspect of learning over another. We would assert that doing so is especially important because it allows for a more balanced curriculum that includes knowledge of facts, terms, and concepts as well as the performance of more complex cognitive operations. The use of concept maps to assess both levels of knowledge may also provide answers to questions about the differences in map scores and how these differences may relate to the constructs they are purported to measure. Another implication is that concept maps might be used in monitoring learning outcomes as instructional emphases shift from more basic to higher levels of performance, rather than employing a variety of assessment procedures. Science educators might also gain some insight into how both declarative and procedural learning outcomes might be measured without usurping more instructional time to carry out multiple assessments. These last two suggestions raise the possibility of applying multiple scoring procedures to the same task and student response.

The objectives of this study were (a) to develop a method of scoring concept maps that would assess students' declarative knowledge relative to specific instructional objectives, (b) to establish the reliability and validity¹ of the resulting scores, and (c) to evaluate the relative merits of the scoring method developed in the study and other methods of scoring maps as measures of student classroom achievement.

At the outset, it is important to establish that some of the procedures employed in this research were designed primarily to support the research goals of the study. The development of curriculum referenced tables of specifications (Linn & Gronlund, 1995) and comparisons of concept maps with multiple choice tests, for example, were necessary to examine the content and concurrent validity of concept maps scored for declarative knowledge. These steps would not be required or effective if the method of scoring concept maps described in the study were to be used in the regular science classroom in support of instruction. Possible adaptations of the method would be needed and will be discussed. The remainder of this article is organized into

the following sections: a brief summary of pertinent research on the use of concept maps, methodology, results, discussion, and conclusions.

Background on the Use of Concept Mapping

Concept mapping has been defined as a “metalearning strategy” (Wandersee, 1990, p. 927), the development of which can be traced back to the well-known work of Ausubel, Novak, and Gowin. The research base on concept mapping shows that the use of concept maps is not limited to any particular group of learners. Children as young as primary grades have been found to be capable of developing and explaining concept maps (Novak & Gowin, 1984; Novak, 1990; White & Gunstone, 1992). A number of researchers have reported the successful development of concept maps by middle school-age children (Novak & Gowin, 1984; Novak et al., 1983; Symington & Novak, 1982; White & Gunstone, 1992; Willerman & Mac Harg, 1991). The works of Anderson and Huang (1989) and Novak et al. have also shown that students of varying ability can become good concept mappers.

Prior research also indicates that concept mapping can be easily and quickly taught to students and that once taught, the technique can be used with large groups with minimal assistance from teachers (Wallace & Mintzes, 1990; White, 1987). This latter feature is particularly important if assessment techniques such as concept mapping are going to be feasible in the typical classroom setting or in large-scale assessment projects (Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993).

As indicated previously, a variety of methods for scoring concept maps have been presented in the research literature (Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993). Some researchers (White and Gunstone, 1992) recommended that maps should not be scored, while at the other end of the spectrum, Novak and Gowin (1984) provided the basis for quite elaborate scoring systems. An important point to note is that it seems the criteria for selecting from among the many possibilities is “the purpose for which the scores are wanted” (White & Gunstone, 1992, p. 39).

While early research that found low correlations between maps and other measures of achievement created some concern about the validity of using concept maps for assessing achievement, more recent research (cited in Shavelson et al., 1993) reported correlations between concept map scores and measures of science achievement and aptitude “at or above 0.50” (p. 22). Of particular interest is another set of studies by Anderson and Huang (1989) in which a correlation between concept maps and corresponding short answer tests of .69 was found. Anderson and Huang also demonstrated that concept map scores correlated highly with standardized measures such as Otis-Lennon ($r = .74$) and Stanford Science Achievement ($r = .66$).

This disparity between correlations obtained in these more recent studies (Anderson & Huang, 1989; Shavelson et al., 1993) and those reported in earlier concept map research (Novak et al., 1983) support the contention that scores resulting from different scoring methods reflect different types of abilities. The scoring criteria employed in the studies by Anderson and Huang did not rely upon characteristics used in more traditional scoring methods such as hierarchy, crosslinks, or branching. Their scoring system focused upon the degree of accuracy of the relationship described in each proposition (pair of concepts), a method favored by Ruiz-Primo and Shavelson (1996). The standards of comparison for evaluating the achievement of students in the Anderson and Huang studies were the “right” answer, that is, an “expert map,” and one might assume scores obtained by this method reflected the quality of students’ understanding of the concepts.

The high correlations between concept maps and classroom tests and various standardized measures of achievement obtained in the studies of Anderson and Huang (1989) suggest that the method of scoring maps, particularly the criteria for allotting points, is a critical determinant of the strength of the relationship between map scores and scores on other assessments. It follows also that the purpose for which the map scores are to be used must be considered when selecting or developing the scoring method. These observations underscore Ruiz-Primo and Shavelson's (1996) contention that the relationships among the task, the response format and the scoring system are essential to the use and interpretation of concept maps as assessments.

Study Methodology

Overview

A method of scoring concept maps based upon a table of specifications for a unit of instruction was developed. Tables of specifications, or test blueprints, are two-way charts used to indicate the objectives that have been stressed and the content that has been emphasized during instruction (Linn & Gronlund, 1995). The number of test items falling into each cell should proportionally reflect the instructional emphasis placed on each category. By basing the scoring criteria on a table of specifications, the scoring method is more closely linked to specific curricular and instructional objectives than are more traditional concept map scoring methods that reflect the form and structure of the concept map itself (for example, see Novak & Gowin, 1984; Starr & Krajcik, 1990; Wallace & Mintzes, 1990).

Study Participants

The study took place in a large suburban middle school located near a medium-sized Southeastern city. The school is located in a district noted for excellence in academics and in sports and serves a diverse mix of socioeconomic groups. The student body of just over 1,600 was composed of seventh, eighth, and ninth graders.

One seventh-grade team of 113 students participated in the study. For science instruction, students were divided into five heterogeneously grouped classes. The group was 81% White, 15% African American, 2% Hispanic, and 2% Asian. These percentages reflected the approximate schoolwide racial makeup of the student body. About 60% of the students were male and 40% were female.

The first author was the students' regularly assigned life science teacher, on leave from a university position. Study procedures were carried out as part of the regular instructional program in life science.

Study Treatment Procedures

The study began about 5 weeks into the school year and spanned a 23-week period during which eight life science instructional units were completed. This period can be roughly divided into two parts: the concept map training phase and the concept map data collection phase. The training phase of the study was initiated after completion of the first three units and focused on teaching students how to construct concept maps. Three units were studied during this period: Unit 4 was The Cell, Unit 5 was Classification, and Unit 6 was Protozoans. Students had expe-

rience during this phase of the study in drawing concept maps during whole-class instruction, as a small-group activity, and as homework. Students also had their first experience in drawing concept maps as part of the summative evaluation process during this phase.

During the concept map data collection phase, emphasis shifted from the use of maps during instruction and teaching students how to draw maps to assessment using concept maps. During this period, five units of instruction were completed: Unit 7 was Fungi, Unit 8 was Plants, Unit 9 was Invertebrates, and Units 10 and 11 were Vertebrates. At the end of each unit, a test was given that consisted of multiple choice items, short answer questions, and one concept map item. In addition, a cumulative examination covering material from Units 5–8, was given between Units 8 and 9. The data collection ended following the Unit 11 test.

Procedures during the Concept Map Training Phase

Many science educators recognize the importance of aligning instruction and assessment, and Shavelson and Baxter (in Shavelson et al., 1993) suggested that such an alignment is developed if concept mapping is embedded within the curriculum. Thus, in the current study students were taught how to make concept maps as part of regular classroom instruction in a progressive manner over a period of nearly 7 weeks using techniques similar to those suggested by White and Gunstone (1992). It was also expected that students' learning of mapping skills would be facilitated by using concept maps as part of instruction (Anderson & Huang, 1989). Training students to use concept mapping over several weeks also allowed the teacher/researcher to reflect upon the progress being made at each step and to adjust instruction accordingly.

The teacher/researcher, who had some 13 years of teaching experience prior to the study, followed detailed lesson plans during the entire course of the study to help ensure consistency in instruction across all classes. These plans were reviewed by another member of the research team at the beginning of each unit. No other steps were taken to establish fidelity of instruction or to ensure that students were not exposed to any form of expert map during the data collection phase.

Instruction in concept mapping began with a homework assignment requiring each student to develop a list of the important words (concepts) from Unit 4, The Cell. The day preceding the unit test, these concepts were the topic of class discussion and review. During this session, the teacher/researcher and the students first selected a smaller number of concepts that the students agreed were the most important words. Students were aided by the teacher as they worked in small groups to make group maps. This activity was facilitated by writing the terms on small pieces of paper which students could move around on desktops (White & Gunstone, 1992). Group maps were then synthesized through class discussion to produce a class map that was drawn on the board by the teacher/researcher based upon students' suggestions.

Unit 5, Classification, provided the context for the development of a class map during whole-class instruction. Upon answering questions during review, students were prompted to contribute to the class map with a follow-up question such as, "Where and how would you put that on the map?" The teacher/researcher added students' responses to the class map, again drawing it on the board. Decisions by the students about exactly how new concepts would be added to the map were often preceded by intense discussion and debate.

During the sixth unit, Protozoans, students were again asked to develop lists of terms they thought were most important in the unit. The following day, a short list of 30 concepts was produced during class discussion. This final selection of concepts for the list was guided by the teacher/researcher based on knowledge of the table of specifications (Linn & Gronlund, 1994) for the upcoming test (Table 1). Students worked in cooperative groups the next day to draw

Table 1
Sample table of specifications (Unit 6)

Objectives content	Knows		Applies	Analyzes	Total	%
	Basic terms	Specific facts				
Protozoans	1, 4, 6	2, 9	16, 18, 19, 20	21, 22, 23	12	53
Sarcodines		7, 15, 17			3	13
Ciliates		3, 11			2	8
Flagellates		10, 13, 14			3	13
Sporozoans	5, 12	8			3	13
Total	5	11	4	3	23	100

maps of these concepts on large pieces of newsprint. Group maps were then used during whole-class review to develop the class map on the board in a manner similar to that used during the Unit 5 review. Subsequently, students were given the opportunity, announced in advance, to draw a map of the unit concepts as an extra credit question on the unit test. These maps provided the database for the subsequent development of an alternative strategy, or rubric, for scoring concept maps.

Procedures during Concept Map Data Collection Phase

The seventh unit, Fungi, could be characterized as a transitional unit. By the end of Unit 6, students had demonstrated their facility in constructing concept maps. As Unit 7 progressed, the teacher/researcher developed a concept map based on student suggestions during a series of brief summary lecture/discussion sessions, much as one would develop an outline. Students were encouraged to copy the growing map from the board as part of their note taking. It is important to note that this was the last time any form of group map was drawn. Near the end of the unit, the teacher/researcher defined a final list of 22 concepts, again based on the unit table of specifications. Students were given the list of concepts and asked to construct a concept map showing how the 22 concepts were related as one of the short discussion questions on the unit test.

During the final four units, 8–11, covering plants, invertebrates, and cold- and warm-blooded vertebrates, the teacher/researcher would make references to where concepts might be placed in a concept map during lectures, questioning, and class discussions. For example, when students asked questions about relationships among concepts, the teacher/researcher's responses were in terms of "If you were drawing a map, where do you think [the concept] would fit? At no time, however, did the teacher/researcher actually draw a concept map or have students draw concept maps during these four units, except for those maps drawn by students on the unit tests.

For each of these units, the selection of the concept list for the concept map test item was guided by the table of specifications for the unit test and the teacher's expert map for the unit. While the expert map included all major unit concepts, the number of concepts on the students' lists was typically between 20 and 30. Students were given the concept list in advance of each unit test. The final item on the unit test required students to draw a map correctly connecting as many of the concepts as possible. These maps were graded for purposes of regular class assessment based on the number of correct propositions, defined as a linkage between two terms or concepts via a directional arrow on which an appropriate label had been placed (Shavelson et al., 1993). Sample student maps are shown in Figures 1–3. For these

last four unit tests, students received extra credit for including additional concepts not on the unit list. During the course of the study, students demonstrated some opposition to or confusion about drawing arrows; subsequently, little emphasis was placed on the presence of an arrow-head on connecting lines.

It is important to note again that beginning with Unit 8, students never saw the teacher/researcher's expert map. Even during instruction, when the teacher/researcher asked students where a concept might fit if they were drawing a map, students in the class made these decisions without intervention or suggestion by the teacher. As a result, maps produced by the students on their tests, or during instruction, were completely of their own design and construction. This procedure was very important to the purposes of the study and addressed warnings regarding teaching to the test—that is, the teacher's presenting an expert map which students then memorize for regurgitation during assessment (Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993).

Other researchers have recommended that the number of concepts used in developing concept maps be limited (Anderson & Huang, 1989; Shavelson et al., 1993; White & Gunstone, 1992), particularly when introducing students to concept mapping. Based on observations of students' mapping during the concept map training phase, and given the age of the participants in this study, the list of concepts was kept to what was considered by the teacher/researcher to be a reasonable number, generally about 20 and no more than 30 terms.

Photocopies of all maps were made for use in the study prior to "grading," and original maps returned to students along with the rest of the unit test. Maps were identified only by assigned student numbers.

Designing Unit Multiple Choice Tests

In designing each unit test, the teacher/researcher, in consultation with one of the other research team members, first reviewed all pertinent information: instructional objectives, teacher/researcher's class notes, lesson plans, and study guides given to the students. Based on this information and the teacher's knowledge of what had actually transpired in class, a table of specifications was constructed. Using this table, the teacher/researcher wrote all test items and developed a list of concepts to be given to students in preparation for the concept map question on the test. The resulting unit tests covering Units 5–8 consisted of between 15 and 26 multiple choice items, three to five short discussion questions, and a concept mapping item. The 45-item cumulative exam on Units 5–8 was developed similarly.

An effort was made in designing these unit tests to maximize the number of multiple choice items which incorporated concepts from the unit concept list, while still building a test which validly reflected the table of specifications. The stem concept(s), correct answer, and wrong answers were all selected from the concept list whenever possible, thus providing a strong link between the multiple choice items and students' concept maps. As a result, each unit test included several multiple choice items that were composed of concepts that were also on the concept map list. The number of related multiple choice items ranged from 4 (18%) items on Unit 7 and 4 (27%) on Unit 11 to a high of 11 (50%) items on Unit 10. Answering a corresponding multiple choice item on a concept map required the use of concepts found in the multiple choice item. Since the numbers of concepts on the concept map lists were small, a limited number of multiple choice items could be addressed on the concept maps. On each unit test, two subgroups of multiple choice items were identified: map related, or those items built from several concepts on the concept list; and other, or those items that could not be linked as completely to the con-

cept list. The importance of the relationship between these two subgroups of multiple choice items and the concept maps will be explained more fully subsequently.

Other Data Collection

In addition to the unit tests, data on students' race and gender, scores on two standardized achievement tests, and final course grades were recorded. The standardized test data included students' sixth-grade scores on the statewide criterion-referenced Basic Skills Assessment Program (BSAP) tests in mathematics, reading and science and the results from the Stanford-8 Achievement Test (SAT), administered to all seventh graders in the state approximately 2 weeks after the completion of the study.

Affective data were collected by including one item related to the concept mapping experience on an open-ended course evaluation given to all students approximately 8 weeks after the end of the study. These evaluations, identifiable only by student numbers, were collected by a student in each class, placed in an envelope, and held by another teacher until after the school year ended and final grades were recorded.

Development of the Scoring Rubric

The principle underlying the development of the scoring rubrics in this study was that concept map scores should reflect the degree to which students mastered expected learning outcomes. As described earlier, concurrent with the development of the multiple choice items, the teacher/researcher's expert map was used to identify a list of concepts for students to use on the unit test mapping item. Therefore, a number of multiple choice items on each test, referred to earlier as map related, could also be answered by concept mapping. That is, the concept in the question stem and, if possible, all of the answer choices were on the list of concepts developed for the concept mapping test question. For example, the following question is from Unit 6, Protozoans: Which of the following is *not* a type of protozoan? (a) paramecium; (b) amoeba; (c) Euglena; (d) bacteria. The correct answer to this multiple choice question is "bacteria" because they are in the kingdom Monera, not the kingdom Protozoa. To demonstrate knowledge of the same elements of content on the expert map, the teacher/researcher, with knowledge of the instructional objectives and instructional emphasis, would link the term "bacteria" with "kingdom Monera," not with "kingdom Protozoa" or any protozoan such as Euglena, paramecia, and amoeba. The list of concepts given to the students for mapping included the relevant terms "kingdom," "Protozoa," "Monera," "Euglena," "paramecium," "amoeba," and "bacteria." Thus, there was the potential for students to answer the question inherent in the corresponding multiple choice item by correctly associating these terms on their concept maps.

To achieve the study objectives, three categories of scoring rubrics were used to rate students' concept maps, identified as "A," "B," and "C." The application of the three rubrics in scoring concept maps based on Unit 6 provides an example.

Rubric Category A Scoring Example. The most basic criterion in scoring concept maps obviously is whether the pertinent terms are on the map (Novak & Musonda, 1991). Category A was used to indicate whether stem and/or correct answer concepts were on the map. If students had all of the pertinent stem and correct answer concepts somewhere on their map, they received a score of 0—that is, nothing was missing. If any of the essential concepts were missing, the

Table 2
Category A example from Unit 6

Score	Qualification
-1	Does not have "bacteria" and/or "kingdom" "Protozoa" or "Monera" on map
0	Has key stem and answer terms on map

students' knowledge of the question could not be determined. Students then received a score of -1 in Category A, just as would occur if students failed to mark an answer on a multiple choice test. Category A of the scoring rubric for the Unit 6 question is found in Table 2.

Rubric Category B Scoring Example. Category B reflects the relationship between the concepts in the stem and the correct answer. If one or more of the concepts were missing from the map, a score of 0 was given. To obtain a score other than 0 in this category, students had to have all essential stem and answer concepts somewhere on their maps. A positive score (+1) was given if the stem concept was linked accurately to the correct answer, and a negative score (-1) recorded if a linking error was made.

The scoring of maps relative to Category B of the scoring rubric for the preceding sample question from Unit 6 is described in Table 3. For students to receive a score of +1 on this item, the terms "bacteria," "kingdom Protozoa," and "kingdom Monera" had to be on their maps. In addition, the term "bacteria" had to be connected to "kingdom Monera" and not to "kingdom Protozoa" (except to indicate that bacteria are a type of food for protozoans). If the concepts were present but incorrectly linked, students received a score of -1. If any of these key concepts were missing, a score of 0 was given. It is evident that Category A is mutually exclusive of Category B. That is, a score of 0 in Category B indicates that the student received a -1 in Category A.

Rubric Category C Scoring Example. Category C was designed to assess misinformation. Scores in this category convey information about connections between concepts in the stem and one or more of the distracters. If students incorrectly linked the stem concept to one or more of the distracters, a score of -1 was given in this category. A score of 0 was recorded if the distracters were correctly linked to the stem concepts or were not linked to the stem at all, as appropriate. Category C for the Unit 6 question is found in Table 4.

Table 3
Category B example from Unit 6

Score	Qualification
+1	Must correctly link "bacteria" with "kingdom Monera" (and not with "kingdom Protozoa")
-1	Does not correctly link "bacteria" with "kingdom Monera" or incorrectly links it with "Protozoa"
0	Does not have "bacteria" and/or "kingdom Monera" and/or "kingdom Protozoa" on map

Table 4
Category C example from Unit 6

Score	Qualification
-1	Links "paramecium," "Euglena," and/or "amoeba" incorrectly to <i>anything</i> in the "kingdom Monera"
0	Does not incorrectly link distractors with the "kingdom Monera"

As described earlier, during the development of the unit tests and concept map lists, a number of multiple choice items were written whose answers could be described on the unit concept map. Using the corresponding multiple choice item as a template, rubrics for scoring maps in Categories A, B, and C were generated for each of these items. Using rubrics developed for each unit concept map, two of the researchers separately scored the maps from the five unit tests (7–11) and the exam. Interrater agreement on the scoring was 98% (Rice, Ryan, & Samson, 1992). This result compares very favorably with those reported in the literature on scoring concept maps (Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993).

In scoring maps in this study, the development of rubrics was generally the most time-consuming part of the process. Time required for the entire scoring process, including writing the rubrics, was about the same as one would need for grading a couple of essay questions for the same size class, the difference being that several concept map items could be graded in a similar time frame.

Results

Scoring Unit Tests and Concept Maps for Assessment of Student Learning

The unit tests and Exam 1 were analyzed using a classical item analysis approach. Item difficulties (p values) and discriminations (point biserials) were calculated and Kuder–Richardson Formula 20 was determined for each as an estimate of test reliability. These reliability coefficients are reported in Table 5, along with basic descriptive statistics. This analysis indicated that these tests were reliable and thus might be used in establishing the validity and reliability of the method of scoring concept maps developed in the study.

As previously described, three concept map scores were developed. The related multiple choice questions were used as a referent in determining these three scores: A = concept map information missing/present; B = correct concept map information; and C = incorrect concept map information. For each unit concept map, A, B, and C scores were separately totaled and subsequently analyzed. Table 6 reports the means and standard deviations for each of the scoring categories (A, B, and C) by test and includes the number of multiple choice items that were related to the corresponding concept maps. The results show that the total A and total C scores are small. In this study, the B scores are of greatest interest as they represent the number of correct answers found on the maps—that is, the number of correct propositions.

To take greater advantage of the information provided by concept maps, an additional variable, percentage of map correct, was defined. This variable reflects the proportion of correct information (Category B) to the total information on each map—that is, correct plus incorrect

Table 5
Descriptive statistics for multiple choice tests

Test	Multiple choice test				
	Total no. of items	<i>M</i>	<i>SD</i>	% <i>M</i>	KR-20
7	22	16.2	3.1	.74	.63
8	21	14.2	3.5	.68	.69
9	26	17.4	5.0	.67	.83
E1	45	34.7	7.2	.77	.88
10	22	17.4	3.0	.79	.67
11	15	12.7	2.2	.85	.68
Multiple choice test: items "related" to CM					
Test	Total no. of items	% total multiple choice items	<i>M</i>	<i>SD</i>	
7	4	18	3.5	0.8	
8	7	33	4.8	1.3	
9	9	35	6.3	1.8	
E1	8	18	6.0	1.6	
10	11	50	9.2	1.8	
11	4	27	3.7	0.5	
Multiple choice test: items not "related" to CM "other"					
Test	Total no.	% total multiple	<i>M</i>	<i>SD</i>	
7	18	82	12.8	2.6	
8	14	67	9.6	2.5	
9	17	65	11.2	3.3	
E1	37	82	28.6	5.9	
10	11	50	8.0	1.8	
11	11	73	8.9	1.9	

(Category B + Category C). The means for the percentage of map correct ranged from 77% to 95%, with an average of 85% (Table 6).

Relationship between Multiple Choice and Concept Map Assessments of Student Learning

Students' concept map Category B scores, correct answers, for each test were correlated with scores on the corresponding multiple choice items on the test to provide a comparison of the concept map as a test with the multiple choice test. Map scores were correlated with the total scores on the multiple choice test (all items), with the scores identified as map-related (multiple choice questions which could also be answered on the maps), and with scores on other remaining items on the multiple choice test (those which could not be answered on the map).

The correlations between Category B map scores (correct answers) and the total scores on the multiple choice items on the five units tests and the exam ranged from .41 to .70, with an

Table 6
Concept map (CM) results

Test	Total no. of MC items	No. (%)MC related to CM	Total A score		Total B score		% map correct		Total C score	
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
7	22	4 (18%)	-0.1	0.5	3.6	1.1	89	0.19	-0.3	0.5
8	21	7 (33%)	-1.2	1.6	4.3	1.9	77	0.25	-1.3	1.6
9	26	9 (35%)	-1.7	2.5	6.5	3.4	85	0.20	-1.4	1.6
E1	45	8 (18%)	-1.4	2.1	3.6	1.9	85	0.23	-0.5	0.9
10	22	11 (50%)	-1.1	2.1	7.6	3.7	82	0.22	-1.4	1.6
11	15	4 (27%)	-0.4	0.7	3.1	1.8	95	0.16	-0.1	0.4
Avg.	26	8 (31%)	-0.9	1.5	4.7	2.3	85	0.20	-0.8	1.1

average of .55 (Table 7). The correlations when B scores were compared with the scores on the corresponding related multiple choice subset varied from .22 to .68. Because the multiple choice tests were much longer than the two subsets of items, the Spearman–Brown prophecy formula (Pedhazur & Schmelkin, 1991) was used to adjust these correlations for test length to achieve more applicable comparisons. Once adjusted, the range of correlations between Concept Map B scores and scores on the related multiple choice items was .51–.91. The initial correlations of B scores with the subsets of other multiple choice items varied from .36 to .66; when adjusted, these correlations ranged from .44 to .74.

The average correlations between the B map scores (correct answers) and total multiple choice scores ($r = .55$) and between B scores and scores on the other multiple choice items ($r = .56$) were modest and quite similar in value. In contrast, the average correlation between B map scores and scores on the related multiple choice items was much higher ($r = .75$) (Table 7). An examination of r values for each unit test and the exam reveals that these values were also higher for correlations of B scores with related multiple choice items. These results show that, as expected, students' map scores were more highly related to multiple choice items identified as map related than to either the total multiple choice scores or the subset of unrelated multiple choice items.

A t statistic for dependent correlations was calculated comparing the correlations of the B score with related multiple choice items and the B scores with other multiple choice items to determine the significance of these differences. The adjusted correlation values were used, and the t statistics are reported in Table 7. Significant differences were found in all instances except Tests 10 and 11. These results indicate that the concept map scores were significantly more highly related to scores on the related multiple choice items than to the other multiple choice items.

A second relationship was determined by comparing the percentage of map correct with the total score for the related multiple choice item subset. The resultant correlations, which were generally high, are found in the last column of Table 7. As expected, they were lower than the correlations for the B score (correct answers) with related multiple choice items, since the percentage of map correct, defined earlier, included both students' misinformation (C scores) and the correct answer (B scores) in the calculation.

A composite view of the relationship between the concept maps and the multiple choice tests can be gained by creating three variables: MCSUBTOT = total number correct on related multiple choice items for all tests; TOTBP = total B score—that is, total number of correct an-

Table 7
Concept Maps correlated with course-based multiple choice tests

Test	"B" score with total MC items		"B" score with related MC items		"B" score with "other" MC items		Differences	Correlation: related MC items to % correct
	No. items	r	No. items	r	No. items	r		
7	22	.46	4	.47	18	.40	6.12*	.53
8	21	.58	7	.59	14	.45	4.57*	.42
9	26	.70	9	.68	17	.66	3.19*	.50
E1	45	.68	8	.67	37	.63	7.35*	.33
10	22	.49	11	.47	11	.36	1.41	.48
11	15	.41	4	.22	11	.42	0.22	.02
Avg.	25	.55	7	.51	18	.48	3.81*	.38
Total	151	.88	43	.90	108	.85	5.84*	.65

*Significant at $p < .05$

swers for all maps; and TOTPRMP = total score of B/(B + C), percentage correct for all maps. When students' total scores for these three categories were compared, the correlation of MC-SUBTOT with TOTBP was .90 and with TOTPRMP, .65, as reported in the bottom row of Table 7. As with all of the other comparisons described previously, these results indicate that students were performing quite similarly on the concept map items and multiple choice items designed to measure similar content.

Students' map scores were also compared with scores on the state criterion-referenced BSAP tests and the nationally normed SAT for this group of students (Ryan, Rice, & Samson, 1993). Correlations between B scores and BSAP science, reading, and mathematics scores were .82, .84, and .82, respectively. Similarly, correlations of students' B scores with SAT reading and mathematics subtests were .85 and .87. Such correlations compare very favorably with similar analyses recently reported in the literature (Anderson & Huang, 1989). No SAT science scores were available.

It should be noted that an examination of concept map scores for possible gender or race differences revealed no significant differences (Ryan et al., 1993). The lack of difference in the performance of girls and boys and that of White and African American students may reflect equal levels of attainment for the two groups, the ability of the classroom teacher to implement instruction fairly and equitably, the nature of the concept-mapping process itself, or some combination of the three.

Scoring Students' Concept Maps: Examples from Unit 8

The Unit 8 test demonstrates more clearly the method of scoring developed in the study. Seven items on the multiple choice portion of the Unit 8 test were considered answerable by concept mapping. That is, the stem concept, the correct answer, and the wrong answers, "distracters," were all included on the concept list used by the students in drawing their concept maps. Scoring rubrics A, B, and C were developed for each of these seven items in the manner previously described and were used in scoring students' Unit 8 maps. The results of this scoring for the maps of three students, Student X, Student Y, and Student Z (Figures 1-3, respectively) are reported in Table 8. These particular student maps were not selected randomly, but were chosen because they typify three different levels of student performance on the mapping task.

The scoring rubrics for two of the seven items, Items 4 and 6, are described in Table 9 and will be used to provide a more detailed description of the application of the scoring method in the study. Frequent reference will be made subsequently to the three students' A, B, and C map scores for Items 4 and 6 found in Table 8.

On Item 4, all three student maps (Figures 1-3) contained the stem concepts and correct answer as indicated by the zero for the A score (Table 8). Only on Student Y's map (Figure 2)

Table 8
Concept map scores for seven Unit 8 items

Item	3			4			6			9			13			14			15				
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C		
Student X	0	+1	0	0	-1	-1	0	-1	0	0	-1	0	0	+1	0	0	+1	-1	-1	0	0	0	
Student Y	0	+1	0	0	+1	0	0	+1	0	0	+1	0	0	+1	0	0	+1	0	0	0	+1	0	0
Student Z	0	-1	-1	0	-1	0	-1	0	0	0	+1	0	0	-1	0	0	+1	0	0	-1	0	0	0

were the stem terms “monocot” and “dicot” correctly connected to the answer, “angiosperm” as indicated by the B score, +1. The B scores of -1 indicate that Student X and Student Z failed to correctly connect these concepts on their maps (Figures 1 and 3, respectively). In addition, one finds the stem concepts “monocot” and “dicot” to be incorrectly linked to one of the distracters, “gymnosperm,” on Student X’s map (Figure 1), an error indicated by a C score of -1 on Item 4. Student Z did not connect the stem concepts incorrectly to any of the distracters (Figure 3), as reflected in the C score of 0.

The A, B, and C rubrics for Item 6 were set up slightly differently than those for Item 4 (Table 9). Since this item was stated in the negative, the correct answer to the multiple choice item was in fact the one concept that would not be expected to be linked to the stem concept, “gymnosperm.” One would expect the other three answers, “cycad,” “conifer,” and “ginkgo,” to be linked to “gymnosperm” to indicate the correct relationships. Therefore, the A, B, and C rubrics for Item 6 (Table 9) were based on this relationship but still reflect the same patterns for development of rubrics described previously.

The A map scores of 0 for Item 6 indicate that the maps of Students X and Y contained the key stem and the three correct answer concepts, “cycad,” “conifer,” and “ginkgo” (Figures 1 and 2). The B scores on Item 6 for Students X and Y were -1 and $+1$, respectively (Table 8). An examination of Student X’s map (Figure 1) shows that “ginkgo,” “cycad,” and “conifer” were not correctly linked to “gymnosperm,” ($B = -1$), while on Student Y’s map (Figure 2) these terms were correctly linked ($B = +1$).

The A map score of -1 (Table 8) indicates that at least one of the four essential terms was missing from Student Z’s concept map, and an examination of Student Z’s map reveals that the concept “conifer” was the missing term (Figure 3). The B map score of 0 indicates similarly that Student Z’s map did not contain the essential pieces of information. All three of the students’ C scores were 0, indicating that no incorrect links were made between “spore producer” and “gymnosperm” (Table 8).

Table 10 provides a comparison of the three students’ scores on the multiple choice items and the corresponding map scores for all seven items from Unit 8 that were answerable using both assessment techniques. One may confirm that the B map scores (indicating correct answers) for Items 4 and 6 found in Table 8 correspond to the concept map results for these two items

Table 9

Sample concept map scoring rubrics for Items 4 and 6 on Unit 8

4. Dicots and monocots are the two groups that make up the _____.	
A. Concepts monocot, dicot or angiosperm not on map.	-1
Concepts monocot, dicot, and angiosperm on the map.	0
B. Dicot and monocot correctly linked to angiosperm.	$+1$
Dicot and monocot not correctly linked to angiosperm.	-1
Dicot, monocot, or angiosperm not on map.	0
C. Dicot and/or monocot linked incorrectly to gymnosperm, ferns, or conifers	-1
Dicot and/or monocot correctly linked or not linked at all to gymnosperm, ferns, or conifers.	0
6. Which of the following is not a gymnosperm group?	
A. Cycad, conifer, ginkgo, or gymnosperm not on map.	-1
Cycad, conifer, ginkgo, and gymnosperm are on map.	0
B. Cycad, conifer, and ginkgo are correctly linked to gymnosperm.	$+1$
Cycad, conifer, and ginkgo not correctly linked to gymnosperm.	-1
Cycad, conifer, ginkgo, or gymnosperm not on map.	0
C. Spore producer incorrectly linked to gymnosperm.	-1
Spore producer correctly linked or not linked at all to gymnosperm.	0

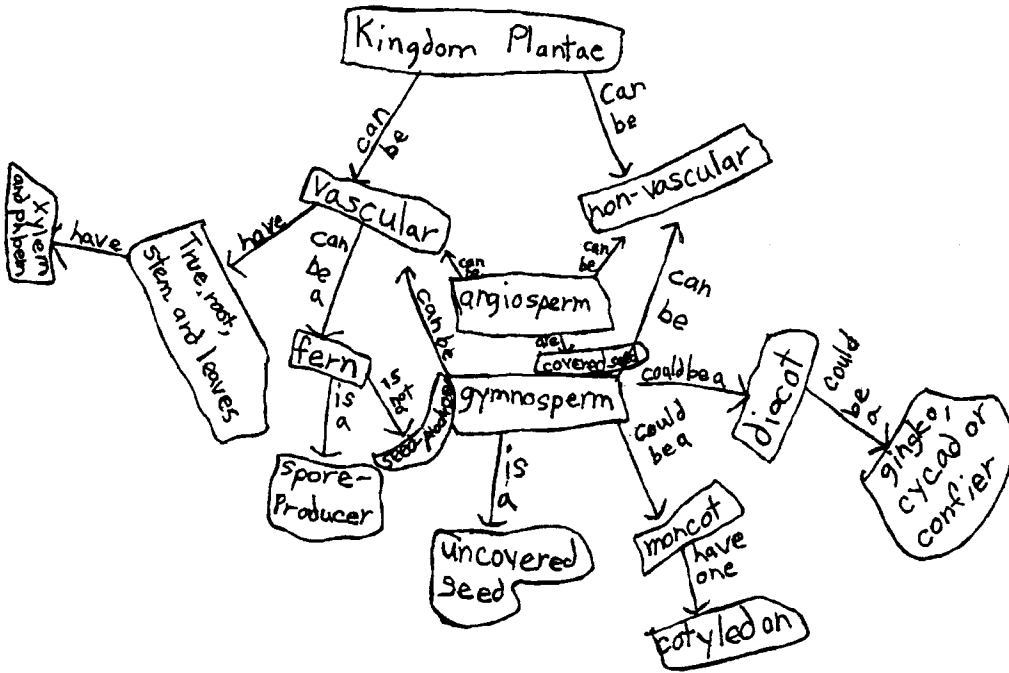


Figure 1. Student X's Unit 8 concept map.

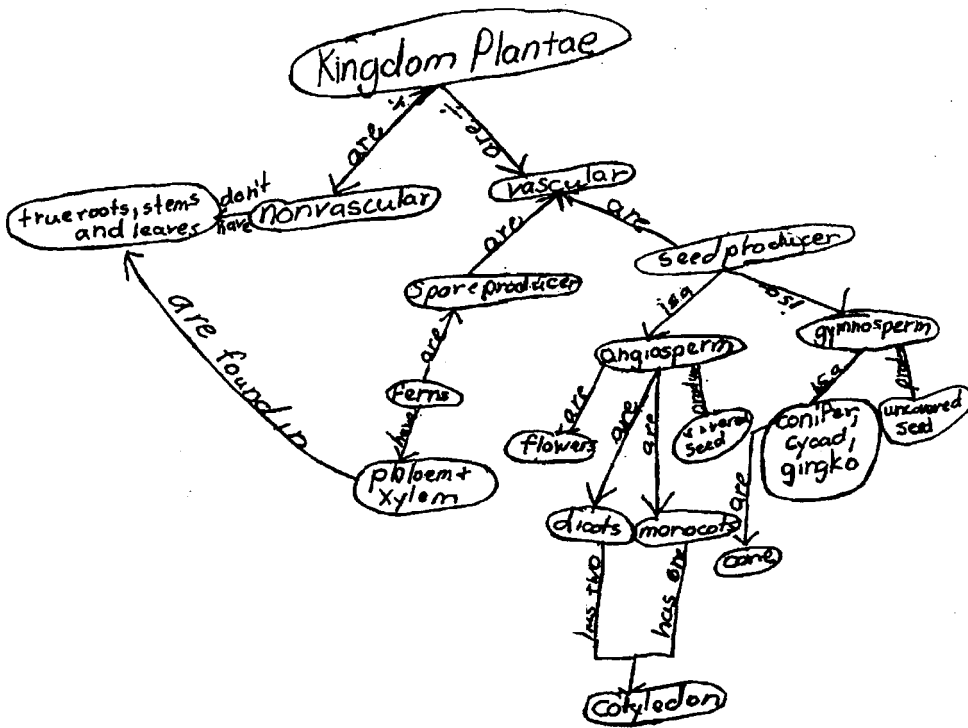


Figure 2. Student Y's Unit 8 concept map.

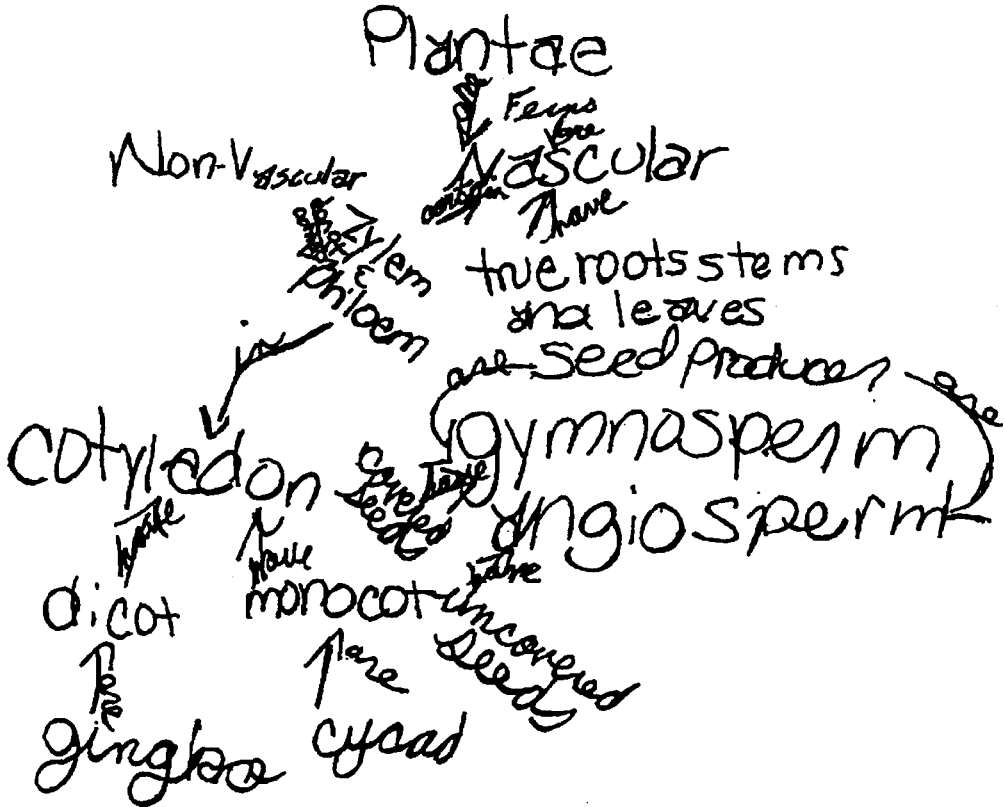


Figure 3. Student Z's Unit 8 concept map.

described in Table 10. The adjusted correlation between scores on these seven related multiple choice items and the concept map scores for Unit 8 for all students in the study was .81 (Table 7).

Affective Factors and the Scoring of Concept Maps for the Assessment of Student Learning

It has been suggested that some students do not develop concept mapping skills because of low “task motivation” (Anderson & Huang, 1989, p. 12). Therefore, it is important in estab-

Table 10
 Comparison of answers to related multiple choice items with concept map answers

Item	3		4		6		9		13		14		15	
	MC	CM	MC	CM	MC	CM	MC	CM	MC	CM	MC	CM	MC	CM
Student X	C	C	I	I	C	I	C	I	C	C	C	C	I	N
Student Y	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Student Z	C	I	I	I	I	N	C	C	I	I	I	C	I	N

Note. C = correct; I = incorrect; N = not on map.

Table 11
Summary of student reactions to concept mapping in science class (n = 95)

Student responses	<i>n</i>	% respondents
Helpful	39	41
and also boring	2	2
and also so hard	3	3
and also enjoyable	6	6
and also easy	1	1
Boring	18	16.8
Fun/enjoyable	20	21
Hated it	3	3
Difficult/complicated	10	10.5
Easy	4	4
Misc. positive	11	11.5
Misc. negative	13	13.6
Neutral	2	2
Overall positive	74 (62%)	
Overall negative	42 (35%)	

lishing the validity of any scoring method to determine whether attitude influences performance in concept mapping, particularly whether low concept map scores are the result of poor attitude or a lack of understanding of concepts. To explore possible effects of student attitudes on concept map scores, the following item was included on the open-ended course evaluation: "Drawing maps in science class was . . ." All responses to this question were categorized as positive or negative (Table 11) by the teacher/researcher and another member of the research team working together. Totals for each of these two categories were then determined for each student. Comments such as "boring" or "hated it" were labeled negative and those such as "helpful" or "easy" were labeled positive. Students sometimes expressed contrasting feelings such as, "It [concept mapping] was hard but it helped me learn better." In such cases, students were given credit for one positive comment and one negative comment.

Students were then divided into three groups: those who had more positive than negative comments; those who had equal numbers of positive and negative comments; and those who had more negative than positive comments. One-way analysis of variance revealed no significant differences in course grades, concept map scores, or multiple choice scores for the three groups.

Conclusions about the effect of attitude on students' concept mapping performance in this study must be viewed very conservatively because of the limited nature of the one-item instrument used in characterizing students' attitudes toward mapping. However, these results provide some support for the contention that attitude, particularly a negative one, was not a factor in concept map scores, and thus aid in the establishment of concept mapping as a valid measure of student learning in the context of this study. The integration of the process of teaching concept mapping into the existing curriculum, as described earlier, produced less disruption to normal class routines than would more invasive research procedures, and may account for the fact that concept mapping was generally perceived in a positive light by students in the study. This effort to downplay the research project was also a primary consideration in choosing not to use a more extensive, formal attitude survey focusing on concept maps.

Unintended Outcomes

Maps produced by students provided a rich source of information about their understanding or misunderstanding of concepts. These insights go beyond the right or wrong information produced by scoring multiple choice tests. While scoring maps, researchers easily developed a qualitative sense of where instruction had been unsuccessful. One such example was noted in scoring maps from Unit 10, Cold-Blooded Vertebrates. The process of examining a large number of maps during the scoring process revealed that students had apparently confused "salamander" with "lamprey." A large number of students interchanged the two terms in connecting them to their taxonomic classes, amphibian and jawless fishes. The teacher/researcher suggested that students were confused about the two animals, in part because, unlike other representative animals, specimens of these two organisms had not been available for examination by students. Without the hands-on observation, it would appear that students failed to grasp the difference between the two and confused the names that they apparently had simply memorized. While item analysis of multiple choice tests might also provide some indication of this confusion, teachers often do not have the time or inclination to carry out such analyses. Using concept maps for assessment, this failure of instruction was easily recognizable, and adjustments were made in subsequent instruction to remedy the problem.

Discussion and Conclusions

Certain questions about the reliability and validity of concept maps as an assessment remain unanswered even though they have been used extensively in science education over some 25 years. Ruiz-Primo and Shavelson (1996) recently provided a clear and comprehensive examination of these issues. The research presented here explored the use of concept maps in the assessment of seventh-grade students' declarative knowledge of concepts in life science and explored a number of the issues discussed by Ruiz-Primo and Shavelson.

The goal of the study was to develop not the definitive method for scoring concept maps for assessment in seventh-grade life science classes, but a method that assessed students' achievement of instructional objectives related to knowledge of terms and concepts. The relationship between instruction and concept map scores was established by developing a table of specifications to reflect instruction, constructing multiple choice test items based on the table of specifications, and then developing concept map scoring rubrics based upon the multiple choice items. Since tables of specifications by design reflect objectives covered during instruction (Linn & Gronlund, 1995), the resulting concept map scores were indicators of students' knowledge of content which had been emphasized during instruction.

This method of scoring maps represents a distinct departure from those that focus on criteria such as hierarchy and branching. Scoring of maps in the current study focused on the correctness of the propositions, a method that Ruiz-Primo and Shavelson (1996) suggested is preferable to those that "simply count the number of map components" (p. 595). Map scores obtained using this method were highly correlated with classroom tests as well as with state and national standardized tests. For example, when map scores were compared with scores on multiple choice test items on chapter and unit tests, correlations ranged from .51 to .86, and when compared with scores on state science, reading, and mathematics tests and with those on the reading and mathematics portions of the SAT, correlations were consistently in the .82-.87 range. The strength of the relationship between concept map scores and multiple choice scores in the study is not surprising since both reflect the same table of specifications, and it provides strong evidence for the content validity of the concept map scores. Further, the high correlations

between concept map scores and a number of standardized tests comprising the state-mandated testing program demonstrate that these map scores also have a high level of concurrent validity.

Generally, reliability data reported in studies using concept maps are in the form of interrater reliabilities (Ruiz-Primo & Shavelson, 1996). As previously mentioned, the two researchers scoring maps in the current study demonstrated an interrater agreement of 98%. This level of agreement using a scoring method focusing on the correctness of students' declarative knowledge rather than on counting parts of the maps such as branches and levels of hierarchy is surprisingly high (Ruiz-Primo & Shavelson, 1996) and provides strong evidence of the consistency of the scoring method.

The method of scoring maps which emerged from this study did not lend itself to the establishment of reliability through traditional methods such as alternate form, internal consistency, or test/retest. As indicated previously, the number of questions per test for which concept map scores were obtained was quite small, between 4 and 11 (Table 6). The concept map scores were, however, highly correlated with scores on corresponding multiple choice items on the five tests and the exam. These six multiple choice instruments had reliabilities (KR-20s) between .63 and .88 (Table 5). The strength of the relationship between scores obtained from concept maps and those for corresponding items from demonstrably reliable criterion multiple choice tests, combined with the relative stability of the correlations between these sets of scores (Table 7) and the very high interrater agreement, suggest that the method of scoring developed in this study yields reliable scores. Further research is needed to obtain more traditional measures of the reliability of scores obtained using this scoring method.

Studies to determine the reliability and validity of concept map scores must be implemented in the context of instruction if concept maps are to be used in assessing students' achievement of instructional objectives. Such an instruction based approach more appropriately reflects the interactive relationships among curriculum, instruction, and assessment. Indeed, much of the recent research and theoretical developments in the study of validity emphasize the centrality of examining the measurement context and the consequences of the measurement as part of validity studies (Linn, Baker, & Dunbar, 1991; Moss, 1995; Shepard, 1993). The evaluation of concept maps in a real-world context was seen as a vital dimension of the current study.

While the establishment of the reliability and validity of scores is a critical component in establishing the usefulness of any assessment, the degree to which methods of assessment can provide fair and just assessment of diverse student populations is also of vital importance (Malcom, in Collins, 1993; Shavelson et al., 1993). Before information obtained from any assessment procedure—in this case, concept maps—may be used for any purpose from the classroom to policy levels, the issue of equity must be addressed. It is therefore very important to note that an analysis for race and gender differences in concept map scores obtained in this study revealed no significant differences (Ryan et al., 1993). This finding should be viewed as a critical outcome by science educators, particularly in light of the increasing diversity of school populations in this country.

The scoring method developed in this study is fairly complex, very likely too much so to be used in the regular classroom setting. However, as indicated in the beginning of this article, certain procedures were implemented to achieve the research objectives of the study. For example, the development of A, B, and C rubrics served an important purpose in validating the method of scoring. Having achieved this objective, the next step would be to simplify the method to make it more practically useful in the regular classroom setting. Anderson and Huang (1989), whose scoring method was more similar to the one developed in this study than to the traditional scoring methods, used a relatively simple checklist approach in scoring student maps. Our experiences in this research project suggest that concept map scoring rubrics focusing on

instructional objectives can easily be designed without the intermediary use of multiple choice items. In fact, even a formal table of specifications may not be necessary. The key would be to relate concept map scores directly to outcomes specified in the objectives. With this approach, concept map scoring would still be directly related to curricular goals and instructional objectives, yet simple enough to be used in regular science classrooms.

The question of why various concept map scoring methods provide somewhat different scoring outcomes, yet purport to measure the same thing, was raised in two previous articles (Ruiz-Primo & Shavelson, 1996; Shavelson et al., 1993). As noted earlier, Novak et al. (1983), in comparing map scores to other measures of achievement, pointed out that “the [low] correlation coefficients suggest that concept mapping and Vee mapping tap abilities that are not well measured by standardized achievement tests or conventional course performance measures” (p. 638). Our research adds to the evidence supporting the observation that in fact, different methods of scoring maps measure different constructs or different aspects of the construct domain.

In comparing more traditional methods of scoring concept maps with the one developed in this study, it is important to note again the differences in criteria for assigning point values relative to information on the maps. As described previously, using the method in this study, points are awarded relative to the degree to which the map communicates understanding of concepts covered during instruction, as defined by tables of specifications. As such, the method is reminiscent of scoring maps relative to an expert teacher’s map that is assumed to include information that the teacher thought was important and that the teacher actually taught (Lomask et al., 1992). Using the traditional scoring methods based on the work of Novak and Gowin (1984), points are awarded based on a number of criteria including specific relationships, hierarchy, branching, crosslinks, and general to specific (for example, see Malone & Dekkers, 1984; Starr & Krajcik, 1990; Wallace & Mintzes, 1990).

As mentioned earlier, White and Gunstone (1992) suggested that the choice of a scoring method is directly related to the use for which the scores are intended. Traditional methods of scoring concept maps have been used in exploring complex constructs—for example, in exploring conceptual change or defining conceptual frameworks (Markham et al., 1994; Novak and Musonda, 1991; Powers and Wright, 1992; Roth and Roychoudhury, 1993; Wallace and Mintzes, 1990). On the other hand, the scoring method developed in this study is demonstrably effective in assessing students’ declarative knowledge relative to specific instructional objectives. We would argue that both levels of learning are necessary for scientifically literate citizens as well as scientists, and our findings suggest that both types of information could be obtained from a single response format—a concept map—by varying the method of scoring. This application of two scoring methods would be analogous to scoring essays using two analytic profiles (Lomask et al., 1992).

A significant practical implication of such an assessment approach is that it would support assessment-driven science curriculum and instruction that encourages students’ mastery of both declarative and procedural learning outcomes (Marzano, 1992). Currently, most science educators face the dilemma of restricting the science curriculum by the choice of the assessment approach. Use of selected response formats (multiple choice) tends to limit or exclude emphasis on higher-order learning targets, while focusing on constructed response formats (performance assessments) deemphasizes knowledge of important scientific facts, terms, and concepts. Educators interested in a curriculum that includes both declarative and procedural learning outcomes face a significant problem when selecting an assessment approach, in part because few schools are willing (or able) to give up instruction time to accommodate two different approaches to assessment. The ideological dichotomy of declarative versus procedural learning outcomes, mir-

rored in the juxtaposition of traditional versus alternative methods, is thereby perpetuated, all with the best of intentions.

As described earlier, this study suggests that such a dichotomy is unnecessary. The bulk of the research on concept mapping supports their value in assessing more complex learning outcomes, while the research described in the current study demonstrates clearly that concept maps can be scored for declarative knowledge of the sort commonly assessed with multiple choice tests. While more research needs to be conducted to further develop practical and efficient scoring rubrics to assess both types, the current study suggests that concept maps may become very useful as a single-format assessment technique with multiple scoring approaches.

The goal of this study was to determine whether science achievement can be measured when defined specifically as students' knowledge of information outlined in the instructional objectives. The study demonstrated that such assessment is possible, and at the same time, provided strong evidence that the match between method of scoring maps and the construct to be measured is critical, at least when measuring knowledge and comprehension level outcomes (Bloom, 1964). Whereas a number of studies have failed to produce concept map scores which were considered acceptable measures of classroom science achievement, in this study the match between method of scoring and the definition of achievement was the key to obtaining valid, useful, equitable, and reasonably reliable information about student learning in real-world seventh-grade life science classes.

Note

¹ The readers should bear in mind that the meaning of "validity" and the various approaches to developing measures of validity are currently undergoing a major reconsideration in the measurement/psychometric community. The work of Messick (1993, 1994, 1995), Linn et al. (1991), Moss (1995), and Shepard (1993) captured the major issues in these developments. It is important to note that the measurement community has not reached consensus on these new definitions and measures of validity; hence they are not incorporated into this article. The measurement terms used are those defined by Linn and Gronlund (1995).

In this study, "reliability" is used in two ways. Interrater reliability is used as an index of rater agreement in scoring the concept maps, and the internal consistency reliabilities (KR-20s) are reported for the multiple choice tests. Construct or content validity is addressed by referencing the multiple choice tests and concept maps to a common table of specifications that describes the intended content and cognitive levels for the students' learning outcomes. The intercorrelations among the various measures are reported as a form of concurrent validity.

References

Aldridge, B., Aiuto, R., Ballinger, J., Barefoot A., Crow, L., Feather, R.M., Kaskal, A., Kramer, C., Ortleb, E., Snyder, S., & Zitzewitz, P.W. (1995). *Science interactions*. New York: Glencoe-McGraw Hill.

Al-Kunifed, A., & Wandersee, J. (1990). One hundred references related to concept mapping. *Journal of Research in Science Teaching*, 27, 1069–1075.

Anderson, T., & Huang, S.-C.C. (1989). *On using concept maps to assess the comprehension effects of reading expository text* (Tech. Rep. No. 483). Cambridge, MA: Center for the Study of Reading. (ERIC Document Reproduction Service No. ED 310 368)

Bloom, B.S. (Ed.). (1964). *Taxonomy of educational objectives: The classification of educational goals: Handbook I: Cognitive domain*. New York: McKay.

Carin, A. (1997). *Teaching science through discovery* (8th ed.). New York: Merrill-Macmillan.

Collette, A.T., & Chiappetta, E.L. (1994). *Science instruction in the middle and secondary schools* (3rd ed.). New York: Merrill-Macmillan.

Collins, A. (1993). Issues in assessment: Purpose, alternative assessment and equity. *School of Education Review*, 5, 68–77. (ERIC Document Reproduction Service No. 370 771)

Jegede, O., Alaiyemola, F., & Okebukola, P. (1990). The effect of concept mapping on students' anxiety and achievement in biology. *Journal of Research in Science Teaching*, 27, 951–960.

Lay-Dopyera, M., & Beyerback, B. (1993, 11–15 April). *Concept mapping for individual assessment*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec. (ERIC Document Reproduction Service No. ED 229 399)

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex, performance-based assessment: Expectation and validation criteria. *Educational Researcher*, 20, 5–21.

Linn, R.L., & Gronlund, N.E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Merrill-Prentice Hall.

Lomask, M., Baron, J.B., Greig, J., & Harrison, C. (1992, 21–25 March). *ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science*. Symposium presented at the annual meeting of the National Association for Research in Science Teaching, Cambridge, MA.

Malone, J., & Dekkers, J. (1984). The concept map as an aid to instruction in science and mathematics. *School Science and Mathematics*, 84, 221–231.

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

Martin, D.J. (1997). *Elementary science methods: A constructivist approach*. Albany, NY: Delmar.

Martin, R., Sexton, C., Wagner, K., & Gerlovich, J. (1997). *Teaching science for all children*. Boston, MA: Allyn and Bacon.

Marzano, R.J. (1992). *A different kind of classroom*. Alexandria, VA: ASCD.

Messick, S. (1993). Validity. In R.L. Linn (Ed.) *Educational Measurement* (3rd Ed.). Phoenix, AZ: Oryx Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13–23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into meaning. *American Psychologist*, 50, 741–749.

Moss, P.A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, 14, 5–13.

Novak, J.D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, 27, 937–949.

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

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

Novak, J.D., & Musonda, D. (1991, Spring). A twelve-year longitudinal study of science concept learning. *American Educational Research Journal*, 28, pp. 117–153.

Novak, J., & Wandersee, J. (Eds.). (1990). Concept mapping [Special issue]. *Journal of Research in Science Teaching*, 27(10).

Pedhazur, E.U., & Schmelkin, L.P. (1991). *Measurement design and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Powers, D., & Wright, E. (1992, March). *The effects of hands-on science instruction on students' cognitive structures as measured by concept maps*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Cambridge, MA.

Rice, D., Ryan, J., & Samson, S. (1992, March). *Concept maps as an instructional and assessment tool in middle school science classes*. Summary of research presented at poster session at the annual meeting of the National Association for Research in Science Teaching, Cambridge, MA.

Roid, G.H., & Haladyna, T.M. (1982). *A technology for item writing*. Orlando, FL: Academic Press.

Roth, W.M., & Roychoudhury, A. (1992). The social construction of scientific concepts or the concept map as conscription device and tool for social thinking in high school science. *Science Education* 76, 531–557.

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, pp. 503–534.

Ruiz-Primo, M.A., & Shavelson, R. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33, pp. 569–600.

Ryan, J., Rice, D., & Samson, S. (1993, April). *Achievement and demographic correlates of students' performance on concept maps*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Atlanta, GA.

Shavelson, R., Lang, H., & Lewin, B. (1993). On concept maps as potential “authentic” assessments in science. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 367 691)

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*. Washington, DC: American Educational Research Association.

Starr, M.L., & Krajcik, J.S. (1990). Concept maps as a heuristic for science curriculum development: Toward improvement in process and product. *Journal of Research in Science Teaching*, 27, 987–1000.

Symington, D., & Novak, J. (1982). Teaching children how to learn. *Educational Magazine*, 39, 13–16.

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.

Wandersee, J. (1990). Concept mapping and the cartography of cognition. *Journal of Research in Science Teaching*, 27, pp. 923–936.

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

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

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