Concept Maps as Assessment in Science Inquiry Learning - A Report of Methodology

Trish Stoddart, Robert Abrams, Erika Gasper & Dana Canaday

University of California, Santa Cruz

Contact information:
Robert Abrams
382 Central Park West #8K
New York, NY 10025
email: bob@robertabrams.net
www.RobertAbrams.net

This paper has been published in

The International Journal of Science Education

Volume 22, Number 12, December 2000, p. 1221-1246

Completion Date: 10-8-99

Concept Maps as Assessment in Science Inquiry Learning - A Report of Methodology

Trish Stoddart, Robert Abrams, Erika Gasper & Dana Canaday

Summary

Increasing availability of technologies, such as CD-ROMs and the WWW, in schools means that more teachers will have the potential to implement student-centered, inquiry-based approaches to learning. Assessing what each student knows in a broad subject area, such as science, is difficult. Assessing students' understanding in circumstances where each student may pursue different topics of study, where there is no way to predict in advance what those topics of study will be, and where the possible topics of study include natural phenomena which are only beginning to be studied by professional scientists is more difficult. The authors recently faced such a challenge. To meet the challenge, the authors chose to assess student learning using an open-ended concept map activity combined with a rubric which extracts quantitative information about the quality of understanding from each map. This article describes the method the authors developed, including tests of reliability and validity.

Concept Maps as Assessment in Science Inquiry Learning - A Report of Methodology

Trish Stoddart, Robert Abrams, Erika Gasper & Dana Canaday

Introduction

Over the past ten years in the United States, there has been a radical shift in views of K-12 teaching and learning. New curriculum standards have been developed that emphasize the development of students as autonomous learners in active inquiry-based constructivist instructional environments (AAAS 1989; NSTA 1990). Educators emphasize the importance of learning science through doing science, using authentic scientific methods. During the same period there have been rapid advances in multimedia and web-based computer technology. Recently it has been recognized that this technology can provide the tools for open-ended student-driven constructivist learning (Savery and Duffy 1995; Rakes 1996). Computer tools can make abstract concepts manipulable and allow students to act like scientists, learning content in the context of real world problems.

Constructivist recommendations for education have suggested that the best way to teach is to figure out what the student already knows and then teach from there (Ausubel, Novak and Hanesian 1978). This suggestion can be extended by giving students the opportunity to study topics of their own choosing. The increasing availability of computer technologies, such as CD-ROMs and the World Wide Web, in schools means that more teachers will have the potential to implement such a student-centered, inquiry-based approach to learning.

Internet based technology provides students with access to a vast reservoir of information, including data bases that scientific institutions choose to make available. Students have the freedom to independently explore these large databases, and design and conduct their own investigations and experiments. This freedom results in autonomous student learning, which includes students pursuing a wide range of interests. This student autonomy, in the scope and

type of information they access, however, creates an assessment dilemma. Traditional closed-ended methods of assessment, such as multiple choice tests, do not adequately measure student learning in open-ended inquiry learning environments (Escalada and Zollman 1997), although attempts are being made to develop assessments which reach a middle ground, such as computer adaptive testing. When used to measure change in understanding of content, traditional tests can have problematic validity because "under regular school conditions, it is unusual and almost impossible to administer the final test at the beginning of a course; it forces students to make guesses and yeilds an invalid indicant of their knowledge structures" (Hoz, Bowman and Chacham 1997, p. 292) Close-ended tests do not capture students' ability to develop and carry out independent investigations nor do they measure the development of student conceptual understanding (Ruiz-Primo and Shavelson 1996a). This situation is compounded in projects where each student may pursue different topics of study, where there is no way to predict in advance what those topics of study will be, and where the possible topics of study include species and phenomena which are only beginning to be studied by professional scientists. Assessment tools should be sensitive to the content that the students are studying. Ruiz-Primo & Shavelson (1996a, 1996b) propose the use of concept maps and performance based assessments as alternatives to the use of multiple choice tests.

The authors of this paper were recently faced with the challenge of evaluating student learning outcomes in the Virtual Canyon project (www.virtual-canyon.org), an on-line and CD-ROM hybrid learning environment with participating schools in both low and middle socio-economic status areas, as well as both elementary and high school students. The Virtual Canyon project is based upon the Monterey Bay Aquarium Research Institute's ongoing ocean science research. As a result, the possible topics of study on the Virtual Canyon include species and phenomena that are only beginning to be studied by professional scientists. The project is intended as a support structure for learning across a diverse range of topics and student grade levels within the domain of marine science and the region of the Monterey Bay (California). As such, it incorporates features of a dry laboratory (Kirschner and Huisman 1998),

particularly with regards to giving students experience with the development of research questions and testable hypotheses. This student research is supported by an on-line library of resource materials drawn from the Monterey Bay Aquarium Research Institute's (MBARI) recent research, digital video of MBARI scientists and other staff which give students a tacit feel of a research institute, selected MBARI datasets (i.e., raw data collected by scientists with which students can search for patterns), and a communication system. Students are given the opportunity to study topics that reflect their own areas of interest. Students operate as independent researchers who use on-line tools to explore a deep underwater canyon, conduct research experiments in a virtual research laboratory, and post the methods and findings of their projects on a Web page for other students to review. Students aged from eight years to eighteen years participated in the project. The projects they developed were highly individualized. The assessment system needed to measure science learning in students in third through twelfth grade and be responsive to the breadth and depth of content they were studying.

The authors looked for existing assessment instruments, but responses to queries to established centers of marine science education, such as the Shoals Marine Laboratory, indicated that no general assessment of marine science existed (Rivest 1996). Moreover, the researchers needed an assessment that would capture student learning about very specific content: the Monterey Bay Submarine Canyon in particular and the deep ocean in general. Concept mapping was selected because the technique could be used with a wide range of content and with students at all grade levels. This paper describes the development of the method and discusses issues of reliability and validity.

Concept maps

Theoretical underpinnings

Concept maps are a procedure that is used to measure the structure and organization of an individual's knowledge (Novak and Gowin 1984; Ruiz-Primo and Shavelson 1996a). Concept

mapping was originally developed by Novak and the members of his research group as a means of representing frameworks for the interrelationships between concepts (Stewart, Van Kirk and Rowell 1979; Novak and Gowin 1984). Concept maps as originally developed have been grounded in a psychological theory which focuses on individuals and how they integrate new learning into existing conceptual frameworks (Ausubel 1968; Ausubel et al. 1978; Novak and Gowin 1984), by making explicit, conscious connections between concepts as a way to integrate information into memory (Anderson 1992; Bruer 1993; Vosniadou 1996). The basic element of a concept map consists of concept words or phrases that are connected together with linking words or phrases to form complete thoughts called 'propositions' (e.g. Concept --> linking word --> Concept). Researchers have continued to develop and refine this technique for use in teaching, learning, research and assessment. Concept maps have been used for many instructional purposes, in many subjects, and with many levels of students.

Previous uses of concept maps

Concept maps have been demonstrated to be a powerful instructional tool which assists students in clarifying their understandings and makes connections between concepts explicit (Markow and Lonning 1998). Educators have found concept maps useful to assess prior student knowledge, to identify gaps in student knowledge, to help teacher education students identify key concepts to target in their teaching, and as an assessment tool to determine the extent and quality of new connections students are able to make after instruction (Mason 1992).

Concept maps have been used in the study of physics (Roth and Roychoudhury 1994; Gangosa 1996), chemistry (Stensvold and Wilson 1990; Markow and Lonning 1998), ecology and environmental education (Brody 1993; Heinze-Fry 1997), biology (Heinze-Fry and Novak 1990; Jegede, Alaiyemola and Okebukola 1990; De Groot 1993; Markham 1993; Songer and Mintzes 1993; Farrokh and Krause 1996; Coleman 1998), history (Baldissera 1993), astronomy (Zeilik, Schau, Mattern, Hall, Teague and Bisard 1997), veterinary medicine (Edmondson 1995), engineering (Moreira and Greca 1996), literature (Leahy 1989; Moreira 1996), geology (González

1993), and mathematics (Khan 1993; Moreira and Motta 1993). More recently, concept maps have been adapted for use in business settings (Novak 1998, p. 120).

Concept maps have been used with elementary students (Eschenbrenner 1994), middle school students (Sizmur and Osborne 1997; Coleman 1998), high school students (Stensvold and Wilson 1990), and college students (Heinze-Fry and Novak 1990; Pearsall, Skipper and Mintzes 1997) including teacher education students (Mason 1992). While concept mapping was originally developed for use with the English language, recent work has begun to explore how concept mapping might be adapted for use with English Language Learners whose first language uses a sentence structure different from that used in English. In one such recent study, Korean speaking middle school students used concept mapping with and without various language accommodations (Lee 1999).

Methodological Issues in Concept Mapping as Assessment

Concept mapping as assessment has two components: a task that students perform to demonstrate and record their knowledge, and a scoring system which a researcher or teacher uses to evaluate the students' knowledge.

Concept maps are typically produced in one of two ways: the students create their own maps, or the students demonstrate their knowledge in another format, such as an interview or a writing assignment, which the researcher uses to develop the concept map. The latter task type is characteristic of the early history of concept mapping. It is useful for certain kinds of in-depth studies, but is not practical for large scale assessments. The methodology reported in this paper is of the type where students produce their own concept maps as there were several hundred students participating in the Virtual Canyon project.

If the task is one where students draw their own concept map, the task format can be constrained or open-ended, with various intermediate possibilities. Constrained tasks are tasks which restrict the mapper to a supplied list of concepts and/or link words (Markham, Mintzes and Jones 1994; Osmundson, Chung, Herl and Klein 1999), or use a fill in the blank approach (Zeilik et al. 1997; Coleman 1998). Open-ended tasks supply a small number of

prompt concepts, and otherwise do not restrict how the map may be drawn. Intermediate tasks are those that specify a list of concepts to be used, but place little or no other restrictions on how the map can be drawn.

The approaches to scoring concept maps generally combine an interest in the content validity or accuracy of the content displayed in the map with an interest in the elaborateness of the map as measured by counting various map components, such as concepts or links. Early scoring systems tended to place much emphasis on elaborateness. Novak and Gowin (1984) originally proposed a scoring system in which the number of valid propositions, levels of hierarchy, examples, and cross-links are counted. Each of these counts is given a weight (for instance levels of hierarchy might be multiplied by 5, while number of valid propositions might be multiplied by 1), and then the weighted counts would be added to obtain a final score.

More recent scoring systems show a trend towards more sophisticated ways to assess a concept map's content validity with a relative de-emphasis on a count of map components (Ruiz-Primo, Schultz and Shavelson 1997, p. 7; Rice, Ryan and Samson 1998). The student learning, as represented in the concept maps, is often measured by comparing a student's map to an expert map. The expert map can be a single concept map produced by a recognized expert in a given topic (Coleman 1998), the concept map on the given topic produced by the students' teacher, the concept maps produced by the students' teacher and a group of experts (Osmundson et al. 1999), or through the use of what could be called an expert link matrix (Ruiz-Primo et al. 1997, p. 12). This latter option consists of a process in which one or more experts on the given topic produce an exhaustive set of possible relationships between each pair of concepts in the allowed set. These possible relationships can then be categorized in various ways.

As shown in table 1, elaborateness scoring systems focus on the number of map components and not the content validity, or accuracy. Validity scoring systems are those in which content validity, or accuracy are the sole criteria. Mixed scoring systems are systems that use both elaborateness and accuracy with roughly equal weight.

Table 1: Comparison of concept map assessment systems.

	Scoring System		
Task/Response Format	**Emphasizes Elaborateness/Map components**	**Uses a mix of Elaborateness and Validity Criteria**	**Emphasizes Validity/Accuracy**
Constrained	(McClure and Bell 1990; Baker, Niemi, Novak and Herl 1991)	(Anderson and Huang 1989)	(Zeilik et al. 1997; Coleman 1998; Osmundson et al. 1999)
Intermediate	(Wallace and Mintzes 1990; Markham et al. 1994; Wilson 1994)	(Champagne, Klopfer, DeSena and Squires 1978; Novak et al. 1983; Hoz, Tomer and Tamir 1990; Mahler, Hoz, Fischl, Tov-Ly and Lernau 1991; Nakhleh and Krajcik 1991; Schreiber and Abegg 1991; Roth and Roychoudhury 1993)	(Ruiz-Primo et al. 1997, p. 10 - task b & c; Hewson and Hamlyn Date Unknown)
Open-ended	(Lay-Dopyera and Beyerbach 1983; Heinze-Fry and Novak 1990; Barenholz and Tamir 1992)	(Beyerbach 1988; Lomask, Baron, Greig and Harrison 1992; Pearsall et al. 1997)	(Ruiz-Primo et al. 1997, p. 10 - task a) The approach to concept mapping as assessment described in this article also is an example of Open-ended task with scoring that emphasizes Validity.

The variety of concept map assessment systems can be understood by using tasks as one scale and using scoring systems as a second scale. These two scales define table 1. With the exception of five articles (Pearsall et al. 1997; Ruiz-Primo et al. 1997; Zeilik et al. 1997; Coleman 1998; Osmundson et al. 1999), all of the studies in table 1 were placed based upon the information supplied in Ruiz-Primo & Shavelson's (1996a, table 1) review of the literature.

In this project, the authors chose to use an open-ended task and response format to parallel the inquiry-based curriculum used in the Virtual Canyon project. While the authors felt that an emphasis on validity and accuracy for the scoring system, as suggested by Ruiz-Primo, Shavelson, Rice and others, would be appropriate for assessing the Virtual Canyon project, the

wide range of possible topics of study on the Virtual Canyon project precluded the use of an expert link matrix. Instead, the scoring system reported in this paper was developed by deriving categories arising from the student data (see table 2 and appendix B to see the final scoring system). Such an approach is consistent with a grounded theory approach to analysis (Patton 1990; Strauss and Corbin 1990; Aikenhead and Ryan 1992; Sizmur and Osborne 1997).

Correlation between concept map scores and conventional tests

Concept maps assess many of the same aspects of learning that conventional tests measure, but they also measure aspects of learning which conventional tests do not measure particularly well (Ruiz-Primo et al. 1997, p. 23). There are moderate correlations between concept map assessments and conventional tests. Students' performance on concept map assessments has been found to be significantly correlated with more conventional assessments such as multiple choice tests (Liu and Hinchey 1993; Liu and Hinchey 1996; Rice et al. 1998). A moderate correlation was found between concept map scores and course grades in a college biology course (Farrokh and Krause 1996). Concept maps have been found to be predictors of student performance on traditional school based tests and national standardized tests (Wilson 1993). A study with ten and eleven year old students found a correlation between concept map content scores and essay scores that was significant and of moderate magnitude (Osmundson et al. 1999).

The strength of the correlations have been found to vary depending on three factors: the type of conventional test, the type of the concept mapping task, and the type of the concept mapping scoring system. The correlation between conventional tests and concept map tests have been found to vary with the type of conventional test. Wilson found higher correlations between concept map scores and test scores that measure application of knowledge compared to lower correlations between concept map scores and test scores that measure recall of knowledge (Wilson 1993). This is consistent with earlier results (Novak, Gowin and Johansen 1983).

Work has also been done to show that the type of concept map task format (open-ended or constrained) affects the strength of the correlation. In a study on general science classes, Liu & Hinchey (1996) found that the correlation between student scores on concept map assessments

which used open-ended tasks and conventional test (multiple choice and short answer items) scores was higher than the correlation between student scores on constrained concept map tasks and the conventional test scores.

The type of concept map scoring system (elaborateness or accuracy) affects the strength of the correlation. Scoring systems that emphasize content validity are highly correlated with conventional tests. Hoz et al. (1997) found that students' rank score on an objective test of geomorphology was significantly correlated with the accuracy of their concept maps: the more accurate the concept map, the higher the test score. Stensvold & Wilson (1990) found that the number of accurate links made on a concept map predicted students' comprehension test scores.

Concept maps measure aspects of learning which conventional tests do not measure particularly well. For example, concept mapping can provide information about students' misconceptions and incorrect conceptions, which are usually unavailable in conventional tests (Liu and Hinchey 1993; Rice et al. 1998). For instance, Rice et al. found that their students had confused ‘salamander’ and ‘lamprey’. The researchers were able to find a possible flaw in their curriculum (lack of specimens of these species for students to examine, in contrast to other species for which specimens were available) with the aid of their students' concept maps (Rice et al. 1998).

Finally, the authors would like to echo Mintzes and colleagues' view that further correlation studies are needed if the concept map is to become a standard procedure in large scale assessments (Pearsall et al. 1997). To date, there have been no correlation studies which explicitly consider all three factors. Based upon the data available to date, the authors would predict that higher correlations would be found between conventional tests that measure application of knowledge and concept map assessments that use relatively open-ended tasks and emphasize accuracy in their scoring systems.

The method

The method is intended to measure change in understanding over time in the context of an inquiry learning project where the specific content to be learned could be different for each

student, and could not be predicted in advance. The method consists of a data collection phase (the tasks given to the students), a scoring phase (coding and analysis of the data), and a verification phase (a statistical test used to confirm the trustworthiness of the results).

Data collection phase

The task used in the data collection phase is an open-ended concept mapping activity (see appendix A for example activity sheets). Students are given paper, stickies, and a pen with which to draw their concept maps. Students are instructed to return the activity sheet and their concept map. Except for a slightly different wording of the activity sheet, the training and assessment protocol is the same for both elementary and high school students.

The activity has several key characteristics: it uses a minimum of prompt concepts which are representative of the most general level of the curriculum being assessed. All versions of the activity (training, pre, post; High School, Elementary) use similar task demands. This activity is administered three times. The first administration is a training session, the second is a pre-assessment, and the third is a post-assessment.

The first administration, the training session, requires a full 45-50 minute class period. During this time, the researchers introduce all of the components of a concept map, lead the class in producing a group map on the blackboard, and administer a practice concept mapping activity. The topic used for the practice activity was a non-marine science topic students had been studying recently. The researchers consulted with each teacher to determine what their students had recently studied so that an appropriate practice topic could be selected. The authors can not stress enough the importance of a consistent protocol for training students in concept mapping before administering the assessment activity. Anything less than a full 45 minute training session is probably inadequate. Ruiz-Primo et al. (1997, p. 15) also used a 50 minute training protocol. One common problem is that students will draw maps without linking words. The importance of linking words should be emphasized repeatedly since a concept map without linking words is mostly unscorable.

The second administration, the pre-assessment, is conducted within a week of the training session, and ideally, before students' initial exposure to the curriculum being studied. The pre-assessment session also requires a 45-50 minute class period. During the first 10 minutes of this period, the researcher reviews the main components of a concept map. Students have the remaining 35 minutes to produce maps, although they rarely take this much time.

The third administration, the post-assessment, is completed within a week of completion of final student projects using nearly identical procedures as the pre-assessment.

Scoring phase

Scoring for scientific discourse is conducted in three stages: vocabulary review, content scoring, and a content validity check. The score sheets can be found in appendix B. An example concept map showing marking for scoring can be found in appendix C. The use of multiple stages during scoring reduces errors while maintaining the efficiency of the scoring process. One of these stages concerns the measurement of student use of scientific vocabulary, which will be discussed in a forthcoming methodology article.

The reviewers score each concept map for content. Concept maps can provide multiple insights into student understanding in part because concept maps can be analysed at multiple levels. For the analysis described here, each concept map was scored for content using criteria assessed at the level of each proposition within the map. Use of analysis at the proposition level generates greater variability than analysis for the same component of learning at the map level, resulting in more effective statistical analysis. Map level analysis can complement proposition level analysis under some circumstances, but was of lesser importance for the Virtual Canyon project (see endnote 1).

The criteria are category systems which arise from the data (Patton 1990, p. 390; Strauss and Corbin 1990). This approach to category development has been used in the study of concept mapping (Sizmur and Osborne 1997), as well as in other large scale studies, such as Views on Science Technology and Society (VOSTS) (Aikenhead and Ryan 1992).

Each proposition is numbered and scored on three variables: Accuracy, Explanation, and Proposition Structure. A summary of these variables with examples is shown in table 2.

There are four levels of Accuracy used for analysis: (i) scientific accuracy, (ii) common knowledge, (iii) inaccurate statements, and (iv) affective statements. The scoring rubric for the accuracy variable includes an additional three categories which help to remove scoring errors. 'Question' is a category for propositions in the form of a question. Such propositions can not be said to be either accurate or inaccurate. 'Makes No Sense' is used to keep track of propositions whose accuracy can not be scored because the handwriting is unintelligible, the spelling is very poor, or the grammar is very poor. 'Don't Know' is a category used when the reviewer does not have sufficient knowledge to judge the accuracy of the proposition. Because students were studying new research, there were some items which were beyond the reviewers' science background. A list of all propositions marked 'Don't Know' is compiled so that a check of content validity can be conducted. This list of all propositions which were marked 'Don't Know' in the Accuracy variable is brought to a reviewers meeting. Any propositions for which the accuracy can not be resolved at the reviewers meeting are sent to a subject matter expert (in this case the Monterey Bay Aquarium Research Institute's liaison to the Virtual Canyon project) for evaluation. The subject matter expert's evaluation is then used to establish the accuracy scores for those propositions.

The levels of accuracy are defined as follows:

(i) Scientific accuracy is defined as correct statements about scientific content, with 'scientific' meaning content which is typically learned in K-12 school science curricula, content of a particular field of science, and content learned from the scientific process. The latter includes specific observations such as exact measurements. Examples of propositions that would be scored as 'scientifically accurate' include, 'whales are mammals', 'Pressure increases with depth in the

ocean', and 'chemosynthetic bacteria live in the gills of clams found in cold seep sites'.

(ii) Common knowledge is defined as non-scientific, everyday knowledge. Examples of propositions that would be scored as 'common knowledge' include, 'whales are big', 'dolphins live in the sea', vague statements about things 'living in the ocean', and 'shells are on the beach'.

(iii) Inaccurate statements are those that are commonly accepted by scientists to be incorrect, at the level of complexity appropriate to K-12 school science curricula. Examples of inaccurate propositions are, 'sharks are mammals', and 'seals eat clams'. Inaccurate statements comprise less than 10% of the overall group of maps. This is most likely due to the open-ended nature of concept mapping, where students are asked to report what they do know, rather than address specific content areas that may be beyond their expertise.

(iv) Finally, affective statements are defined as those that express emotions, feelings or personal thoughts. Examples of affective propositions include, 'dolphins are pretty', 'deep sea creatures are cool', and 'I love whales'.

The Depth of Explanation criterion differentiates between (i) basic descriptions, and (ii) higher-order explanations:

(i) Basic descriptions are defined as factual statements, often answering 'what' questions. Examples of propositions that are scored as 'basic description' include, 'whales are mammals', 'siphonophores eat jellies', 'Anglerfish have bioluminescent danglers', and 'hatchetfish have upturned eyes'.

(ii) 'Higher-order' is defined as explanations that describe function or purpose. They often address 'how' or 'why' questions. Examples of propositions scored as 'higher-order explanations' include, 'shining tubeshoulders squirt a bioluminescent cloud to

confuse predators', and 'anglerfish have bioluminescent danglers above their mouths that are used to attract prey'. (The rubric distinguishes between How and Why, but these categories are combined for analysis.)

The Complexity of the Proposition Structure criterion is used to assess the elaboration of an idea within a proposition, and uses two levels which are (i) simple, and (ii) compound:

(i) A simple proposition is defined as a proposition containing only one subject-object clause. Examples of propositions scored as 'simple' include, 'whales are mammals', 'dolphins eat fish', 'Shining tubeshoulders have photophores', and 'fangtooth belongs to species angloplogasteridae'.

(ii) A compound proposition is defined as a proposition containing one or more dependent clauses. Examples of compound propositions are, 'anglerfish reproduce by the male bonding to the female and staying there for life', and 'Shining tubeshoulders have photophores on their undersides and heads'.

Table 2: Proposition Variable Categories

Variable	Category	Example
Accuracy	Scientifically Accurate	Pressure increases with depth in the ocean
	Common Knowledge	Whales live in the ocean
	Inaccurate	Sharks are mammals
	Affective	Dolphins are pretty
Depth of Explanation	Descriptive	Anglerfish have bioluminescent danglers
	Higher-order explanation (answers 'how' or 'why')	Anglerfish have bioluminescent danglers above their mouths that are used to attract prey
Complexity	Simple	Shining tubeshoulders have photophores
	Compound	Shining tubeshoulders have photophores on their undersides and heads

The scoring rubric produces data with a relational structure. A relational database is a data structure in which two or more databases are linked to each other. In this case, one database contains information about each concept map, while the second database contains information about each proposition. The data must be transformed into a flat data structure before statistical analysis so that each student's pre and post concept map can be compared. The authors created a custom data processing application to flatten the data into both proposition level and map level databases. The application also matches pre and post maps so that only students who drew both maps are included in the database used for statistical analysis, and it creates a variety of computed variables, such as number of scientifically accurate propositions per map. The custom data processing application would not be needed if there were a statistics program that was built around a relational database.

Finally, statistical analysis is performed on the data. To assess the quality of student learning, the authors calculated the proportion of scientifically accurate propositions relative

to all propositions, and the proportion of higher order explanation propositions (how or why) relative to all propositions, in each concept map. The authors chose to use proportions as their measure because they wanted to be able to compare elementary and high school student concept maps. Older students tend to produce maps with more propositions. The authors were also interested in a weighted measure of the quality of student understanding, not necessarily the quantity. Users of the method can also report raw category counts.

Verification phase (content validity test)

The reviewers who scored the concept maps were confident that their own scientific background in combination with the assistance of the subject matter expert resulted in accuracy scores which were valid. To be absolutely confident, content validity was confirmed by verifying the scientific accuracy of a random sample of 327 propositions against scientific texts (Press and Siever 1986; Garrison 1993; Davenport 1998). A minimum of 15% of the maps from each Virtual Canyon class was selected using a random numbers table (Moore and McCabe 1989). Since the number of maps selected was always rounded up, there were 17.4% of all maps in the random sample for the validity test (17.4% = 126 maps). 'All maps' in this case refers to the set of 724 maps composed of all maps drawn by students who drew both a pre and post concept map. All 327 propositions which were scored as 'Scientifically Accurate' in these 126 concept maps were checked against the scientific texts and resources. The same data was used to calculate the concept map results reported in the Virtual Canyon research team's final report to the National Science Foundation and in a forthcoming paper. 266 responses were verified by Garrison (1993), 15 responses were verified by Press & Siever (1986), and four responses were verified by Davenport (1998).

Due to the cutting edge nature of some of the content that students were learning, 42 of the responses could not be verified by text resources. Of these, 15 were verified using the Virtual Canyon web site (MBARI 1998), three using the Pelagic Shark Foundation web site (Foundation 1998), and the remaining 24 were verified by a MBARI scientist. Validity was determined to be 97.5% (319 of 327) agreement between the reviewers' original scores and the information found

in the scientific texts and other resources. Of the eight propositions for which there was disagreement between the original score and the scientific texts and resources, three of the propositions would be classified as 'Commonly Accurate' based on the information in the texts, and the other five would be classified as 'Inaccurate'. Of the latter five, one of the propositions (Black dragon females swim to the surface at night to feed) would have been considered 'Scientifically Accurate' a couple of years ago, but would be classified as 'Inaccurate' today based on updated scientific research.

Inter-rater agreement and reliability of the method

The process of carefully deriving categories which emerge from the data is as much a part of the method as the final rubric itself. This is an inductive, qualitative process in which initial measurement ideas are tested against the data through review, coding, and identification of dominant themes (Bogdan and Biklen 1992). This process of category development centered on three tests in which the inter-rater agreement and reliability of the categories was determined with a random selection of student concept maps. The results of the final test are discussed here. This final test was conducted on the penultimate rubric.

The test was conducted for each of the three proposition level variables that form the core of the rubric (Accuracy, Level of Explanation, and Complexity). To conduct the test, a stratified random sample of 15% of the concept maps from each class were selected using a random numbers table (Crowl 1996). This resulted in a sample containing 139 concept maps and 548 propositions. All three researchers individually scored every map in this sample.

The Agreement test revealed that the Accuracy variable needed improvement. A review of the propositions where there was disagreement revealed that most of the disagreement concerned the distinction between commonly accurate and scientifically accurate. 128 such propositions were identified and were examined by the three reviewers. This review of 128 propositions revealed criteria which could be used to operationalize the distinction between commonly accurate and scientifically accurate. The agreement status of 112 propositions was resolved by the refined distinction. This improved distinction raised the inter-rater agreement

for accuracy from 57% to 78%. This refined distinction is the main difference between penultimate version of the rubric and the final version.

Inter-rater reliability

The SPSS 6.1 statistical package was used to calculate Cronbach's Alpha coefficient, a measure of inter-rater reliability (Markow and Lonning 1998). Cronbach's Alpha coefficient is a measure of the consistency across raters.

For the purposes of the reliability test, the Accuracy and Level of Explanation variables had to be reduced from a categorical variable to a series of dichotomous variables. This was necessary because Cronbach's Alpha requires one to calculate means, which would be an inappropriate calculation to make on a categorical variable. The original Accuracy variable was reduced to four variables: Affect, Inaccurate, Commonly Accurate, and Scientifically Accurate. In each case, the proposition would be marked as '1' if it had been scored as the response in question, or a '0' if it had been scored as anything else. For example, a proposition that had been scored 'affect' in the original variable would be marked '1' in the 'Affect' variable, and '0' in the other three. The original 'Level of Explanation' variable was reduced to a single dichotomous variable in which 'How' or 'Why' would be marked '1', and 'What' would be marked '0'. The Complexity variable was already a dichotomous variable, so it did not need to be reduced.

The Inter-rater reliability results are at an acceptable level (traditionally 0.70) or better, for all the variables except Inaccurate and Commonly Accurate, as shown in table 3. However, there are relatively few cases in which a proposition was scored as being 'Inaccurate'. In other words, this variable has a somewhat skewed distribution, which can affect Cronbach's Alpha. The somewhat low Commonly Accurate result is likely due to this being a test of the penultimate version of the rubric. The distinction between common and scientifically accurate propositions was improved for the final version. The other components of the Accuracy variable's reliabilities are high. Cronbach's Alpha was calculated using a random sample from both pre and post concept maps in order to make sure that the reviewers were reliable

across all types of propositions. Some researchers prefer to calculate Cronbach's Alpha using only post concept maps. Such a calculation is shown in table 3 for reference purposes.

Table 3: Inter-rater Reliability of Proposition Level Variables

Variable	Cronbach's Alpha pre & post	Cronbach's Alpha post only
Affect	0.90	0.88
Inaccurate	0.60	0.46
Commonly Accurate	0.64	0.61
Scientifically Accurate	0.74	0.70
Level of Explanation (how & why)	0.81	0.78
Complexity	0.76	0.74

Inter-rater Agreement

The original categorical data was used to calculate inter-rater agreement for the three primary variables. Inter-rater agreement was defined as the percentage of cases in which all three reviewers classified a specific variable (Accuracy, Level of Explanation, Complexity) for a given proposition in exactly the same way (e.g. 'affect' for the Accuracy variable). The Inter-rater agreement results are all relatively high for the number of categories in each variable, as shown in table 4.

While inter-rater agreement is often reported as the raw percent of agreement, the multi-rater Cohen's Kappa is a more robust measure of inter-rater agreement because it accounts for the number of response options, the number of raters, and the likelihood that some of the agreement is due to chance. 'The kappa coefficient will equal 1 if there is perfect agreement, whereas 0 is what would be expected by chance alone' (Vermeylen 1998). Fleiss (1981) has

suggested that 'for most purposes, values greater than 0.75 or so may be taken to represent excellent agreement beyond chance, values below 0.40 or so may be taken to represent poor agreement beyond chance, and values between 0.40 and 0.75 may be taken to represent fair to good agreement beyond chance.'

Table 4: Inter-rater Agreement of Proposition Level Variables (Cohen's Kappa & Raw Percent)

Variable	Kappa	Percent of Agreement	
	Penultimate	Version	Final Version
Accuracy (all response options)	0.45	0.57	0.78
Accuracy (Scientifically Accurate Only)	0.47		
Accuracy (Scientific and Common combined)	0.63		
Explanation (all response options)	0.48	0.86	0.86
Explanation (How and Why combined)	0.58		
Complexity (all response options)	0.47	0.70	0.70

Kappa for the three variables exceeds the minimum acceptable value even in the penultimate version of the rubric (table 4: Accuracy - all response options = 0.45, Explanation - all response options = 0.48, Complexity - all response options = 0.47). Two calculations were performed which provide Kappa in the context of how the variables were ultimately used. Since the accuracy variable is primarily intended to separate scientifically accurate propositions from all other kinds of propositions, Kappa was calculated where the accuracy variable was reduced to two categories: Scientifically accurate as one category, and everything else in another. When calculated this way, Kappa = 0.47. Similarly for the Explanation variable, when the How and Why categories are combined, Kappa = 0.58. Finally, the authors wanted to know Kappa for the final version of the rubric. Since an inter-rater test (where multiple raters all score a sub-sample of the maps) was not conducted using the final version, the best that can be done is to reduce the penultimate version inter-rater test data in a way

which reflects the improvement made to the Commonly Accurate / Scientifically Accurate distinction. Thus, the Kappa for Accuracy in the final version can be estimated by combining Commonly Accurate and Scientifically Accurate into one category, and everything else into a second category, which yields an estimate of Kappa for the final version of 0.63, which is comfortably above the minimum. While the authors would recommend that future work include an inter-rater test along with final applications of this type of scoring rubric, they are not too concerned vis a vis this current study because a validity test using final version data was also conducted (see above). Since the data used in this validity test was randomly selected, the work of all three reviewers is included in this validity test. Since the result of the validity test was very strong, it follows that all reviewers contributed to the result that the scoring is valid when checked against scientific texts.

In addition, since the concept maps were assessed at the map level by calculating the proportion of scientifically accurate propositions to all propositions in each map, it is likely that some of the disagreement among raters at the proposition level would not appear at the map level because while individual propositions may have been scored differently, the total number of scientifically accurate propositions found in a given map by each rater would be the same. For the penultimate version, the mean difference between the reviewer who found the most scientifically accurate propositions and the reviewer who found the least scientifically accurate propositions was 1.3 scientifically accurate propositions per map. Since 112 propositions' agreement was resolved by the refined distinction between common and scientific accuracy used for the final version, the mean difference for the final version should be 0.8 lower (112 propositions / 139 concept maps), or an average disagreement of 0.5 propositions per map.

Based upon these reliability, agreement, and content validity tests, the authors feel that this concept map scoring rubric extracts important components of learning from the concept maps in a way which is practical, reliable, and valid.

Opportunities for future work on the method

Opportunities for future work on the method include increasingly refined measurement of scientific discourse, studies of the interaction effects of concept maps as instruction crossed with concept maps as assessment, tests of the reliability and validity of the method under typical teaching conditions, tests of the feasibility of concept mapping as assessment where n > 1000, and research which examines the complementarity of constrained and open-ended approaches to concept mapping as assessment.

Increasingly refined deconstruction and measurement of scientific discourse

The Accuracy variable in the concept map scoring rubric for the Virtual Canyon study was designed to evaluate whether students had crossed a certain threshold of sophistication in the information they were able to communicate via their concept maps. When the rubric was developed, the primary concern was to be able to detect change in the quality of student understanding. Statements were characterized as affective, inaccurate, commonly accurate, or scientifically accurate and then the proportion of scientifically accurate statements to all statements was computed.

It should be noted that the current category of scientifically accurate statements might fruitfully be subdivided for future studies of student science learning. As students develop increasingly sophisticated scientific knowledge, it might be useful to distinguish their statements via a more elaborate hierarchy of Accuracy than the authors have applied in this study. To some extent, the Depth of Explanation variable accomplishes the goal of more finely distinguishing the quality of a student's statement, and thereby provides an additional threshold level for evaluation of student understanding. Statements which answer questions of How or Why tend to be more sophisticated than statements which are purely factual. As one attempts to classify student statements ever more finely, however, the classification may be increasingly ill-defined, so one might easily reach a point of diminishing returns as expressed in the reliability of the classification scheme. Discussions with Virtual Canyon scientist

partners were especially valuable in helping the authors strengthen the distinction between commonly accurate and scientifically accurate statements. Future discussions, between physical science and education researchers, as well as teachers, could help all participants on a project such as the Virtual Canyon develop a more explicit understanding of scientific discourse and how to best teach the components of that discourse to students.

Interaction effects of concept maps as instruction crossed with concept maps as assessment

More work needs to be done on the interaction effects of concept maps as instruction and concept maps as assessment, with particular attention to the effects upon students with prior records of high achievement compared to students with prior records of low achievement. One early study of this type examined the correlations between college students' SAT scores, their concept maps drawn during a biology course, and grades obtained on conventional tests during the course. The results indicated that the experience of students drawing their own concept maps had an immediate and positive impact on students with high SAT scores, and a delayed and positive impact on students with low SAT scores. The study also found that concept maps were associated with improved learning efficiency for concept mappers compared to the matched non-mapping control group (Heinze-Fry and Novak 1990). By contrast, another study found that students with low vocabulary scores who used concept mapping did better on the final comprehension test than their similarly able peers in the non-mapping group, but students with high vocabulary scores who used mapping did less well than their similarly able peers in the non-mapping group (Stensvold and Wilson 1990) . A hypothesis which would resolve this apparent contradiction is that concept mapping contributes to an improvement in low performing students' study strategies, and also contributes to improvements for high performing students, except for those high performing students whose high performance is strongly dependent upon rote mode learning strategies. Furthermore, concept maps may constitute a kind of activation energy barrier in the sense that they require more work up front than rote mode learning strategies, but have a greater long term learning potential. An ideal study would include subject matter studied among the conditions included in the study, and learning attitude

as well as achievement variables among the outcome measures by drawing on meta-analysis of concept mapping as instruction (Horton, McConney, Gallo, Woods, Senn and Hamelin 1993). Future studies should also follow up on the work of Mintzes and colleagues who have used concept maps to assess student understanding at several points within a course (Pearsall et al. 1997). Such longitundinal studies provide richer data than pre-post studies, but also make it more likely that the assessment concept maps will have an instructional effect. Future studies should also consider the possible need to amend the rubric described in this paper in order to ensure that assessments used are sensitive to the curriculum being studied. The most likely place the rubric would be amended is in the statements which specify the distinction between commonly accurate and scientifically accurate. The rubric could also be modified by adding a variable to measure the relevance of the concept map to the curriculum. This would allow studies which used the method to investigate student choice of topic as a measure of the dynamics within an inquiry based classroom.

Reliability and validity tests under typical teaching conditions

It should be noted that the inter-rater agreement results, the inter-rater reliability results, and the validity results reported above apply to the team of three researchers who were primarily responsible for developing the rubric and then scoring the concept maps. The next step, both for the continued development of the rubric and for a further demonstration of the reliability of the rubric, would be to train additional researchers in the use of the rubric and then run another reliability test with that larger group of researchers. If possible, these tests should be performed using final versions of the rubric, as well as development versions of the rubric. Also, the rubric was designed to be of practical use to teachers and other educational practitioners. Further tests of the rubric would be needed to determine whether the rubric is reliable when used by teachers within the constraints of typical teaching loads. In other words, such a future study needs to ask the question, Can teachers score concept maps themselves using this method in a timely manner? We are confident that the answer to this question is

'Yes', and that the time demands are equivalent to a teacher scoring written responses to short answer questions.

The authors could have chosen to compute an overall score for each map based upon weighted accuracy and explanation subscores, but the accuracy and explanation scores were sufficiently revealing on their own that they didn't think an overall score was necessary. If the method were to be used as an in class assessment replacing a multiple choice test, an overall score would be called for. In this case, the equation used to generate the score would have to be carefully justified and validated.

Feasibility of concept mapping as assessment in studies where n > 1000

This project used the hand-drawn / stickies method of concept mapping with approximately 400 students. The authors feel that this approach is feasible with an n in this range. Scoring initially took about five minutes per map, and took less time as the reviewers gained experience with the use of the rubric. If the number of students increased above 400, such as to several thousand, the logistics of processing hand drawn maps might become too unwieldy. If concept maps could be drawn using a computer program (such as LifeMap, which is available free of charge from the authors at www.mlrg.org), this approach should be usable with very large numbers of students. Such computer based data collection, of course, would present a different set of logistical challenges which will need to be tested in the future.

Complementarity of constrained and open-ended approaches

The authors believe that the constrained / expert-knowledge and open-ended / student-knowledge approaches are complementary. This rubric has components which are similar to those derived independently for other studies (Hoz et al. 1997, p. 932; Ruiz-Primo et al. 1997, p. 14), such as a variable with three or four categories which assess a proposition's accuracy. The difference between other such category systems and this system is threefold. First, this system was derived from student data, whereas the other systems are derived from data generated by selected experts (although part of this system is checked by experts, and some other systems

incorporate some student data (Ruiz-Primo et al. 1997, p. 13), so the difference is mostly in the emphasis). Second, other systems are intended to be used as a scale, whereas this system is most appropriately used as a threshold. In other words, the reviewers try to reliably remove several sets of statements which are not considered sufficiently scientific or accurate. What remains are statements which are at least accurate when considered in the context of the students' developmental level. Third, this system measures student responses along multiple dimensions of scientific discourse (accuracy and level of explanation), whereas other systems only look at accuracy. Although level of explanation has not been used in other concept map scoring systems, it has been used in the tradition of which concept maps are a part: in a study using reverse Vee diagrams to infer students' epistemological stances from interview data (Ault, Novak and Gowin 1988).

Open-ended concept map assessments can reveal the topics that students are likely to study when given a choice, and the concepts that students are likely to learn in that study. Panels of experts could be convened to help refine the curricula available to support such likely topics of student inquiry. The combined conceptual maps of student and expert understanding of science would improve teachers' ability to assess their students, and would contribute to further refinements to concept map assessment rubrics for both constrained and open-ended activities. Such a combined approach would be compatible with recent calls for ‘evaluation which is dynamic and ongoing’ (Roth and Roychoudhury 1994), with the idea that concept mapping is or should be ‘a recursive not a linear process’ (Rafferty and Fleshner 1993), and with the suggestion that concept maps could be used in such a way that a single evaluation activity could be used to assess multiple levels of learning (Rice et al. 1998, p. 1124).

References

AAAS, 1989, Science for all Americans. American Association for the Advancement of Science.

Aikenhead, G.S. and Ryan, A.G., 1992, The Development of a New Instrument: "Views on Science-Technology-Society" (VOSTS). Science Education, **76**,5, 477-491.

Anderson, O.R., 1992, Some interrelationships between constructivist models of learning and current neurobiological theory, with implications for science education. Journal of Research in Science Teaching, **29**,10, 1037-1058.

Anderson, T.H. and Huang, S.-C.C., 1989, On using concept maps to assessthe comprehension effects of reading expository text (Technical Report No. 483). ERIC Document Reproduction Service No. ED 310 368, Center for the Studying of Reading, University of Illinois at Urbana-Champaign.

Ault, C.R.; Novak, J.D. and Gowin, D.B., 1988, Constructing Vee Maps for Clinical Interviews on Energy Concepts. Science Education, **72**,4, 515-545.

Ausubel, D., 1968, Educational Psychology: A Cognitive View, (New York, NY: Werbel and Peck).

Ausubel, D.; Novak, J.D. and Hanesian, H., 1978, Educational Psychology: A Cognitive View, 2nd Edition (New York, NY: Holt, Rinehart and Winston).

Baker, E.L.; Niemi, D.; Novak, J. and Herl, H., 1991, Hypertext as a strategy for teaching and assessing knowledge representation. Paper presented at NATO Advanced Research Workshop on Instructional Design Models for Computer-Based Learning Environments, Enschede, The Netherlands.

Baldissera, J.A., 1993, Misconceptions of Revolution in History Textbooks and Their Effects on Meaningful Learning. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

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

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

Bogdan, R. and Biklen, S., 1992, Qualitative research for education: An introduction to theory and methods, (Nedham Heights, MA: Allyn & Bacon).

Brody, M., 1993, Student Misconceptions of Ecology: Identification, Analysis and Instructional Design. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Bruer, J.T., 1993, Schools for thought, (Massachusetts: The MIT Press).

Champagne, A.B.; Klopfer, L.E.; DeSena, A.T. and Squires, D.D., 1978, Content structure in science instructional materials and knowledge structure in students' memories (Report No. LRD-1878/22). ERIC Document Reproduction Service No. ED 182 143, Learning Research and Development Center, University of Pittsburgh.

Coleman, E.B., 1998, Using Explanatory Knowledge During Collaborative Problem Solving. The Journal of the Learning Sciences, **7**,3&4, 387-427.

Crowl, T.K., 1996, Fundamentals of Educational Research, 2nd Edition (New York: Brown & Benchmark Publishers).

Davenport, J.C., 1998, Sanctuary Explorations: An access guide to the Monterey Bay National Marine Sanctuary, (Berkeley, California: UC Printing).

De Groot, S.S., 1993, Concept Mapping with computer support, laser disc and graphics applied to Microbiology. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Edmondson, K., 1995, Concept Mapping for the Development of Medical Curricula. Journal of Research in Science Teaching, **32**,7, 777-793.

Escalada, L.T. and Zollman, D.A., 1997, An investigation of the effects of using interactive digital video in a physics classroom on student learning and attitudes. Journal of Research in Science Teaching, **34**, 467-489.

Eschenbrenner, M. (1994). Concept Mapping in the Primary Grades. Unpublished Master of Science. Education. California State University, Fullerton.

Farrokh, K. and Krause, G., 1996, The Relationship of Concept-Mapping and Course Grade in Cell Biology. Meaningful Learning Forum, **1**.

Fleiss, J., 1981, Statistical Methods for Rates and Proportions, 2nd Edition (New York: John Wiley and Sons).

Foundation, P.S.R., 1998, Pelagic Shark Research Foundation web site (www.pelagic.org).

Gangosa, Z., 1996, Meaningful Learning Based Instructional Design. Meaningful Learning Forum, **1**.

Garrison, T., 1993, Oceanography, (California: Wadsworth, Inc.).

González, F.M., 1993, Diagnosis of Alternative Conceptions in Science in Spanish Primary School Students. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Heinze-Fry, J.A., 1997, Concept Mapping: Weaving Conceptual Connections. Paper presented at Weaving Connections: Cultures and Environments - Environmental Education and Peoples of the World, Vancouver, British Columbia, Canada, (North American Association for Environmental Education).

Heinze-Fry, J.A. and Novak, J.D., 1990, Concept Mapping Brings Long-Term Movement toward Meaning Learning. Science Education, **74**,4, 461-472.

Hewson, M.G. and Hamlyn, D., Date Unknown, The influence of intellectual environment on conceptions of heat. ERIC Document Reproduction Service No. ED 231 655, National Institute of Personnel Research.

Horton, P.B.; McConney, A.A.; Gallo, M.; Woods, A.L.; Senn, G.J. and Hamelin, D., 1993, An Investigation of the Effectiveness of Concept Mapping as an Instructional Tool. Science Education, **77**,1, 95-111.

Hoz, R.; Bowman, D. and Chacham, T., 1997, Psychometric and Edumetric Validity of Dimensions of Geomorphological Knowledge Which Are Tapped by Concept Mapping. Journal of Research in Science Teaching, **34**,9, 925-947.

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

Jegede, O.J.; Alaiyemola, F.F. and Okebukola, P.A., 1990, The Effect of Concept Mapping on Students' Anxiety and Achievement in Biology. Journal of Research in Science Teaching, **27**,10, 951-960.

Khan, K.M., 1993, Concept Mapping as a Strategy for Teaching and Developing the Caribbean Examinations Council (CXC) Mathematics Curriculum in a Secondary School. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Kirschner, P. and Huisman, W., 1998, 'Dry laboratories' in science education; computer-based practical work. International Journal of Science Education, **20**,6, 665-682.

Lay-Dopyera, M. and Beyerbach, B., 1983, Concept mapping for individual assessment. ERIC Document Reproduction Service No. ED 229 399, School of Education, Syracuse University.

Leahy, R., 1989, Concept Mapping: Developing Guides to Literature. College Teaching, **37**,2, 62-69.

Lee, J.J., 1999, The Impact of Korean Language Accomodations on Concept Mapping Tasks for Korean American English Language Learners. Paper presented at American Educational Research Association Annual Meeting, Montreal, Canada, (National Center for Research on Evaluation, Standards, and Student Testing (CRESST, UCLA)).

Liu, X. and Hinchey, M., 1993, The Validity and Reliability of Concept Mapping as an Alternative Science Assessment. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Liu, X. and Hinchey, M., 1996, The internal consistency of a concept mapping scoring scheme and its effect on prediction validity. International Journal of Science Education, **18**,8, 921-937.

Lomask, M.; Baron, J.B.; Greig, J. and Harrison, C., 1992, ConnMap: Connecticut's use of concept mapping to assess the structure of students' knowledge of science. Paper presented at Annual Meeting of the National Association of Research in Science Teaching, Cambridge, MA.

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

Markham, K.M.; Mintzes, J.J. and Jones, M.G., 1994, The Concept Map as a Research and Evaluation Tool: Further Evidence of Validity. Journal of Research in Science Teaching, **31**, 91-101.

Markham, K.M.M., Joel J. & Jones, M. Gail, 1993, The Structure and Use of Biological Knowledge about Mammals in Novice and Experienced Students. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Markow, P.G. and Lonning, R.A., 1998, Usefulness of Concept Maps in College Chemistry Laboratories: Students' Preceptions and Effects on Achievement. Journal of Research in Science Teaching, **35**,9, 1015-1029.

Mason, C.L., 1992, Concept Mapping: A Tool to Develop Reflective Science Instruction. Science Education, **76**,1, 51-63.

MBARI, 1998, Virtual Canyon Project Web Site (www.virtual-canyon.org).

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

Moore, D.S. and McCabe, G.P., 1989, Introduction to the Practice of Statistics, (New York: W.H. Freeman and Co.).

Moreira, M.A. and Greca, I., 1996, Concept Mapping and Mental Models. Meaningful Learning Forum, **1**.

Moreira, M.A. and Motta, A.M.B., 1993, Concept Mapping in 7th Grade Mathematics: An Exploratory Study. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Moreira, M.M., 1996, The Use of Concept Maps in an EFL Classroom. Meaningful Learning Forum, **1**.

Nakhleh, M.B. and Krajcik, J.S., 1991, The effect of level of information as presented by different technology on students' understanding of acid, base, and pH concepts. Paper presented at National Association for the Research in Science Teaching, Lake Geneva, WI, (ERIC Document Reproduction Service).

Novak, J.D., 1998, Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations, (London: Lawrence Erlbaum Associates).

Novak, J.D. and Gowin, D.B., 1984, Learning how to Learn, (Cambridge, England: Cambridge University Press).

Novak, J.D.; Gowin, D.B. and Johansen, G.T., 1983, The use of concept mapping and knowledge vee mapping with junior high school science students. Science Education, **67**,5, 625-645.

NSTA, 1990, The content core: Guide for curriculum designers. National Science Teachers Association.

Osmundson, E.; Chung, G.K.W.K.; Herl, H.E. and Klein, D.C.D., 1999, Concept Mapping in the Classroom: A tool for examining the development of students' conceptual understandings. Paper presented at American Educational Research Association Annual Meeting, Montreal, Canada, (National Center for Research on Evaluation, Standards, and Student Testing (CRESST, UCLA)).

Patton, M.Q., 1990, Qualitative Evaluation and Research Methods, 2nd Edition (London: Sage Publications).

Pearsall, N.R.; Skipper, J.E. and Mintzes, J.J., 1997, Knowledge Restructuring in the Life Sciences: A Longitudinal Study of Conceptual Change in Biology. Science Education, **81**,2, 193-215.

Press, F. and Siever, R., 1986, Earth, (New York: W.H. Freeman and Co.).

Rafferty, C.D. and Fleshner, L.K., 1993, Concept Mapping: A Viable Alternative to Objective and Essay Exams. Reading Research and Instruction, **32**,2, 25-34.

Rakes, G.C., 1996, Using the internet as a tool in a resource-based learning environment. Educational Technology,Sept-Oct 96, 52-56.

Rice, D.C.; Ryan, J.M. and Samson, S.M., 1998, Using Concept Maps to Assess Student Learning in the Science Classroom: Must different methods compete? Journal of Research in Science Teaching, **35**,10, 1103-1127.

Rivest, B., 1996, Personal Communication: Marine Biology Assessments.

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

Roth, W.-M. and 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**,5, 503-534.

Ruiz-Primo, M.A.; Schultz, S.E. and Shavelson, R.J., 1997, Concept Map-Based Assessment in Science: Two Exploratory Studies. CSE Technical Report 436, National Center for Research on Evaluation, Standards, and Student Testing.

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

Ruiz-Primo, M.A. and Shavelson, R.J., 1996b, Rhetoric and Reality in Science Performance Assessments: An Update. Journal of Research in Science Teaching, **33**,10, 1045-1063.

Savery, J. and Duffy, T., 1995, Problem-based learning: An instructional model and its constructivist framework. Educational Technology,Sept-Oct 95, 31-38.

Schreiber, D.A. and Abegg, G.L., 1991, Scoring student-generated concept maps in introductory college chemistry. Paper presented at Annual Meeting of the National Association for Research in Science Teaching, Lake Geneva, WI.

Sizmur, S. and Osborne, J., 1997, Learning processes and collaborative concept mapping. International Journal of Science Education, **19**,10, 1117-1135.

Songer, C.J. and Mintzes, J.J., 1993, Understanding Cellular Respiration. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

Stensvold, M.S. and Wilson, J.T., 1990, The Interaction of Verbal Ability with Concept Mapping in Learning from a Chemistry Laboratory Activity. Science Education, **74**,4, 473-480.

Stewart, J.; Van Kirk, J. and Rowell, R., 1979, Concept Maps: A tool for use in biology teaching. The American Biology Teacher, **41**,3, 171-175.

Strauss, A. and Corbin, J., 1990, Basics of Qualitative Research: Grounded Theory Procedures and Techniques, (London: Sage Publications).

Vermeylen, F., 1998, StatNews #31: Assessing Agreement Between Raters. Office of Statistical Consulting, Cornell University.

Vosniadou, S., 1996, Learning environments for representational growth and cognitive flexibility. In Vosniadou, S.; DeCorte, E.; Glaser, R. & Mandel, H., International perspectives on the design of technology-supported learning environments (New Jersey: Lawrence Erlbaum Associates), 129-148.

Wallace, J.D. and 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.

Wilson, J.M., 1993, The Predictive Validity Of Concept-Mapping: Relationships To Measures Of Achievement. Paper presented at Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Ithaca, NY.

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

Zeilik, M.; Schau, C.; Mattern, N.; Hall, S.; Teague, K. and Bisard, W., 1997, Conceptual astronomy: A novel model for teaching postsecondary science courses. American Journal of Physics, **65**,10, 987-996.

Appendix A: Sample concept map activities

Virtual Canyon Concept Map Activity (High School, pre)

Your Name: _____
Your Virtual Canyon Teacher's Name: _____

Some general concepts to get you started

1} Marine Science
2} Monterey Bay Submarine Canyon

The topic you are planning on studying

3} _____

Please list three important concepts which relate to your topic

4} _____ 5} _____ 6} _____

Please list four more concepts that you feel are important for other students to know about 'marine science' in general.

7} _____ 9} _____
8} _____ 10} _____

Don't forget that your concept map can include any other concepts that you need to explain the above concepts.

Now please write your concepts on stickies and draw a concept map.

Virtual Canyon Concept Map Activity (Elementary, post)

Your name: _____
Your teacher's name: _____

One line 1 below please write the topic you studied about on the Virtual Canyon.

1. _____

On lines 2, 3 and 4 below please list three things you know about this topic:

2. _____ 3. _____ 4. _____

You want to tell your friend about the ocean. Think of three things that you think would be important to tell them.

1. _____ 2. _____ 3. _____

Appendix B: Concept Map Assessment Rubric Final Version

Final Version (Stage 1): Identifying Scientific Vocabulary
Instructions: Examine all of the assigned concept maps. Write down all of the scientific vocabulary terms you find in the maps. Terms can be found in any part of the maps. List a given term only once. This stage should include ALL scientific vocabulary, including common, specific, and Latin species names, as well as terms which are recognizably abstract, specialized, or sophisticated. This vocabulary includes, but is not limited to, Common Species names for a group of species, Common Technologies, Common parts of species or technologies, Science Occupations and Fields, Common functions or abstract ideas. Also included are Common Species names for a specific species, Latin species names, Common names for categories of species (e.g. 'Mammals'), and Other specialized scientific language. After you have identified the scientific vocabulary in the maps, bring your list to the Reviewers Meeting. At this meeting, we will decide which terms count for 'Scientific Language' in Stage 2.

Final Version (Stage 2): Scoring of Each Concept Map		
Variable	**Response Options**	**Usage guidelines**
Map Data		
Reviewer	Open cell	Enter the Reviewer ID you were assigned
School	Open cell	Enter School ID as marked on the map
Class	Open cell	Enter Class ID as marked on the map
Student	Open cell	Enter Student ID as marked on the map
Map	Open cell	Enter activity ID as marked on the map
Vocabulary	Large open cell	Does the map include scientific language? • Only list words in the agreed upon list. Remember to separate each term with a comma, and only list each term once. Do not use trailing spaces.
Supra-Proposition Explanation Typologies		
Classification included	No Yes	
Awareness of Impact		
Human Impact included	No Yes	'Human Impact' covers both impact by humans on other spheres, as well as impacts of other spheres on humans.
Geosphere / Climate Impact included	No Yes	'Geosphere / Climate Impact' covers any large scale system interactions, especially those which cross system boundaries, such as Climate-Biosphere.
Overall Impression		
Overall Impression	Open cell	What is the reviewer's overall impression of the understanding demonstrated by the map? • Enter a number from 1 to 10, with 10 being best.

Proposition Data		
Proposition Number	Open cell	On your copy of the concept map, hilight each proposition. Try to start from the main idea and work down the map. Mark each proposition with an ID number at both ends of the proposition. Err on the side of marking compound propositions. In other words, don't score dependent propositions separately. Technically, any proposition with two boxes and a linking word is a simple proposition, but if the linking word appears to contain a concept, you can use your judgment to mark it as compound. On the other hand, don't combine propositions that could be scored separately.
Accurate	Don't Know	Reviewer has insufficient knowledge to judge the proposition's accuracy. • Only use this option if you are very uncertain about the accuracy of the proposition. If you are pretty sure of the accuracy of the proposition, be decisive and make a judgment.
	Affect	Statement of emotion
	No	Inaccurate
	AccCom	Commonly Accurate = (e.g. 'Whales are big') • Common Animal / Plant / Common Group of Animals lives in Ocean / Water • El Nino affects people generally, Vague statements about El Nino • General Human Activity,not specifically referring to scientific process (trawling / fishing / eating) • Ocean has Water / Salty Water / Sand • Something gets Washed Up On Beach / Found on Shore • Common Animal is a Sea Creature • Common Animal / Object is a Vague Size / Measurement • Unclear Proposition because of unspecified sequence (e.g. 'Temperature Decreases in Midwater')
	AccSci	Scientifically Accurate = Categorical Observations • Vague, but gets at important idea • Common Animal uses ocean as habitat, scientific terms or more specific process than 'lives' • Common Animal eats Fish / Meat • Common Animal has Common Animal Part • Common Animal is a Color • Marine Science studies Ocean / Common Animal • Common Animal lives in Somewhat Specific Place (Monterey Bay, Pacific Ocean) • Pollution affects Blank • Pollution gets in Blank • Uncommon Animal / Object is Vague Size / Measurement • Seals lay on Rocks • Sharks attack Humans • Something is a food to Many Animals
	Q	Proposition is a Question
	MNS	Proposition Makes No Sense
Explanation	What	
	How	
	Why	
Proposition Structure	Simple	
	Compound	
Final Version (Stage 3): Consultation with Subject Matter Expert		
Instructions: Create a list of all propositions which were scored 'Don't Know' in the Accuracy variable. Bring these propositions to the reviewers meeting. Any propositions which can not be scored by consensus at the reviewers meeting should be sent out to a subject matter expert for suggested scoring. At this point the concept map data will be ready for statistical analysis.		

Final Version (Stage 2) - Data Entry form

Map Data Version 10

Reviewer | School | Class | Student | Map

Vocabulary: Does the map include Scientific Language? (1)

1) Only list words in the agreed upon list. Remember to separate each term with a Comma, and only list each term once. Do not use trailing spaces.

Supra-Proposition Explanation Typologies

No Yes Classification included

Awareness of Impact

No Yes Human Impact included (3)

No Yes Geosphere/Climate Impact included

3) "Human Impact" covers both impact by humans on other spheres, as well as impacts of other spheres on humans. • "Geosphere/Climate Impact" covers any large scale system interactions, especially those which cross system boundaries, such as Climate-Biosphere.

Overall Impression

What is the reviewer's overall impression of the understanding demonstrated by the map? (Enter a number from 1 to 10, with 10 being best.)

Proposition Data

Proposition Number (4)

4) On your copy of the concept map, hilight each proposition. Try to start from the main idea and work down the map. Mark each proposition with an ID number at both ends of the proposition. Err on the side of marking compound propositions. In other words, don't score dependent propositions separately. Technically, any proposition with two boxes and a linking word is a simple proposition, but if the linking word appears to contain a concept, you can use your judgement to mark it as compound. On the other hand, don't combine propositions that could be scored separately.

DontKnow Affect No AccCom AccSci — Accurate (2)

Q MNS

What How Why — Explanation

Simple Compound — Proposition Structure

2) DontKnow = Reviewer has insufficient knowledge to judge the proposition's accuracy, Affect = statement of emotion, No = Inaccurate, AccCom = Commonly Accurate (e.g. "Whales are big"),AccSci = Scientifically accurate, Q = Proposition is a Question, MNS = Proposition Makes No Sense.

Usage guidelines:

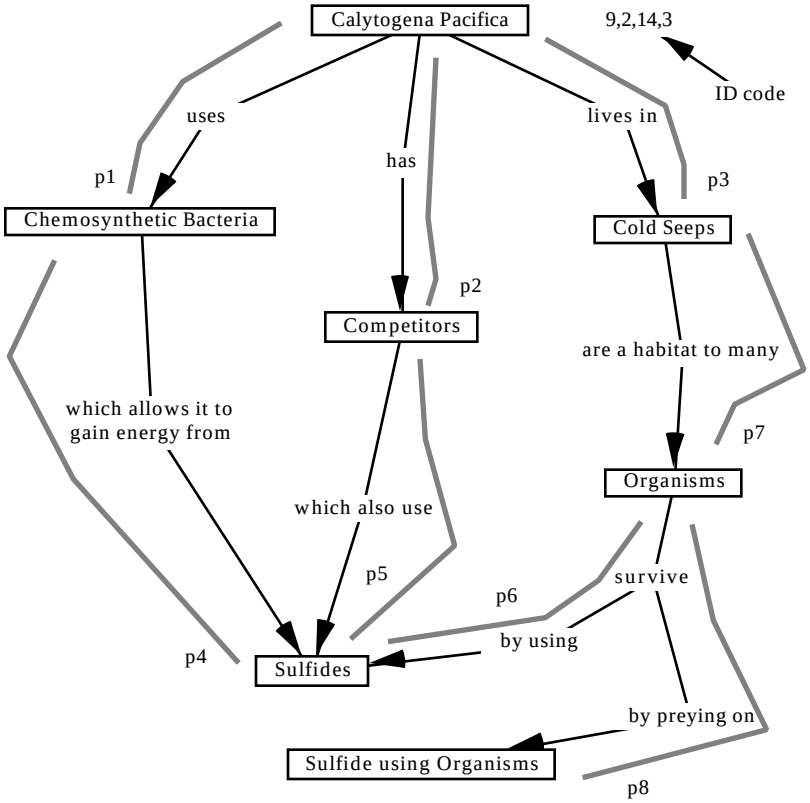
DontKnow = Only use this option if you are very uncertain about the accuracy of the proposition. If you are pretty sure of the accuracy of the proposition, be decisive and make a judgement.

The following guidelines help to distinguish between Commonly Accurate and Scientifcally Accurate.

AccSci = Categorical Observations • Vague, but gets at important idea • Common Animal uses ocean as habitat, sci terms or more specific process than "lives" • Common Animal eats Fish/Meat • Common Animal has Common Animal Part • Common Animal is a Color • Marine Science studies Ocean/Common Animal • Common Animal lives in Somewhat Specific Place (MB,Pacific Ocean) • Pollution affects Blank • Pollution gets in Blank • Uncommon Animal/Object is Vague Size/Measurement • Seals lay on Rocks • Sharks attack Humans • Something is a food to Many Animals

AccCom = Common Animal/Plant/Common Group of Animals lives in Ocean/Water • El Nino affects people generally, Vague statements about El Nino • General Human Activity,not specifically referring to scientific process (trawling/fishing/eating) • Ocean has Water/Salty Water/Sand • Really Unclear Proposition • Something gets Washed Up On Beach/Found on Shore • Common Animal is a Sea Creature • Common Animal/Object is a Vague Size/Measurement • Unclear Proposition because of unspecified sequence (e.g. "Temperature Decreases in Midwater")

Appendix C: An example concept map showing marking for scoring



Headings

Reviewer	School	Class	Student	Map	Vocab	Classif ication	Human	Geo	Over -all
Prop#	Accurate	Explain	Prop Struct						
EOM (this is an end of map marker used by the program which transforms the data to a flat structure)									

Sample Data

1	9	2	14	3	Cold seeps, ... *	No	No	Yes	6
1	AccSci	What	Simple						
2	AccSci	What	Simple						
3	AccSci	What	Simple						
4	AccSci	How	Com-pound						
5	AccSci	What	Simple						
6	AccSci	How	Simple						
7	AccSci	What	Com-pound						
8	AccSci	How	Simple						
EOM									

* organisms, sulfides, chemosynthetic bacteria, energy, competitors, Calytogena Pacifica, habitat, preying

A concept map was transcribed from the original hand drawn form using LifeMap. The gray lines show how propositions would be marked. If all concept maps can be transcribed into the computer, some kinds of time consuming data analysis are made much easier later on, but this marking process can be done by hand if needed. Map level variables, including the ID code, are recorded on the score sheet, followed by the proposition level variables. All variables should be entered as numeric codes to facilitate statistical analysis, but they are shown here written out for simplicity. The proposition numbers recorded on the map are in an arbitrary order, but they are important in case scores have to be rechecked at a later date. The authors have provided a sample score record for a high school map as entered into a spreadsheet.

End notes

(1) Map level variables are Review ID (consisting of Reviewer, School, Class, Student, and Map (training, pre, post)), Scientific Vocabulary, Classification, Human Impact, Geosphere / Climate Impact, and Overall Impression. These variables, other than Overall Impression, are intended to detect knowledge structures which can not be expressed in a single proposition. Classification records the presence or absence of classic classification structures. These are often biological in nature, but similar structures describing other kinds of objects would qualify (such as a classification of Remote Operated Vehicles). Human Impact and Geosphere / Climate Impact record the presence or absence of knowledge structures which discuss large scale system interactions. In the Virtual Canyon study, Human Impact, Geosphere / Climate Impact, and Classification did not appear in enough maps to be useful variables.

Overall Impression is an intentionally undefined variable. It is included primarily to confirm the need for detailed scoring rubrics: inter-rater agreement for Overall Impression in the agreement test of the penultimate version of the rubric was 7% when calculated using the full ten point scale, and between 43% and 57% when calculated using a three point scale.

Acknowledgments

This research was funded in part by grant # RED-9554325 under the National Science Foundation Networking Infrastructure for Education Program. The authors would like to thank the participating teachers and students, without whom this work would not be possible, Susan Lasky for her help in developing the training protocol used in this study, and Richard Shavelson and Maria Ruiz-Primo for generously sharing their experience with concept maps as assessment.