JOURNAL OF RESEARCH IN SCIENCE TEACHING VOL. 36, NO. 4, PP. 475–492 (1999)

Concept Map Assessment of Classroom Learning: Reliability, Validity, and Logistical Practicality

John R. McClure,[1] Brian Sonak,[2] Hoi K. Suen[2]

[1]*Department of Educational Psychology, Center for Excellence in Education, Northern Arizona University, P.O. Box 5774, Flagstaff, Arizona 86011-5774*

[2]*Department of Educational Psychology, Pennsylvania State University, State College, Pennsylvania 16802*

Received 13 January 1997; revised 17 November 1997; revised 23 March 1998; accepted 27 April 1998

Abstract: The psychometric characteristics and practicality of concept mapping as a technique for classroom assessment were evaluated. Subjects received 90 min of training in concept mapping techniques and were given a list of terms and asked to produce a concept map. The list of terms was from a course in which they were enrolled. The maps were scored by pairs of graduate students, each pair using one of six different scoring methods. The score reliability of the six scoring methods ranged from $r = .23$ to $r = .76$. The highest score reliability was found for the method based on the evaluation of separate propositions represented. Correlations of map scores with a measure of the concept maps' similarity to a master map provided evidence supporting the validity of five of the six scoring methods. The times required to provide training in concept mapping, produce concepts, and score concept maps were compatible with the adoption of concept mapping as classroom assessment technique. © 1999 John Wiley & Sons, Inc. J Res Sci Teach 36: 475–492, 1999

In December 1990, the *Journal of Research in Science Teaching* presented a special issue devoted to the topic of concept mapping. In the lead article of this issue, Novak (1990) outlined the potential uses of concept mapping for the improvement of learning and teaching in science classrooms. From Novak's remarks, we may organize the potential of concept mapping to improve science education into four categories: (a) as a learning strategy, (b) as an instructional strategy, (c) as a strategy for planning curriculum, and (d) as a means of assessing students' understanding of science concepts. This article is concerned with the last of these four categories.

In the same special issue of the *Journal of Research in Science Teaching,* Wallace and Mintzes (1990) presented evidence for the concurrent validity of concept map assessments of students' learning, and concluded that concept mapping tasks are a valuable tool for educational researchers. However, use of concept maps to assess students' understanding and learning was not a new idea. Novak (1990) described the development of concept maps in the late 1960s and

Correspondence to: J.R. McClure

 CCC 0022-4308/99/040475-18

early 1970s as part of a longitudinal research project that assessed changes in children's understanding of science concepts over a 12-year period. Surber and Smith (1981) investigated concept mapping tasks as a means of investigating students' misunderstandings.

Since 1990, concept mapping tasks have been used in a variety of ways to research topics in science education. Barenholz and Tamir, (1992) and Trowbridge and Wandersee (1994) used concept mapping tasks to assess the effects of science instruction. Hegarty-Hazel and Prosser (1991) used a concept mapping task to assess the relationship between conceptual understanding and the use of study strategies by science students.

In the introduction to their contribution to the 1990 special issue of this Journal on concept mapping, Wallace and Mintzes stated that ". . . as researchers we have an obligation to be cautious and circumspect about any new investigative technique" (Wallace & Mintzes, 1990, p. 1034). However, a review of the empirical evidence for the reliability and validity of concept map assessment tasks by Ruiz-Primo and Shavelson (1996) suggests that the time for caution and circumspection is not past.

Ruiz-Primo and Shavelson (1996) suggested several areas with regard to the use of concept map assessments that should be more thoroughly researched. The investigation of (a) the reliability of various concept mapping tasks, (b) the validity of inferences drawn from various concept mapping tasks, and (c) the practical application of concept mapping to classroom and large-scale assessment were among the areas recommended for further study. The primary purpose of this study was to contribute the investigation of these three areas.

Potential Advantages of Concept Map Assessment Tasks

Goldsmith and Johnson (1990) described the ideal assessment task as one that is objective and reliable, minimizes the influence of context on responses, and captures something of the structural nature of the subjects' knowledge. Intuitively, it seems that the first two characteristics of the ideal assessment are at odds with the last two. Traditional objective test formats (alternative response, short answer, etc.) may be objective and reliable; however, responding to these formats depends on cued recall or recognition processes. The result is that students' responses are strongly constrained by the context imposed by the test items. This limitation on students' responses may mask important individual differences in the organization of students' knowledge.

The traditional alternative to objective tests is tasks that are scored subjectively, such as essays, reports, presentations, or some type of project. The format of these tasks may reduce the constraint on students' responses, allowing something of the structure of their knowledge to be expressed. However, the quality of the responses may be influenced by a variety of factors that have nothing to do with the knowledge being assessed. Specifically, the students' level of skill at producing some artifact such as an essay, speech, or poster may unduly influence the assessment degrading the validity of subsequent decisions. In addition, the subjective nature of this type of assessment introduces the possibility of error due to the inconsistency of scores assigned by raters.

Assessments based on concept mapping tasks may strike a balance between desired objectivity and sensitivity to the structure of students' knowledge. Because of this potential balance, the addition of concept mapping tasks to teachers' repertoire may improve their classroom assessment in two ways. First, concept mapping tasks may be more useful for the diagnosis of students' misunderstandings owing to their sensitivity to (a) the structural nature of student knowledge, (b) intrusions or distortions in students' understanding of content, and (c) errors of omission (Surber, 1984). Second, in comparison to the production skills required by traditional

subjective assessment tasks, those required to produce a concept map are relatively simple, thereby representing less of a threat to the accurate assessment of students' knowledge.

The Nature of Concept Map Assessment Tasks

A concept map assessment is composed of two parts: (a) a concept mapping task, and (b) concept map evaluation. The concept mapping task is defined by those procedures that result in the construction of a concept map representing a student's knowledge. There is a variety of ways such maps may be produced. For instance, a map may be constructed by the evaluator based on student responses to an activity such as an interview or a word association task. Alternatively, students may be asked to construct a concept map themselves using pencil and paper. As this second type of task seems most practical for classroom applications, this type of concept mapping task was used in the assessments evaluated in this study.

A concept map evaluation involves an examination of the content and structure of a concept map. The nature of an evaluation may involve making qualitative and/or quantitative observations. The research reported here compares six different evaluation methods.

Purposes of the Current Research

The purposes of the research presented here are to investigate issues relevant to the adoption of concept map assessment tasks to the classroom. To this end, three questions will be addressed.

1. Is the score reliability of concept map assessments affected by the selection of scoring method?
2. Do concept map assessments provide valid information about student's knowledge?
3. Can concept map assessments be practically applied to classroom situations?

Reliability

In academic situations, the object of measure is an individual characteristic such as students' knowledge or skill in some domain. Often, this individual characteristic is assessed and represented by a single score. It is common to observe variations among the scores received by different individuals. Some of this variation is due to actual differences in the characteristic being measured; however, there may be other factors that contribute to the observed variation. These other factors are sources of error that prevent an accurate assessment of the object of measure. Reliability is an expression of the proportion of the variation among scores that are due to object of measure. As variation due to error goes to zero, the reliability of an assessment goes to 1.

Factors that may serve as sources of error in a concept map test include: (a) variations in students' concept mapping proficiency, (b) variations in the content knowledge (domain expertise) of those evaluating the concept maps, and (c) the consistency with which the concept maps are evaluated. This third factor depends in large part on the selection of a method by which concept maps are scored. The effect of the selection of a scoring method on the assessments score reliability is of primary concern in this study. The first two factors, subjects' mapping proficiency and raters' domain expertise, were assumed to contribute little to the variation of map scores. The concept map scoring method was varied, and it was anticipated that this would affect the consistency of the map evaluation.

The comparative score reliability of six different scoring methods was assessed by calculating a generalizability coefficient for each of the six methods. The six scoring methods eval-

uated were: (a) holistic, (b) holistic with master map, (c) relational, (d) relational with master map, (e) structural, and (f) structural with master map. Descriptions of these are provided in the Methods section.

Validity

"Validity is an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1988, p. 33). Classroom teachers' concerns with validity have to do with the quality of the inferences they make about students' grasp of content, as well as the quality of their pedagogical decisions and actions, including the assignment of grades.

The validity of decisions made using information from a concept map assessment is influenced by the nature of both the concept mapping task and the concept map evaluation. For optimal validity, the concept mapping task must result in an artifact (a concept map) that accurately reflects the content and organization of students' knowledge. If the procedures of a concept mapping task are overly complex, there is a chance that students' focus on the mapping procedures may degrade the quality of their representation.

The concept map evaluation will influence the validity of the assessment by affecting the quality of the information extracted from the concept maps. To some extent, this will be influenced by the nature of the concept mapping task. If the procedures for creating a map are not well specified, the variation in students' maps may make interpretation difficult. Combined with the arguments in the preceding paragraph, we are led to the conclusion that the concept mapping task should not be so complex as to distract the mappers, nor so simple as to sacrifice representational clarity.

The concept map evaluation procedures are also likely to affect the resultant validity of the assessment. The raters' prior knowledge of the content area, with concept maps in general and with evaluation procedures in particular, may affect the quality of the extracted information.

While there is potentially a wealth of information about students' knowledge that may be extracted from a concept map representation, for reasons of practicality, our analysis will focus on the validity of a score derived by each of the six specific scoring techniques. Evidence of the contribution of a score's contribution to a valid decision may be gathered in a variety of ways. One technique is to compare the scores from one assessment task to scores acquired from another form of assessment purported to measure the same characteristic. This is done by calculating a correlation for the two scores. Evidence gathered in this way is typically referred to as supporting concurrent validity of a measure.

In the current study, the analysis of validity involves the correlation of map scores with a measure of similarity between each map and a master map. The similarity of subjects' concept maps to the master map was assessed using a set theoretic method described by Goldsmith and Davenport (1989). To calculate the similarity between two maps using this method, the first step is to establish the set of related concepts for each neighborhood; a neighborhood is defined for each concept represented on the maps. Then, for each neighborhood, a similarity is calculated by dividing intersection of the neighborhood sets for the two maps by their union. Finally, the mean of the neighborhood similarities is calculated. The range of this mean will be between zero and one, and is the measure of the similarity between the two maps.

For example, referring to the maps represented in Figures 1 and 2, the neighborhood for the concept of "learning" for Map A of Figure 1 includes the concepts "knowledge," "expository," and "discovery." For the master map represented in Figure 2, the same neighborhood includes these same three concepts, as well as the concepts "transfer," "practice," "organization," and "elaboration." The intersection of these two neighborhood sets includes the concepts common to both

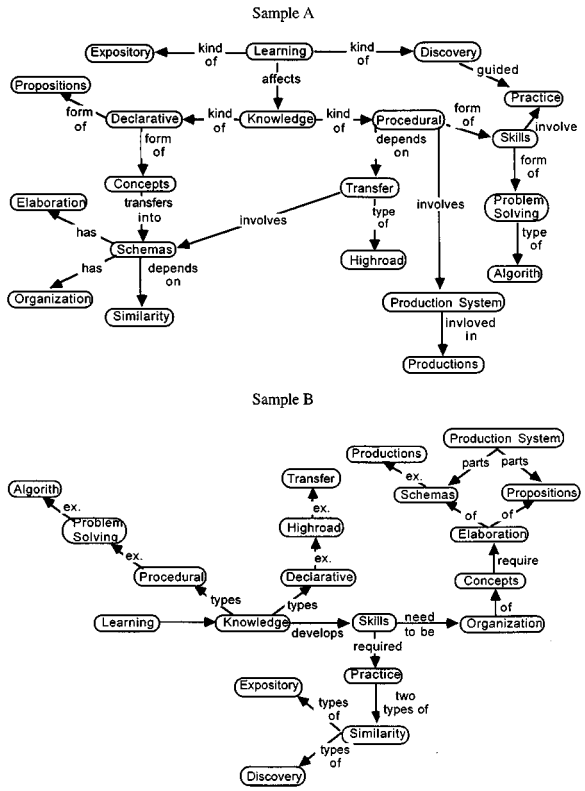

Figure 1. Sample concept maps.

maps. In this case, the intersection is 3. The union of the two neighborhoods is found by adding the number of concepts in the neighborhood for each map and subtracting the number of concepts they hold in common. In this case, the union is 6. The neighborhood similarity is then 3 divided by 6, to obtain 0.5. This procedure is repeated for each of the 20 concepts the maps have in common. This results in 20 neighborhood similarities that are then averaged to find the map similarity. In this example, the map similarity, comparing Sample A with the master map, is 0.3679.

This method of calculating map similarity is possible because all maps were constructed using the same set of concepts. The map similarity measure has been demonstrated to change in a predictable way with instruction and to correlate with traditional measures of classroom learning (Goldsmith & Johnson, 1989). This suggests that this measure is a valid indication of the quality of students' knowledge. Given the validity of the map similarity measure, correlations between concept map scores and the map similarity measures will be taken as evidence of the concurrent validity of the concept map scores produced by the various scoring methods.

Logistic Issues

While there are some potential advantages to assessing classroom learning through concept map tests, the degree of advantage must justify the allocation of resources. Teaching is a complex act requiring the simultaneous allocation of resources in the pursuit of multiple goals. Teachers must consider how the implementation of any new technique will affect the accomplishment of their other goals.

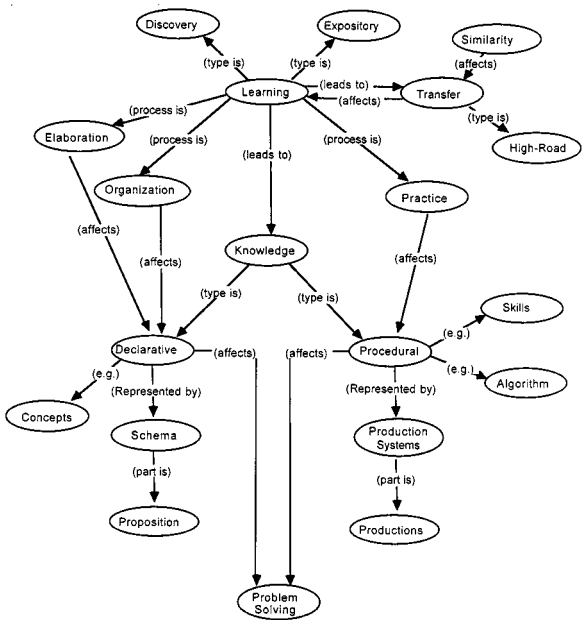

Figure 2. Master map.

The use of concept mapping tasks for classroom assessment will affect teachers in three ways. First, time must be allowed to train students in a concept mapping technique. Students must develop a level of proficiency necessary to produce reliable results (Surber & Smith, 1981). Second, teachers must consider how the time required to produce concept maps compares with traditional assessment tasks. Finally, teachers must consider the time required to score, or otherwise evaluate the concept maps produced by students. This study will attempt to assess the cost, in terms of time, of the adoption of concept map tests to the classroom.

Method

Subjects

Subjects for this study were recruited from two groups of students enrolled in a college of education at a large eastern university: (a) undergraduate education majors, and (b) graduate students enrolled in an educational psychology program. Sixty-three undergraduates volunteered to participate in this study. These volunteers were recruited from students enrolled in an introductory educational psychology course. These undergraduate students received a small amount of extra credit for their participation in the research. The undergraduate students participated as concept mappers, producing concept maps composed of concepts that represented content from the educational psychology course in which they were enrolled.

Twelve graduate students were recruited to serve as concept map scorers. These graduate students were paid a small fee for their participation in the study. All graduate students recruited for this portion of the study were enrolled in the educational psychology program and had earned an "A" in a graduate-level course in the cognitive psychology of learning. Because of the similarity of their education and training, the domain expertise of the raters was considered

to be equivalent. All raters received only limited training in concept mapping or scoring. This is consistent with the expectation that the experience of most classroom teachers with regard to concept map assessment techniques would be limited. The average classroom teacher might properly be considered a content area expert, with limited experience with concept map scoring techniques.

Questionnaires

Questionnaires were administered to both undergraduate and graduate participants to assess their familiarity with concept mapping. The purpose of the questionnaire administered to the concept mappers was to assess their familiarity with concept mapping techniques. The seven items comprising this questionnaire are presented in Table 1. This questionnaire was administered before their training in the concept mapping technique.

The questionnaire administered to the raters was composed of two parts. The first part of the questionnaire was to assess their familiarity with concept mapping techniques. The second was to assess their familiarity with the concepts used in the concept mapping task. The first portion of this questionnaire was similar to that administered to the concept mappers, except that Item 2 was modified to read, "I use concept mapping as a study strategy," and Item 3 was modified to read, "I use concept mapping as an instructional strategy."

The second portion of the questionnaire completed by the raters addressed their perceived familiarity with concepts used to create the concept maps. In this part of the questionnaire, the raters were to indicate if they were (a) not familiar, (b) somewhat familiar, (c) familiar, or (d) very familiar with each of the terms used in the production of the concept maps.

Training in the Concept Mapping Task

The training in concept mapping included a presentation of a concept mapping technique, followed by three guided practice sessions. The mapping technique presented was a modification of the networking technique (Holly & Dansereau, 1984). In this technique, words or phrases are connected by labeled arrows. In the original networking scheme, arrows were to be labeled with letters that represented a limited set of relational categories. However, as presented to the participants of this study, arrows might be labeled with any word or phrase to identify the relationship between two concepts.

Table 1
Survey administered to prospective concept mappers

For each of the items below, indicate whether or not the statement is applicable to you by circling either "yes" or "no."		
1. I never use concept mapping.	Yes	No
2. I use concept mapping as a study strategy for EDPSY 14.	Yes	No
3. I use concept mapping as a study strategy for most of my classes.	Yes	No
4. I use concept maps to prepare for exams and quizzes.	Yes	No
5. I use concept maps to help me understand what I read.	Yes	No
6. I use concept maps to take notes in class.	Yes	No
7. I use concept maps to understand or solve problems.	Yes	No

In the guided practice portion of the training, the participants were provided with a brief text passage and a list of concepts (words) from the passage. Participants were allowed 20 min in which to generate concept maps from the text passage using the listed concepts. While the participants worked on their concept maps, a researcher was available to offer advice and answer questions about the mapping technique. This procedure was repeated for three separate practice exercises. After each practice exercise, students were shown a map created by the researchers and encouraged to compare it to their map and ask questions.

To allow students to concentrate on the mapping technique, the three text passages were selected so as to present material that would be generally familiar to the mappers. The passages presented content common to high school courses in (a) American history: George Washington and the Continental Army; (b) biology: an introduction to bacteria; and (c) physical science: phases of matter. The complexity of the mapping tasks was manipulated across practice sessions by increasing the number of concepts listed for each session: 8 for first exercise, 10 for the second, and 15 for the third.

Map Production

Immediately after completion of the training, the concept mappers were given a set of 20 concepts and asked to create a concept map. These concepts were taken from an instructional unit on cognitive theories, completed as part of the educational psychology course in which the concept mappers were enrolled. Although told they had 40 min to complete this map, concept mappers who had not completed their map in the 40 min were allowed to finish. The result of this portion of the study was the generation of a set of 63 concept maps. Figure 1 provides an

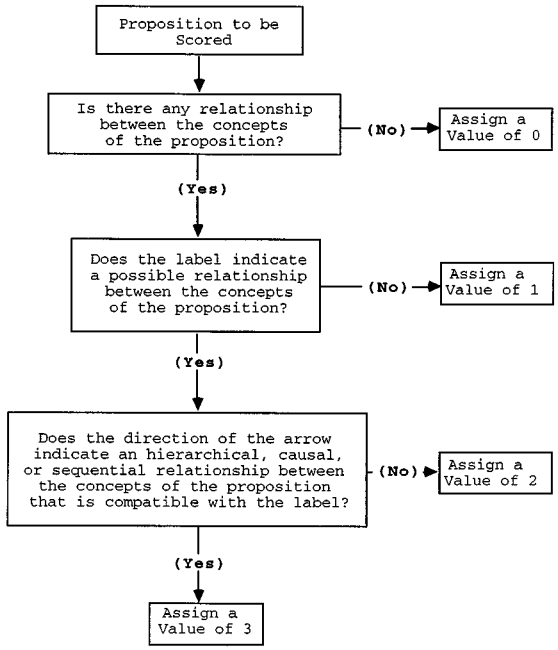

Figure 3. Protocol for the Relational Scoring Method.

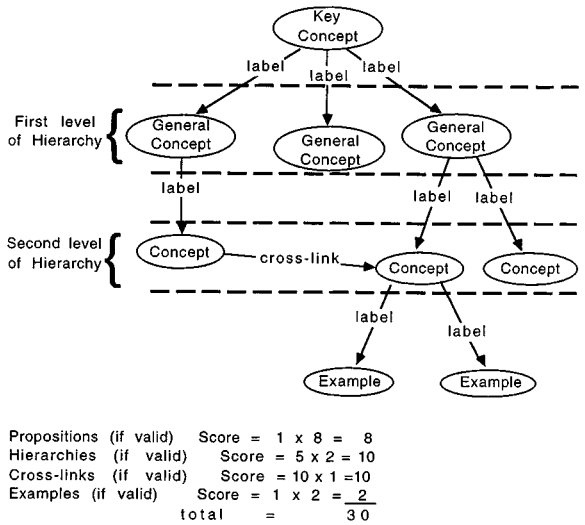

Figure 4. Instructions for the Structural Scoring Method.

example of two of the concept maps produced in this portion of the study. In addition to the concept maps, the time required for each student to complete their concept map was recorded.

Concept Map Scoring

The entire set of 63 concept maps was scored by six pairs of independent raters. Each pair used one of six scoring methods. The six scoring methods were: (a) holistic, (b) holistic with master map, (c) relational, (d) relational with master map, (e) structural, and (f) structural with master map.

Raters assigned the holistic scoring method were instructed to examine each concept map and judge the mapper's overall understanding of the concepts represented by the map. Based on this judgment, each map was to be assigned a score on a scale from 1 to 10. For example, referring to the maps in Figure 1, the two raters assigned the holistic scoring method assigned Map A scores of 8 and 10, and Map B, scores of 6 and 4.

The relational scoring method was adapted from a technique developed by McClure and Bell (1990). In this technique, raters scored individual maps by evaluating the separate propositions identified on the map. A proposition was defined as two concepts connected by a labeled arrow indicating the relationship between the concepts. Each proposition was to be scored from zero to three in accordance with a scoring protocol that considered the correctness of the proposition. The final score for the map was found by summing the scores of all the separate propositions. The protocol used by scorers assigned this condition is shown in Figure 3. Using this scoring protocol, one of the raters identified 4 two-point propositions and 17 three-point propositions in Map A of Figure 1, for a total map score of 59. The same rater found 2 one-point propositions and 12 three-point propositions in Map B of Figure 2, for a total map score of 38.

The structural scoring method was adapted from a method described by Novak and Gowin (1984). This method, in addition to awarding points for identifying correct propositions, also con-

sidered the presence of higher level structures within the concept maps. Points were awarded based on the number of hierarchical levels and crosslinks identified on the maps. Hierarchies are defined as branching structures that show superordinate–subordinate categorical relationships among concepts. Crosslinks are relationships identified between concepts located in different hierarchical branches. Figure 4 is a copy of the scoring instructions provided to raters assigned the structural scoring method. Using these instructions, one of the raters assigned this scoring method identified: 17 valid propositions, 3 levels of hierarchy, 2 crosslinks, and 2 valid examples, for a total score of 36 for Map A from Figure 1. The same rater identified 2 valid propositions, 1 hierarchical level, and 2 valid examples for Map B of Figure 1, for a total score of 14.

The three remaining scoring methods, holistic with master map, relational with master map, and structural with master map, were modifications of the three methods previously described: holistic, relational, and structural, respectively. The scoring methods were essentially identical; however, for these methods, a master map was provided to be used as a guide in scoring. The master map was a concept map constructed by the professor instructing the educational psychology course from which the mappers were recruited. Figure 2 is a copy of the master map used by the raters assigned these three scoring methods.

Each of the six scoring methods was randomly assigned to two raters. Each rater was given a packet of materials, that included (a) an introduction to concept mapping; (b) instructions for either a holistic, relational, or structural scoring method; (c) a questionnaire assessing raters experience with concept mapping and perceived familiarity with the concepts used to produce the concept maps; and (d) a set of 63 concept maps. In addition, raters assigned conditions involving a master map received a master map. This map was to be used as a guide in the application of the assigned scoring method. The written introduction to concept mapping and the instructions for a particular scoring method constituted the only training provided to the raters.

The raters were to read the instructions, complete the questionnaires, and then rate each map in the packet. They were to work at their own pace and at their convenience. In addition to recording the score for each map, the raters were asked to keep track of the time they spent on the rating task. Specifically, they were to record the total time to rate all 63 maps.

Results

The data collected and analyzed included: (a) scores for concept maps rated by each of the six scoring methods, (b) times required to generate and score concept maps, (c) the similarity of rated concept maps to the master map, and (d) responses to the questionnaires completed by concept mappers and the raters.

Questionnaire Responses

The concept mappers' responses to the questionnaire were used to determine whether there were differences in concept mapping experience that may have influenced the quality of the maps produced. To investigate this possibility, the relationship between the questionnaire responses and the quality of concept maps produced was checked. The first step in this analysis was to convert the concept mappers' responses to a single mapping familiarity score. The first item on the questionnaire was scored as 1 for a negative response; all other items were scored 1 for positive responses. The possible range for these scores was 0–7. The mean mapping familiarity score for the concept mappers was 1.59, with a standard deviation of 1.65. These scores were then correlated with the mean map scores for each map scoring method. Table 2 provides

Table 2
Correlations between concept mappers' mapping familiarity scores and the mean map scores for each scoring method

	Scoring method					
	Holistic scoring		Relational scoring		Structural scoring	
	Without master map	With master map	Without master map	With master map	Without master map	With master map
Correlation of map scores with mapping familiarity scores	.000	.004	−.072	−.139	−.006	.138

a summary of the correlations with mean map score for each scoring method. None of these correlations was statistically significant.

The purpose of the questionnaire completed by the raters was to determine whether there were important differences among the raters assigned different scoring methods in either concept mapping experience or domain expertise. The first portion of the questionnaire addressed the issue of differences in familiarity with concept mapping. The raters' responses were converted to a single concept map familiarity score using the same technique applied to the responses of the concept mappers. Table 3 provides a summary of the mapping familiarity scores for each rater assigned each scoring condition.

The second portion of the raters' questionnaire collected information about the raters' perceived familiarity with the terms used in the concept maps. The purpose was to determine whether there were differences in domain expertise among the raters assigned to the various scoring conditions. To analyze these data, the raters' responses were converted to a composite domain expertise score. This was done by assigning a score to each response as follows: "not familiar" = 0; "somewhat familiar" = 1; "familiar" = 2; "very familiar" = 3. Table 4 presents the mean concept familiarity scores for the raters assigned to each scoring condition.

Table 3
Summary of concept map familiarity scores for each rater and scoring method

	Scoring method					
	Holistic scoring		Relational scoring		Structural scoring	
	Without master map	With master map	Without master map	With master map	Without master map	With master map
1st rater	1	5	4	2	3	3
2nd rater	0	3	4	5	2	2
M	0.5	5.0	5.0	4.5	3.5	3.5

Note. There were two raters for each scoring method. The concept map familiarity score presented in the cells represents a composite scores derived from the raters response to questionnaire addressing the raters' familiarity with concept mapping. The minimum possible composite score was 0; the maximum possible composite score was 7.

Table 4
Mean domain expertise scores for raters assigned different scoring methods

	Scoring method					
	Holistic scoring		Relational scoring		Structural scoring	
	Without master map	With master map	Without master map	With master map	Without master map	With master map
1st rater	48	54	40	60	32	41
2nd rater	31	50	53	52	49	48
M	39.5	52.0	46.5	56.0	40.5	44.5

Note. There were two raters for each scoring method. The concept familiarity scores presented in the cells represent a composite scores derived from the raters response to questionnaire addressing the raters' familiarity with the concepts used in the concept mapping task. The minimum possible composite score was 0; the maximum possible composite score was 60.

Reliability of Concept Map Evaluation Methods

Scores from the six different scoring methods were submitted to a generalizability analysis. The data were analyzed through the GENOVA program (Brennan, 1983). Equation 1 describes the variance model used in the analyses:

$$\sigma_x 2 = \sigma_s 2 + \sigma_r 2 + \sigma_{sr} 2$$

where $\sigma_x 2$ is the total variance, $\sigma_s 2$ is the variance associated with the individual students, $\sigma_r 2$ is the variance associated with individual raters, and $\sigma_{sr} 2$ is the variance associated with the interaction between students and raters. The generalizability coefficients were estimated for each scoring method using Equation 2:

$$\text{g–coefficient} = \frac{\sigma_{s^2}}{\sigma_{s^2} + \sigma_{sr^2}} \tag{2}$$

The g-coefficients are an estimate of the score reliability of scores assuming a single rater. Table 5 shows the resulting variance component estimates and g-coefficient for each of the six scoring methods. As can be seen, the score reliabilities estimated by the g-coefficients ranged from a low of .23 for structural scoring with a master map to a high of .76 for relational scoring with a master map.

Validity of Concept Map Evaluation Methods

For each map, a map similarity was calculated by comparing the maps produced by the mappers to the master using the set theoretic method. The Pearson product moment correlation (*r*) was calculated comparing the similarity scores with the map scores generated by each of the 12 raters and with the means of the map scores for each scoring method. Table 6 provides a summary of these correlation. The mean correlations between map scores and map similarities ranged

Table 5
Variance components and g-coefficients for the six scoring methods

Source of variance component	Holistic without master map	Holistic with master map	Relational without master map	Relational with master map	Structural without master map	Structural with master map
Student	3.93	.73	68.45	80.83	37.87	19.51
Rater	0.24	.71	76.42	38.81	15.91	43.86
Student by rater	1.91	1.28	64.01	26.01	54.73	65.26
g-coefficient	0.67	.36	0.51	.76	0.41	.23

from a low of $r = .193$ for scores resulting from the structural with master map method to a high of $r = .608$ for scores resulting from the relational with master map method. The average correlations between map scores and similarity for all methods were statistically significant ($p < .01$), with the exception of the scores derived from the method, structural with master map. Excluding the comparisons with the correlation for the structural with master map method, there were no statistically significant differences among the mean correlation coefficients.

Time Requirements of the Concept Mapping Task and Evaluations

Time Required to Construct Concept Maps. Because of errors in data collection, completion time data were available for only 60 maps. For the students constructing the concept maps, the completion times ranged from a minimum of 16 min to a maximum of 51 min. The average time required for a student to complete a concept map in this study was 29 min ($n = 60$, standard deviation = 8.6 min). The mode and median were both 28 min.

Time Required to Score Concept Maps. The raters were asked to record the time required to score the entire set of 63 concept maps. For each of the six methods, a per-map time was estimated by dividing these times by 63. Table 7 provides a summary of the average times to score a set of maps and a per-map scoring time for each by each scoring method. The per map scoring times ranged from 1.3 min for maps scored using the holistic method to 5.2 min for maps scored using the structural scoring method.

Table 6
Correlation of concept map scores with neighborhood similarity

Rater	Holistic	Holistic w/master map	Relational	Relational w/master map	Structural	Structural w/master map
1st rater	.464**	.335**	.365**	.570**	.467**	.232
2nd rater	.310*	.561**	.392**	.645**	.279*	.153
M	.387**	.448**	.379**	.608**	.373**	.193

Note. Pearson product moment correlations calculated; $n = 63$ are statistically significant as indicated by $*p < .05$, $**p < .01$.

Table 7
Average time required for raters to score concept maps (n = 2)

Scoring method	Time (min) to score set of 63	Per-map scoring time (min)
Holistic	87	1.4
Holistic with master map	150	2.4
Relational	169	2.7
Relational with master map	230	3.7
Structural	327	5.2
Structural with master map	175	2.8

Note. Time to score a set of 63 maps are rounded to the nearest minute. Times to score a simple map are rounded to the nearest 0.1 min.

Discussion

A primary purpose of this study was to assess the relative score reliability and validity of six different scoring methods. An additional objective of this study was to evaluate the time requirements for concept map tests, taking into account (a) the time required to train students to create concept maps, (b) the time required for students to create concept maps, and (c) the time required to score concept maps.

Effects of Scoring Method on Reliability of Concept Map Assessments

Estimates of scoring method reliabilities were made by calculating the g-coefficient for each of the six scoring methods. Table 5 provides a summary of these g-coefficients. The reliabilities for the different scoring methods ranged from a low of .23 to a high of .76. These data suggest that the selection of a scoring method is likely to have an effect on the score reliability and that the relational scoring method used in conjunction with a master map yielded the most reliable scores. The reason for these results may have to do with the relative cognitive load imposed by the different scoring methods on the working memories of the raters.

While the instructions for holistic scoring are simple, the actual scoring task may be cognitively complex. Concept maps vary along many dimensions, and the identification and simultaneous evaluation of these dimensions may place a heavy load on raters' working memory. The task is made more complex when the raters try to compare the quality of many maps. The lack of structure in this scoring procedure leaves it to the rater to devise a way to deal with the complexity of the task. If the rater fails to develop an effective strategy for dealing with this complexity, the consistency of the rater's performance may be degraded, affecting the general reliability of the scores.

The structural scoring method provides raters more guidance than do the holistic methods; however, the cognitive complexity of the task is still high. Raters must determine what constitutes the higher order structures within a concept map (i.e., what constitutes a crosslink or hierarchy). While not as complex as an entire map, hierarchies may still be complex structures and the identification and evaluation of these structures may place a strain on the working memory of the rater.

Of the three methods evaluated, the relational scoring methods are perhaps the most structured. In these methods, raters are directed to consider the propositions represented in the map

separately. The proposition is the least complex structure represented in a concept map; an analysis of concept maps at this level is least likely to tax a rater's working memory.

That the relational scoring method has the lowest cognitively complex is consistent with the observed increase in score reliability, when this method was augmented with a master map. Given the structure provide by this method, the raters were able to accommodate the additional cognitive load imposed by the consideration of the master map. In contrast, the incorporation of a master map with the other two methods actually resulted in scores with lower reliability.

The relative cognitive complexity of a scoring method may be especially critical when, as in the present study, the raters lack familiarity with the specific scoring method and concept maps in general. This is likely to be the case with the average science teacher. If the analysis of the cognitive complexity of the different scoring methods is correct, we might expect that experience with the scoring method might have a differential effect on the reliability of scores. In other words, the performance of raters using cognitively complex scoring methods may be more likely to improve with practice compared to the performance of raters using less cognitively complex methods. The design of the current study did not allow an investigation into this possibility. The change in reliability with raters' experience with specific scoring methods is an area for further study.

Validity of Concept Map Assessments

Concerning the concurrent validity of concept map scores, the evidence provided by this study is encouraging. With the exception of the scores derived from structural with master map, the correlations between map scores and the similarity measure were statistically significant.

The question might be asked, since the similarity measure seems to assess students' knowledge, why not just measure the similarity between students' maps and the teachers'? The answer is that the calculation of this similarity measure is tedious and time-consuming, and it is doubtful whether classroom teachers would be able or willing to invest the time. The scores generated by the relational method with master map correlated most closely with the maps' measured similarity. This is not especially surprising, as both the similarity measure and the relational scoring method are based on an assessment of individual propositions. Given this, scoring maps using a relational method with a master map might be considered a practical alternative to assessing similarity between the teachers' and students' concept maps.

Time and the Adoption of Concept Mapping for Classroom Assessments

A major factor influencing a classroom teacher's adoption of a new technique is the consideration of value of the new technique. The value of a technique may be expressed as the relationship between the potential costs and benefits associated with the technique. In other words, is the effort required by the new technique justified by its usefulness? In this study, time was used as a measure of cost. The cost in terms of time is addressed by looking at (a) the time required to train students in the mapping task, (b) the time required to complete a mapping task, and (c) the time required to rate students' maps.

Students' skill with concept mapping techniques will have an effect on the quality of their products, and therefore affect the score reliability and validity of the assessment. With other more common assessment tasks, teachers may feel comfortable assuming that students are already familiar with the task format and possess the necessary skills to produce reliable responses. However, teachers may not feel the same level of comfort with concept mapping tasks. The question then becomes what kind of training is required for students to produce reliable

concept map representations, and whether this training can be accomplished in a reasonable period of time.

The data presented here suggest that it is possible to train students in a reasonable period using a direct instruction method. In this study, college students were provided approximately 90 min of training and produced maps that could be scored with a score reliability as high as .76. Of that 90 min, most was devoted to guided practice. Classroom teachers may realize some time savings during guided practice if concept mapping is used as a learning activity. Producing maps under the guidance of the teacher, as a means to organize and elaborate course content, can help students learn the material (Novak & Gowin, 1984) while they are acquiring concept mapping skills.

The time required for students to complete a concept mapping task is another consideration affecting the decision to adopt concept mapping as an assessment technique. The average time required for students in this study to construct fairly complex concept map was 29 min. Taken as an estimate of a typical completion time, this easily fits into a standard class period of 40–50 min. The cost in terms of time is comparable to a tradition pencil and paper examination.

The final critical time consideration is the time required for teachers to evaluate and score concept maps. The results of this study indicate that the times required to score concept maps are likely to range from 1 to 5 min, depending on the scoring method selected. These times are likely to compare favorably with objective assessments such as short answer quizzes or tests. It might be more realistic, however, to compare concept mapping to tasks that are similar in the degree to which open-ended responses are permitted, such as essay exams. Based on personal experience, it seems likely that a comparison of the time required to score concept maps with the time required to score responses to essay questions would favor concept mapping.

Recommendation to Science Teachers

Students' acquisition of content knowledge is a common objective of science teachers. Assessing the quality of knowledge acquired by students as they participate in learning various activities is critical to the instructional process. Students' concept mapping tasks may provide science teachers with a unique and valuable source of information.

Science teachers should give careful thought to the exact nature of the concept mapping tasks selected for use in their classrooms. The nature of a concept mapping activity will have an effect on the potential costs and benefits of the assessment process. The primary costs to classroom teachers are the time required to train students to produce concept maps and the time required for teachers to score and evaluate concept maps. The benefit results from the open-ended nature of concept mapping tasks and the opportunity provided by such tasks to assess the idiosyncratic nature of students' knowledge structures. We make three recommendations to minimize costs and maximize the benefits of concept map assessment activities.

The first recommendation is to keep the actual mapping task simple. A complex mapping procedure is likely to require more time for instruction and practice for students to develop the skills to produce concept maps that accurately represent what they know. It is also possible that a complex mapping scheme will constrain students' responses and therefore compromise the information available in the students' maps. The concept mapping task used in this study seemed to work reasonably well.

The second recommendation is to use some form of relational scoring method, preferably with a master map. This recommendation is based on three considerations. First, the method is simple and classroom teachers should be able to master the scoring technique with little or no training. Second, because of the mechanical simplicity of the technique, the scores for maps are

more easily defended. Both holistic and structural methods require broad subjective assessment of complex structures that may be difficult for the teacher to explain or justify to other concerned individuals. Finally, as reported in this study, the relational scoring method combined with a master map yielded the most reliable scores.

The last recommendation is that teachers should use the scores derived from concept mapping tasks and other forms of assessment with caution. The validity of a decision based on a single observation is always suspect. Decisions made based on a concept map score may not be any more valid than decisions based on a single score from a test, essay, or project. A score, whether from a concept map test, standard objective test, or some other type of performance, reduces a complex task to a single number; and while these numbers may be useful in some cases, much information is lost.

Concept maps may be a valuable source of information about both the content and organization of students' knowledge. It is the organizational component captured by concept maps that may allow teachers to identify and correct student misconceptions. However, extracting information about the way individual students organize their knowledge is likely to require a time-consuming analysis, thereby imposing what may be unacceptable costs on a teacher's time.

To maximize the utility of concept map assessments, teachers should approach the scoring tasks as an initial survey that may indicate areas where more in-depth analysis would be profitable. For instance, a low score on a concept map may warrant a closer look to diagnose misconceptions. In addition, as a set of maps is scored, the teacher should be on the lookout for recurring patterns in students' maps that might indicate some problem with instruction. Evaluating students' concept maps in this way may allow teachers to optimize the balance between the costs and benefits concept map assessment tasks.

References

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

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City: American College Testing Program.

Goldsmith, T.E., & Davenport, D.M. (1989). Assessing structural similarity of graphs. In R.W. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization. (pp. 75–87). Norwood, NJ: Ablex.

Goldsmith, T.E., & Johnson, P.J. (1989). A structural assessment of classroom learning. In R.W. Schvaneveldt (Ed.), Pathfinder associative networks: Studies in knowledge organization. (pp. 241–254). Norwood, NJ: Ablex.

Hegarty-Hazel, E., & Prosser, M. (1991). Relationship between students conceptual knowledge and study strategies. Part 1: Student learning in physics. International Journal of Science Education, 13, 303–312.

Holly, C.D., & Dansereau, D.F. (1984). Networking: The technique and the empirical evidence. In C.D. Holly & D.F. Dansereau (Eds.), Spatial learning strategy: Techniques, applications and related issues (pp. 3–19). New York: Academic.

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

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 33–45). Hillsdale, NJ: Erlbaum.

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

Novak, J.D. (1990). Concept mapping: A useful tool for science education. Journal of Research in Science Teaching, 10, 923–949.

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

Surber, J.R., & Smith, P.L. (1981). Testing for misunderstanding. Educational Psychologist, 16, 163–174.

Surber, J.R. (1984). Mapping as a testing and diagnosis device. In C.D. Holly & D.F. Dansereau (Eds.), Spatial learning strategy: Techniques, applications and related issues (pp. 3–19). New York: Academic.

Trowbridge, J.E., & Wandersee, J.H. (1994). Identifying critical junctures in learning in a college course on evolution. Journal of Research in Science Teaching, 31, 459–473.

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.