ED 422 403 TM 028 965

AUTHOR	Ruiz-Primo, Maria Araceli; Shavelson, Richard J.
TITLE	Concept-Map Based Assessment: On Possible Sources of Sampling Variability.
INSTITUTION	Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC. and Student Testing, Los Angeles, CA.
PUB DATE	1997-03-29
NOTE	52p.
CONTRACT	R305B60002
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Chemistry; *Concept Mapping; Educational Assessment; *High School Students; High Schools; *Knowledge Level; *Scoring; Student Evaluation; Tables (Data); *Test Construction

ABSTRACT

A concept-map assessment consists of a task that elicits structured knowledge, a response format, and a scoring system. Variation in tasks, response formats, and scoring systems produce different mapping techniques that may elicit different knowledge representations, posing construct-interpretations challenges. This study examined two mapping techniques: (1) students generated the 10 concepts from chemistry to construct a map; and (2) the assessor provided a list of 10 concepts. Two concept-lists were randomly sampled from the domain to examine the effect of concept sampling on map scores. Forty high school students, two teachers, and one expert participated. Results indicate that: (1) the two mapping techniques were statistically equivalent; (2) students' concept-map scores generalized across samples of concepts; (3) concept maps could be reliably scored, even though they involved complex judgments; and (4) multiple-choice test and concept maps measure somewhat different aspects of science knowledge. Appendixes contain a discussion of compiling the list of concepts, a list of concepts considered for the three conditions, a sample of instructions, and a matrix of the relations between pairs of concepts. (Contains 9 tables and 32 references.) (Author/SLD)

Running Head: CONCEPT-MAP BASED ASSESSMENT

Concept-Map Based Assessment: On Possible Sources of Sampling Variability

Maria Araceli Ruiz-Primo and Richard J. Shavelson

Stanford University

Draft

March 29, 1997

* The work reported herein was supported by the Center for Research on Evaluation, Standards, and Student Testing under the Educational Research and Development Centers Program PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report reflect those of the authors ant not necessarily those of the funding agencies. The authors are deeply grateful to Susan Elise Shultz, for her help in collecting data and scoring students' maps, and Dr. Pinchas Tamir, for his valuable comments.

Abstract

A concept-map assessment consists of a _task_ that elicits structured knowledge, a _response format_, and a _scoring system_. Variation in tasks, response formats, and scoring systems produce different mapping techniques that may elicit different knowledge representations, posing construct-interpretations challenges. This study examined two mapping techniques: (1) students generated the 10 concepts from chemistry to construct a map, and (2) assessor provided a list of 10 concepts. Two concept-lists were randomly sampled from the domain to examine the effect of concept sampling on map scores. Forty high-school students, two teachers and one expert participated. Results indicated that: (a) the two mapping techniques were statistically equivalent; (b) students' concept-map scores generalized across samples of concepts; (c) concept maps could be reliably scored, even though they involved complex judgments; and (d) multiple-choice test and concept map measure somewhat different aspects of science knowledge.

Concept-Map Based Assessment: On Possible Sources of Sampling Variability

The search for "authentic" science assessments that reflect what students know and can do is well underway. As part of this search, educators and researchers are looking for more or less direct measures of students' knowledge structures. Concept maps--structural representations of key concepts in a subject domain--have been claimed to be a potential "find."

The rationale behind this claim is that knowledge has an organizational property that can be captured with structural representations (e.g., Goldsmith, Johnson, & Acton, 1991; Jonassen, Beissner, & Yacci, 1993; White & Gunstone, 1992). Cognitive psychologists posit that "the essence of knowledge is structure" (Anderson, 1984, p. 5). Concept interrelatedness, then, is an essential property of knowledge. Indeed, one aspect of competence in a domain is well structured knowledge; as expertise in a domain grows, through learning, training, and/or experience, the elements of knowledge become increasingly interconnected (e.g., Glaser & Bassok, 1989; Shavelson, 1972).

Assuming that knowledge within a content domain is organized around central concepts, to be knowledgeable in the domain implies a highly integrated conceptual structure. Concept maps, then, may capture important aspects of this interrelatedness between concepts.

Formally, a concept map is a graph consisting of nodes and labeled lines. The nodes correspond to important terms (standing for <u>concepts</u>) in a domain.[1] The lines denote a relation between a pair of concepts (nodes). And the label on the line tells how the two concepts are related. The combination of two nodes and a labeled line is called a <u>proposition</u>. A proposition is the basic unit of meaning in a concept map and the smallest unit that can be used to judge the validity of the relationship drawn between two concepts (e.g., Dochy, 1996). Concept maps, then, purport to represent some important aspect of a student's declarative knowledge in a content domain (e.g., chemistry).

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

Concept maps, as assessments, can be thought of as a set of procedures used to measure important aspects of the structure of a student's declarative knowledge. We use the term "assessment" to reflect our belief that reaching a judgment about an individual's knowledge and skills requires an integration of several pieces of information; we consider concepts maps as potentially one of those pieces (see Cronbach, 1990).

Ideally, before concept maps are used in classrooms or for large-scale assessment, and before concept-map scores are reported to teachers, students, the public, and policy-makers, research needs to provide information about the technical properties of concept maps for representing knowledge structure. Accordingly, this study provides evidence bearing on the reliability and validity of concepts maps as representations of students' knowledge structures. More specifically, this study explores the sampling variability of two concept mapping techniques for use in large-scale science assessment. We examine whether map scores are sensitive to who samples the concepts to be used in the map (student or tester) and to the variability that arises from one random sample of concepts from a domain to the next.

Concept Map-Based Assessment

Intuitively, the use of concept maps to evaluate students' declarative knowledge structure is appealing. A student's map construction directly reflects, to some degree, her or his understanding in a domain. Nevertheless, before adopting maps for assessment use, more needs to known about them. A common understanding is needed as to what a concept map assessment is and whether it provides a reliable and valid measure of students' cognitive structure (Ruiz-Primo & Shavelson, 1996).

A Concept-Map Based Measurement Framework

A concept map measure can be characterized by: (a) a task that invites a student to provide evidence bearing on his or her knowledge structure in a domain, (b) a format for the student's response, and (c) a scoring system by which the student's concept map can be substantively evaluated accurately and consistently. Without these three components, a concept map cannot be considered as a measurement tool (Ruiz-Primo & Shavelson, 1996).

Based on this conceptualization, we found tremendous variation in what counted as concept mapping techniques. This variation emerged from tasks, response formats, and scoring systems (Ruiz-Primo & Shavelson, 1996), and is captured in Table 1.

Insert Table 1 About Here

Concept mapping tasks varied in three ways: (a) demands made on the students in generating concept maps (tasks demands), (b) constraints placed on the task (task constraints), and (c) the intersection of task demands and constraints with the structure of the subject domain to be mapped (content structure). As an example of the third category consider the constraint of building a hierarchical map in a subject domain that is not hierarchical. Methodologically and conceptually, there is no need to impose a hierarchical structure. If the content structure is hierarchical, and the student has mastered the domain, a hierarchical map should be observed.

Response formats vary in three ways: (a) whether the student's response is given with paper-and-pencil, orally, or on a computer (response mode); (b) the link between task and format (e.g., if the task asks the student to fill in a skeleton map, the response format provides the skeleton map; response format); and (c) who draws the map (i.e., most frequently the student; however, teachers or researchers can draw maps from student interviews or essays; mapper).

Three general scoring strategies have been used with maps: (a) score the components of the students' maps (e.g., number of links); (b) compare the students' map with a criterion map (e.g., a map constructed by an expert); and (c) a combination of both strategies.

If each of the 6 task demands (e.g., fill-in-the-blank nodes on a map) is combined with each of the 8 types of task constraints (e.g., hierarchical vs. non-hierarchical), there are no less than 1530 (i.e., 6 x 2^8 - 1) different ways to produce a concept mapping task! Of course not all combinations may be realistic. Regardless, the wide variety of potential maps raises issues about what is being measured. Some examples may help to make clear the problem of the variation in concept mapping techniques. Table 2 presents five examples of different types of tasks, response formats, and scoring systems used in practice and in research on concept maps (see Ruiz-Primo & Shavelson, 1996, for more examples).

Insert Table 2 About Here

We suspect that different mapping techniques may tap different aspects of cognitive structure and lead students to produce different concept maps. Nevertheless, current practice holds that all variations in mapping techniques are interpreted the same way, as representing a student's knowledge structure--the relationship of concepts in a student's memory (Shavelson, 1972). If concept maps are to be used as a measurement tool, we must take the time and effort to provide evidence on the impact of different mapping techniques on representation of a student's knowledge structure.

Unfortunately, cognitive theory does not provide an adequate basis for deciding which techniques to prefer because many of the techniques have no direct connection with one or another theory. Furthermore, current cognitive theories may be limited in their ability to guide mapping techniques because they tend to be middle-range theories focused on particular aspects of cognition. Application of cognitive theory, along with empirical research, should, over time, provide guidelines that would narrow the number of possible techniques to a manageable set.

In the meantime, research on concept maps can proceed by developing criteria that help to discard some techniques. In the study reported here, we applied the following criteria to narrow down alternatives: (a) appropriateness of the cognitive demands required by the task; (b) appropriateness of a structural representation in a content domain; (c) appropriateness of the scoring system used to evaluate the accuracy of the representation; and (d) practicality of the technique. We eliminated, for example, a fill-in-the-blank task because we regarded it as inappropriate for measuring students' knowledge structures since the task itself too severely restricts students' representations. We also think that imposing a hierarchical structure, regardless of content domain, is inadequate since an accurate concept-map representation of a hierarchical domain should be hierarchical itself. Furthermore, not all subject-matter domains have a hierarchical structure (e.g., Shavelson, 1972). We also favored scoring criteria that focus on the adequacy of propositions and eliminated just counting the number of map components (i.e., nodes and links). Finally, since our focus is on large-scale assessment, we eliminated mapping techniques that require one-to-one interaction between student and tester on practical grounds.

Technical Properties of Concept Maps

As mentioned previously, concept maps have been used more for instruction than for formal assessment. The direct consequence is that reliability and validity issues associated with knowledge structure interpretations of concept maps have largely been ignored to date. Here we highlight several important technical issues in the interpretation of concept maps.

Reliability. Reliability refers to the consistency of concept map scores from one sample (of concepts, time, etc.) to another. Some questions that should be raised about the reliability of concept map scores are: Can raters reliably score concept maps? Are students' scores stable across short occasions? Are scores sensitive to the concepts (sampled) from a domain?

Little attention has been paid to reliability issues in the literature. Available research suggests high interrater reliability and agreement (e.g., Barenholz & Tamir, 1992; Lay-Dopyera & Beyerbach, 1983; Lomask et al., 1992; Nakhleh & Krajcik, 1991). However, these findings should be interpreted cautiously. In some studies the scoring system involved only counting the number of certain map components (e.g., number of nodes; Lay-Dopyera & Beyerbach, 1983). In others, raters scored only a small sample of concept maps (e.g., Nakhleh & Krajcik, 1991). It may be that reliability depends on the scoring criteria (e.g., counting nodes vs. judging the accuracy of propositions) and the number of concept maps scored.

We found only one study that reported retest reliability (stability) (Ruiz-Primo & Shavelson, 1996). Lay-Dopyera and Beyerbach (1983) concluded that their study failed to establish concept maps as stable measures (stability coefficients ranged from .21 to .73 for different map component scores).

No study has examined the issue of concept-sampling variability. Hence one purpose of this study was to do so. We randomly sampled concepts from a subject domain (Sample "A" and Sample "B") to examine concept sampling variability of map scores.

Validity. Beyond reliability it is essential to justify an interpretation of a concept map scores as a measure of some aspect of a student's knowledge structure in a knowledge domain--that is, to demonstrate the validity of proposed "cognitive structure" interpretations of concepts map scores. One set of evidence for constructing validity bears on the content represented in the map; evidence of content relevance and representativeness. One source of evidence is expert judgment. Only a few studies reported that "experts" judged the terms and maps as consistent with the subject domain (e.g., Anderson & Huang, 1989, Barenholz & Tamir, 1992, Lomask et al., 1992; Nakhleh & Krajcik, 1991).

Other validity evidence is correlational (e.g., concurrent evidence, and convergent and discriminant evidence). Some studies have shown consistent correlations between concept map scores and measures of student achievement (concurrent evidence; e.g., Anderson & Huang, 1989), while others have suggested that concept map scores seem to measure a different aspect of achievement than that measured by multiple-choice tests (e.g., McClure & Bell, 1990; Novak, Gowin, and Johansen, 1983).

In this study we examined two issues: the sampling variability of scores from two alternative mapping techniques (convergent validity) and the sensitivity of concept map scores to the sampling of concept terms. We varied the mapping technique used: One technique asked students to provide the concepts in a domain with which to construct the map; the other technique provided a set of 10 concepts. When concepts were provided by the assessor, we sampled two sets of concept terms.

More specifically, this study addressed the following questions: Do different mapping techniques provide the same information about a student's knowledge structure? If concepts are provided by the assessor, are map score sensitive to the sampling of concept terms? How reliable are map scores across raters? Do concept maps provide sensible representations of knowledge in a domain as judged by subject-matter experts? Do different assessment technique scores correlate differently with traditional multiple-choice test scores?

Method

Subjects

This study involved two classes of high school chemistry students taught by the same teacher (with 4 years of teaching experience), a second chemistry teacher (with 7 years of teaching experience), and a chemist (expert, with 10 years of experience in research on water quality). All subjects were drawn from the Palo Alto area. The students, the teachers and the expert were trained in the same way to construct concept maps.

Of the original 47 students in the two groups, 4 were dropped because of incomplete data. Another three were randomly dropped in order to equalize cell sizes (see design below) and simplify interpretations. As a result, data were analyzed for 40 students who were assessed on each of three occasions. (Analysis with all 43 students replicated the analysis reported here).

Design

The two classes (groups) were randomly assigned to one of two sequences of concept samples: Sequence 1--Sample A first followed by Sample B (class 1), and Sequence 2--Sample B first followed by Sample A (class 2). Each group of students was tested on three occasions: (1) on the first occasion, students were asked to construct a map with terms they provided; (2) on the second occasion, students were asked to construct a map with the first list of assessor-provided concepts; (3) on the third occasion, students were asked to construct a map with the second list of assessor-provided concepts. The 2 x 3 mixed design had one between-subjects factor, sequence of terms samples, and one within-subjects factor, occasion.

Subject Domain and Material

The topic, "Reactions and Interactions," was selected from the knowledge domain defined by the notion of "big ideas" in physical science contained in the Science Framework for California Public Schools (California Department of Education, 1990). Reactions and interactions involve the study of chemical reactions. This big idea focuses on two issues: "What happens when substances change? What controls how substances change?" (p.49). At the high school level, these issues involve, among other topics, understanding atomic structure and the nature of ions, molecules and compounds. The latter was the topic selected for this study.

The concepts of ions, molecules and compounds were addressed in the unit, "Chemical Names and Formulas," in the chemistry curriculum of the high school where the study was carried out. As with the other units in the curriculum, this unit was taught from the widely used text, Chemistry (Wilbraham, Staley, Simpson, & Matta, 1990). The chapter, "Chemical Names and Formulas," defined the domain for sampling concepts to be used in the study.

We compiled a list of 20 key concepts in two ways: (a) asking the chemistry teachers to provide the concepts they thought were most important in the unit; and (b) reviewing the textbook used in class ourselves. The process followed in compiling the sampling of concepts is described in Appendix A.

Two sub-lists of concepts were created from this list. Both sub-lists contained four control concepts in common (i.e., ions, molecules, compounds, and electrons). Two samples of six concepts each were randomly selected from the other 16 concepts on the list to form the two sets of concepts (see Appendix B).

Instrumentation

Here we describe the concept-mapping techniques (tasks, response format, and scoring system) as well as the multiple-choice achievement test.

Concept map tasks. The two mapping techniques varied the task constraints imposed on students: provision or not of the concepts used in the task. Mapping technique 1--student provides the concepts--asked students to construct a 10-concept map about "Ions, Molecules, and Compounds." Using the three concepts provided by the topic (i.e., ions, molecules, and compounds) students were asked to select another seven concepts that they thought were important in explaining ions, molecules, and compounds and construct the map (see Appendix C). Mapping technique 2--assessor provided concepts--asked students to construct a concept map using 10 concepts provided on the instruction page (see Appendix C). In both mapping techniques students were asked to organize the concepts in any way they wanted; no particular (e.g., hierarchical) structure was imposed. Also, students were encouraged to use as many words as they wanted to label the line between two concepts.

Concept Map Scoring System. The scoring system was based on a criterion map developed by the researchers using the 20 key concepts. The goal in constructing the criterion map was to identify those propositions (nodes and links) considered to be "substantial" to the domain, and that students should know about "Ions, Molecules, and Compounds" at that point in the chemistry course.

Based on the 20 key concepts, a square-matrix was constructed to define all possible links between pairs of concepts. The entries in the cell of the matrix denoted the relation between a specific pair of concepts. Up to 190 links can be drawn between the pairs of 20 concepts (see Appendix D).

To determine the "substantial" links, teachers, the expert, and ourselves constructed concept maps. The teachers and the expert constructed their maps based on the concepts they considered important in the chapter. We used the 20 key concepts. (The concepts selected by the expert are presented in Appendix A and were very similar to those in the key-concept list.) By comparing the concepts selected as key concepts across the three different sources, we concluded that the concept list was discipline valid.

Teachers' concept maps were expected to provide a benchmark for the "substantial" links students were expected to have after studying the chapter and participating in class. The expert's concept map provided the "substantial" links based on the structure of the discipline. Finally, we constructed a third map that was thought to reflect the "substantial" links in the textbook chapter.

An analysis of the four maps identified 48 "substantial" links. About a third of these links were the same across the four maps. We carefully analyzed the rest of the links (some found in the expert's map, others in the teachers' maps, others in our map) and concluded that all of them could be expected from students and justified as valid based on the instructional unit. These 48 links were used to construct a <u>criterion map</u> with the 20 key concepts and were considered as "<u>mandatory</u>"--students should reasonably be expected to provide any one of these propositions at that point in their instruction.

For each link in the criterion map a proposition was developed. The labeled links between pairs of concepts provided by the teachers, the expert and the researchers varied in the quality of their explication of the relationship. For example, the propositions used by the expert more completely explained the links between concept pairs than those used by the teachers. In fact, the propositions found in the expert's map and the researchers' map were more complete and accurate than those found in the teachers' maps. Furthermore, when students' maps were collected, we found that some students provided more accurate and complete propositions than the ones provided by the teachers.

To account for the variation in the quality of the proposition, we developed a Propositions Inventory. This inventory compiled the propositions (nodes and direction of links) provided by the teachers' maps, expert's map, researchers' map and students' maps and classified each proposition into one of five categories: Accurate Excellent, Accurate Good, Accurate Poor, "Don't Care" and Inaccurate. Table 3 presents the definition of each category (see Appendix E). For example, the accurate excellent proposition between acids and compounds should be read, according to the direction of the arrow (<), as follows: compounds that give off H^+ when dissolved in water are acids.

Insert Table 3 About Here

The Propositions Inventory provided propositions not only considered "mandatory," but also propositions for the "other possible" relations between the pairs of concepts in the Key-Concept List. These other propositions were considered as "possible" propositions. In this form any other proposition not contained in the criterion map could also be scored and credit was given if the proposition was valid. The Propositions Inventory was judged by the expert and a science educator to determine whether the classification of the propositions was accurate and appropriate. Both agreed on the classification.

The scoring system, based on the criterion map and the Propositions Inventory, evaluated two components of the map: the propositions and the nodes. The accuracy of each proposition in a student's map was assessed on a five-level scale (from 0 for inaccurate to 4 for accurate excellent) according to the Propositions Inventory. The concepts used as nodes were noted, counted, and classified as contained/not contained in our list of 20 key concepts. This last aspect was especially important for the maps constructed with student provided concepts.

Three map scores were formed: (1) a total proposition accuracy score--the total sum of the scores obtained on all propositions; (2) convergence score--the proportion of valid propositions in the student's map out of all possible propositions in the criterion map (i.e., the degree to which the student's map and the criterion map converge); (3) salience score--the proportion of valid propositions out of all the propositions in the student's map.

Separate score forms were designed for the three conditions (occasions)--No-Concepts, Sample A concepts, and Sample B concepts. Appendix E shows the form used to score the concept maps when Sample A was provided to students. Two raters scored each student map.

Multiple-Choice Test. Prior to administering the concept maps, all students received a short 15-item multiple-choice test on "Chemical Names and Formulas" designed by the researchers and reviewed by both teachers. The internal consistency reliability of the test was .67.

Training

A training miniprogram was designed to teach students, teachers, and the expert to construct concept maps and piloted with high school chemistry students not participating in the study. The same researcher trained both groups of study students to minimize variability. The training lasted about 50 minutes and had four major parts. The first part focused on introducing concept maps: what they are, what they are used for, what their components are (i.e., nodes, links, linking words, propositions), and examples (outside the domain to be mapped) of hierarchical and non-hierarchical maps. The second part emphasized the construction of concept maps. Four aspects of mapping were highlighted: identifying a relationship between a pair of concepts, creating a proposition, recognizing good maps, and redrawing a map. Students were then given two lists of common concepts to "collectively construct" a map. The first list focused on the "water cycle"--a non-hierarchical map; the second list focused on the "living things"--a hierarchical map. The third part of the program provided each individual with 9 concepts on the "food web," to construct a map individually. The fourth part of the program was a discussion of students' questions after they had constructed their individual maps.

After training, a random sample of 10 of the individually constructed maps was analyzed for each group (a total of 20 concept maps) to evaluate the training. This analysis focused on three aspects of the maps: use of the concepts provided on the list, use of labeled links, and the accuracy of the propositions. Results indicated that: (a) 97.8 percent of the students used all the concepts provided on the list, (b) 100 percent used labeled lines, and (c) 93.8 percent provided one or more valid propositions. We concluded that the training program succeeded in training the students to construct concept maps.

Procedure

The entire study was conducted in three 55 minute-sessions during a three-week period. Students were assessed on the same days in their respective classrooms. The first session was training. Two weeks of instruction followed. The second and third sessions were conducted consecutively after instruction.

At the second session, students first took the multiple-choice test (13 minutes, on average). Then, after all the students finished this test, a 15-minute reminder about concept maps was conducted. Finally, students constructed concept maps under the No-Concept-Provided condition. Although students had about 30 minutes to construct their maps, over 90 percent of the students finished in 20 minutes.

At the third session, students constructed maps with concepts provided. Class 1 first mapped with Sample A concepts, whereas Class 2 mapped with Sample B. After students finished their maps, they constructed the next map using the other sample of concepts. Construction of concept-provided maps took 14 minutes, on average, for both groups.

Results and Discussion

This study asked: Do two concept mapping techniques produce scores that can be interpreted in the same way, as representing the same aspect of a student's knowledge structure? If concepts are provided by the assessor, are map scores sensitive to the sampling of concept terms?

Before turning to these questions a preliminary methodological issue needs to be addressed: Does concept sample sequence (Sample A and then Sample B or viceversa) affect map scores? A 2x2 split-plot ANOVA using total proposition accuracy scores revealed no significant differences ($\alpha = .05$)

for sequence (S), concept sample (CS), or their interaction (S x CS; $F_S = .64$, $F_{CS} = .33$, $F_{SxCS} = .18$). Similar results were obtained when convergence ($F_S = .83$, $F_{CS} = .20$, $F_{SxCS} = 3.57$) scores were used. However, a significant interaction, sequence by concept sample was found when salience ($F_S = 01$, $F_{CS} = 1.51$, $F_{SxCS} = 10.28$) scores were analyzed. We decided to collapse the two groups and present results overall for brevity. (Presentation and discussion of results could be difficult to follow if different strategies were used for each type of score. Furthermore, salience scores were found to be the least desirable score to be used in assessing concept maps because their inconsistency in ranking students; see results below).[2]

Comparability of the Mapping Techniques

To evaluate whether the two mapping techniques are equivalent, we compared the scores produced by each technique, and their technical characteristics. For two techniques to be considered equivalent, they should produce similar means and variances, as well as similar indices of reliability and validity.

Mean and Variances. Table 4 presents the means and standard deviations across the three conditions (i.e., No-Concepts, Sample A and Sample B) for the proposition accuracy, convergence score, and the salience scores. Overall, students' knowledge about "ions, molecules, and compounds" was partial and their maps were not close to the criterion map. The low proposition accuracy and convergence scores across the three conditions indicated that students' knowledge was rather weak compared to the criterion map. Salience mean scores showed that about half of the propositions provided by the students were accurate, regardless of condition. Although in the No-Concept condition students had the opportunity to select concepts they felt most confident about to reflect their knowledge of the topic, still only half of their propositions were accurate.

Insert Table 4 Around Here

To test mean differences between the techniques, a repeated measures ANOVA over the three conditions was carried, one for the proposition accuracy score and other for the salience score.[3] We could not calculate differences in convergence means because this score was not available for the No-Concepts conditions since no criterion could be established to determine the expected number of propositions. For the proposition accuracy and salience scores, no significant differences were found among means (Propositions Accuracy: Hotelling's = .05; p = >.05) or the Salience means (Hotelling's = .06; p = >.05). The two concept mapping techniques, student provided or assessor provided concepts, produced similar mean map scores.

To evaluate the equivalence of variances between techniques, a Mauchly's test of sphericity was carried out for the proposition accuracy and salience scores. Results of the test indicated that variances did not differ significatively across the three conditions for the proposition accuracy and salience scores (Proposition Accuracy: W = .94, p = > .05; Salience: W = .90, p = > .05).

Reliability and Validity of Concept-Map Scores. We examined the generalizability of proposition accuracy scores across raters and conditions in a person by rater by assessment condition Generalizability (G) study (see Table 5). The largest variance component was for systematic differences among persons followed by the error component, person by condition interaction;

raters did not introduce error variability into the scores (percent of variability is negligible). Not surprisingly, students' relative standing varied from one condition (or sample of concepts) to the next (some students did better with Sample A, others with Sample B, and still others when they selected the concepts). These results are consistent with what has been found with science performance assessments: The person by task interaction has been a major source of unreliability (e.g., Shavelson, Baxter, Gao, 1993). Both relative and absolute "reliability" coefficients were high, suggesting that concept map scores can consistently rank students relative to one another ($\hat{\rho}^2$ = .90) as well as provide a good estimate of a student's level of performance, independently of how well their classmates performed ($\hat{\phi}$ = .90).

Insert Table 5 Around Here

Another G study was carried out for salience scores. Patterns of variability were the same (i.e., the highest percentage of variability was for persons followed by the person by condition interaction). Relative and absolute coefficients were of the same magnitude (Salience$_{(NC, A, B)}$: $\hat{\rho}^2$ = .79, $\hat{\phi}$ =.79), but lower than those found with proposition accuracy scores.

These results suggest that the type of score selected for scoring concept maps might be an issue. Notice that the percent of variability among persons is higher for the propositions accuracy score than for the salience score (see Table 5). This indicates that proposition accuracy scores reflect the differences in students' knowledge structure better than salience scores.

Table 6 presents a multiscore-multitechnique matrix. Although the three types of scores are presented in this table, we ignored the information related to convergence scores in this section because this type of scores were available only for Technique 2.

Interrater reliability coefficients are enclosed by parenthesis on the diagonal. Coefficients across types of score were high confirming the results obtained from the G studies: Raters can consistently score concepts maps. The lowest interrater coefficients were found in Technique 1--concepts given by students.

Insert Table 6 Around Here

To examine the convergence of the two techniques, we focused on the first two columns of the matrix. Correlations between the same type of score obtained with different techniques are underlined and expected to be high if techniques were equivalent. Coefficients are higher for the proposition accuracy score ($\underline{r}_{avg}$ = .73) than for the salience score ($\underline{r}_{avg}$ = .48). This suggests that the salience score may rank students differently depending on the technique used. This result is confirmed by looking at the correlations between different types of scores using different techniques (correlations in italics). The correlations between salience scores obtained with Technique 1, and proposition accuracy obtained with Technique 2, are lower ($\underline{r}_{avg}$ = .50) than the correlations between salience scores obtained with Technique 2 and the proposition accuracy score obtained with Technique 1 ($\underline{r}_{avg}$ = .63).

Finally, if concept maps measure somewhat different aspects of declarative knowledge than multiple-choice tests, the correlation between these two measures should be positive, because they measure the same knowledge domain, but moderate in magnitude because they tap into somewhat different aspects of students' knowledge in that domain. Furthermore, if Mapping Techniques 1 and 2 are equivalent, correlations between multiple-choice test scores and a particular type of map scores should be of similar magnitude across techniques.

Correlations between multiple-choice test scores and each type of scores for each technique are presented in Table 7. (As before, we focused only on the accuracy proposition and salience scores.) Along with the original correlations, we presented the correlations corrected for attenuation using the interrater reliability coefficient for each score. (This correction may not be accurate and must be interpreted cautiously. Reliability coefficients involved in the correction are not equivalent because the type of errors of measurement differ for each reliability coefficient.) We focused on the original correlations which are positive and moderately high. We interpret these findings to mean that concept maps and multiple-choice tests measure overlapping and yet somewhat different aspects of declarative knowledge. Although correlations between multiple-choice and types of scores are not the same across techniques (correlations are lower for Technique 1 than for Technique 2), no significant difference ($p > .05$) was found across the three conditions (see Meng, Rosenthal, & Rubin, 1992) for the proposition accuracy and the salience score (Proposition Accuracy: $\chi^2 = .792$; Salience: $\chi^2 = 1.844$). We concluded that Techniques 1 and 2 produced similar discriminant coefficients across types of scores.

Insert Table 7 Around Here

In sum, the evidence obtained suggested that Technique 1--asking students to provide the concepts to construct concept maps--and Technique 2--assessor provides a sample of 10 concepts to construct the maps--are equivalent. Both techniques provided similar means and variances. Furthermore, reliability and validity coefficients were similar. However, based on practical grounds, we recommend the use of Technique 2 because is far easier to score. For example, in Technique 1 credit was given to all valid propositions provided by students in their maps, even though some of these propositions were a "don't care" type; that is, the propositions were not essential to the topic assessed. (For example, students used concepts such as "chemistry" or "chemical substances," and related them in a way that the explanation of the relation was valid--chemical substances are studied in chemistry, but do not provide any evidence of knowledge of the topic.) Giving students the opportunity to select the concepts for their maps leaves the door open for this to happen. The issue can be resolved by not considering the propositions that involved concepts not directly related to the topic. However, more decisions would need to be made by the rater (e.g., is "chemistry" a concept that should not be considered as essential to the topic?) that may decrease consistency across raters. Results also indicate that the proposition accuracy is a better score than the salience score. Salience scores proved to be less consistent across techniques.

Comparability of Concept Sample Scores

To evaluate whether different samples of concepts, Sample A and Sample B, influenced students' map scores we compared means and variances across samples, as well as reliability and validity coefficients.

Means and Variances. Results from the analyses across techniques (see Means and Variances in the previous section) indicated no significant difference between means for the proposition accuracy or salience scores (Table 4). For the convergence score, available only for Sample A and Sample B, likewise, no significant difference was found between means (Hotelling's T^2 = .16; p = >.05) or variances ($F_{Max.}$ = .80; p = >.05).[4]

Reliability and Validity. The generalizability of scores across raters and concept samples was examined in three, person x rater x concept-sample, G studies, one for each type of score (Table 8).

Insert Table 8 Around Here

The pattern observed was the same as in the comparison of Techniques 1 and 2. The largest variance component was for persons, followed by the interaction of person by concept-sample; variability between raters was negligible. The proposition accuracy scores had the highest relative and absolute coefficients (.89 for both coefficients). Convergence and salience scores had similar coefficients. However, the convergence score had a higher percent of variability for persons meaning that it better reflected the differences in students' knowledge than the salience scores.

The multiscore-multitechnique matrix (Table 6) shows the interrater reliability coefficients across score types for concept Sample A and B. Interrater coefficients were consistently high (above .90).

To compare coefficients across concept samples we focused on the second and third columns of the matrix. The highest correlations between score types within the same concept-sample were the correlations between proposition accuracy and convergence scores ($\underline{r}_{avg}$ = .96), followed by the convergence and salience score correlations ($\underline{r}_{avg}$ = .93), and then by the propositions accuracy and salience score correlations ($\underline{r}_{avg}$ = .90).

Correlations between the same score type across the two concept-samples are underlined in Table 6. The highest correlations were for the proposition accuracy scores ($\underline{r}$ = .83) and the lowest for the salience scores ($\underline{r}$ = .71). When different scores are correlated across concept-samples (correlations in italics), those correlations involving salience scores were, in general, the lowest ($\underline{r}_{avg}$ = .70). Salience scores were the least consistent.

Table 7 provides the correlations between multiple-choice scores and the different score types across the two concept-samples. Correlations are positive and moderate to high, but far from unity. We interpreted this to mean that multiple-choice achievement tests and concept maps tap somewhat different aspects of students' knowledge.

In sum, concepts randomly selected from a list of key concepts provided roughly equivalent concept map scores. Moreover, proposition accuracy and convergence scores better reflected the differences in students' knowledge structure than salience scores. However, based on the time and effort involved in developing the scoring guide for accuracy scores, we recommend, for large-scale assessment, the use of the convergence score (i.e., the proportion of valid propositions in a student's map out of the possible propositions in the criterion map).

A Closer Look At Students' Maps

In a closer analysis of the students' concept maps, we examined two issues: (1) the characteristics of the concepts selected by the students with

Mapping Technique 1: How relevant were the concepts to the topic assessed? Were the concepts selected by the student the same as those selected in the key concept list (from which Sample A and B were created)? And (2) Could some misconceptions be identified in the students' maps?

Characteristics of the Concepts in Students' Maps. Table 9 presents the mean number of key and "other" concepts, and their total used in the maps along with standard deviations.

Insert Table 9 About Here

About 75 percent of the students selected at least six of the key concepts (including the three concepts provided) in Mapping Technique 1. Only one student selected 10-concepts that were all key concepts. Not all students used the three concepts provided in the instructions (i.e., ions, molecules and compounds); 82.5 percent used the three concepts, 15 percent used at least two, and 2.5 percent (i.e., one student) did not use any of the three.

Besides ions, molecules, and compounds (the three concepts provided in the instructions; see Appendix C), students selected anions, cations, and acids most frequently for their maps. From the other 16 concepts on the list, 14 (e.g., binary ionic compounds, polyatomic ions, bases) were selected by at least one student. Binary molecular compounds and neutral charge were not selected by any student. The last finding was a surprise because the knowledge that compounds have a neutral charge is a key idea in the topic assessed. The correlation between the number of key/core concepts on the students' concept maps with the proposition accuracy score (see Table 4) was moderate and significant (r = .51, p = <.005): Those students who recalled more key/core concepts tended also to have a greater proposition accuracy score.

The "other" concepts provided by the students were classified as related, but not important, and not related (e.g., plants, animals) to the knowledge domain assessed. About 63 percent of the "other" concepts were considered related to the topic "ions, molecules and compounds" (e.g., element, ductile, HCl). Of all these concepts, "element" was the most frequently used as an "other" concept. Forty-seven percent of the students selected it for their maps. "Element" was on our original 23-concept list, however, we decided to drop it for two reasons: teachers did not selected it and only a few links with the others concepts on the list could be drawn. Nevertheless, this concept should have been included on the key-concept list.

Another missing set of items were "examples of compounds." Inclusion of examples of molecular compounds, polyatomic ions, acids or ternary ionic compounds would have allowed students to show their understanding of, say, how HCl is related to anions, cations, and binary ionic compounds, instead of only memorizing the formula. (For example, does the student understand that HCl is an acid and is a binary ionic compound that has one simple anion, Cl^-, and a cation, H^+?)

Not related concepts included, for the most part, "general" chemistry concepts (e.g., symbols, mixtures, substances). Even though they can be related to the other key concepts, they do not reflect students' understanding of ions, molecules, and compounds. The correlation between number of "other" concepts and the proposition accuracy total score (see Table 4) was close to zero ($\underline{r}$ = -.09, $\underline{p}$ = >.05).

Not surprisingly, if a student had an adequate understanding of the topic, she tended to provide topic-relevant concepts and the propositions between pairs of concepts tended to be valid. When a student's understanding was weak, she provided more non-relevant "other" concepts, which resulted in superficial (i.e., "don't care") and/or inaccurate propositions.

An average of about 86 percent of the students over the Sample A and B conditions used the ten concepts provided in the mapping instructions. One student used only 8 of the Sample B concepts. One student used 13 concepts to construct the Sample A map, 10 provided on the list and 3 "other" concepts. Although the three concepts provided by the student were related to the topic (e.g., positive charge) the propositions that included these "other" concepts. In this way, the student's score was not advantaged by the additional concepts. The same was done for the No-Concept condition. When more than ten concepts were provided in the students' maps, concepts were randomly dropped, along with their links, to leave ten concepts.

Misconceptions Evident in the Concept Maps. In reviewing the two teachers' maps, we identified a misconception in the map of the teacher who taught the students in this study. According to this teacher's map, anions and cations lose or gain electrons when, in fact, atoms lose or gain electrons to become ions; and ions, according to their charge--positive or negative--are either cations or anions. We decided to take a closer look at students' maps to find out whether this misconception was reproduced by the students in their maps. About 17 percent of the students showed exactly the same misconception.

Another phenomenon observed but not evident by simply examining the scores, was how the sample of concepts seems to "lead" some students to create connections that, probably due to their partial knowledge, result in invalid propositions. We observed that more students related molecules to ions with Sample B than with Sample A concepts. Sample B's limited selection of concepts that could be related to molecules "forced" students to look for more connections with the concept, even though these connections were incorrect.

Conclusions

This study explored the potential of two concept mapping techniques for use in large-scale science assessment. We examined: (1) whether map scores were sensitive to who samples the concepts to be used in the map (student or assessor) and, if the assessor does the sampling, variation from one random sample of concepts from a domain, to another, and (2) how reliable concept map scores are and how valid the interpretation of this scores is. To this end we constructed two mapping techniques: Technique 1--students provide the concepts with which to construct a map, and Technique 2--assessor provides a set of 10 concepts to construct the map.

Our findings lead to the following tentative conclusions: (1) The two mapping techniques provided equivalent scores reflecting some aspects of students' knowledge structures. However, with Technique 1 students were probably given too much credit for propositions that did not provide evidence of their knowledge of the topic assessed. We plan to explore this issue further before reaching a final conclusion about the equivalence of these mapping techniques. (2) Sampling variability from one random sample of concepts to another provides equivalent maps scores when the concept domain is carefully specified. (3) Concept maps can be reliably scored, even when complex judgments such as proposition quality are required. (4) The relationship between multiple-choice scores and concept map scores suggests that they measure overlapping and yet somewhat different aspects of declarative knowledge. (5) The convergence score--the proportion of valid propositions in the student's map to the number of all possible propositions in the criterion map--may be the most effort and time efficient indicator when scoring students' concept maps.

We also found that students can be trained to construct concept maps in a short period of time with limited practice. From a closer look at students' maps we know that even though practice may improve map characteristics, still students were able to demonstrate their knowledge on the topic assessed.

It seems, then, at least from the technical quality perspective, that the potential of concept maps as an assessment tool in science is high. Nevertheless, still more questions need to be answered before we can conclude that they can reliably and validly evaluate students' knowledge structure, especially in high stake accountability situations.

References

Anderson, R. C. (1984). Some reflections on the acquisition of knowledge. Educational Researcher, 13(10), 5-10.

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

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

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

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

Briscoe, C. & LaMaster, S. U. (1991). Meaningful learning in college biology through concept mapping. The American Biology Teacher, 53(4), 214-219.

California Department of Education (1990). Science framework for California Public Schools: Kindergarten through grade twelve. Sacramento, CA: California Department of Education.

Cronbach, L. J. (1990). Essentials of psychological testing (Fifth ed.). New York: Harper & Row Publishers.

Dochy, F. J. R. C. (1996). Assessment of domain-specific and domain-transcending prior knowledge: Entry assessment and the use of profile analysis. In M. Birenbaum & F. J. R. C. Dochy (Eds.) Alternatives in assessment of achievements, learning process and prior knowledge (pp. 93-129). Boston, MA: Kluwer Academic Publishers.

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

Glaser, R. & Bassok, M. (1989). Learning theory and the study of instruction. Annual Review of Psychology, 40, 631-66.

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

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

Hoz, R., Tomer, Y., & Tamir, P. (1990). The relations between disciplinary and pedagogical knowledge and the length of teaching experience of biology and geography teachers. Journal of Research in Science Teaching, 27(10), 973-985.

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

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

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

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

Meng, X. L., Rosenthal, R. & Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological Bulletin, 111(1), 172-175.

Nakhleh, M. B. & Krajcik, J. S. (1991). The effect of level of information as presented by different technology on students' understanding of acid, base, and pH concepts. Paper presented at the annual meeting of the National Association for the Research in Science Teaching, Lake Geneva, WI. (ERIC Document Reproduction Service No. ED 347 062).

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

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

Pankratius, W. (1990). Building an organized knowledge base: Concept mapping and achievement in secondary school physics. Journal of Research in Science Teaching, 27(4), 315-333.

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(6), 569-600.

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

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

Schmid, R. F. & Telaro, G. (1990). Concept mapping as an instructional strategy for high school biology. Journal of Educational Research, 84(2), 78-85.

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

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(10), 1033-1052.

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

Wilbraham, A. C., Staley, D. D., Simpson, C. J., & Matta, M. S. (1990). Chemistry. Menlo-Park, CA: Addison-Wesley Publishing Company.

Willerman, M. & Mac Harg, R. A. (1991). The concept map as an advance organizer. Journal of Research in Science Teaching, 28(8), 705-711.

Notes

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

[2] All the statistical analyses were carried out for the two groups (Sequence AB and BA). Results are available from the authors.

[3] Although we report regular salience score means, the analyses for this type of score were carried out using both the proportions and the arcsine transformation of proportions. Results were consistent across both types of data.

[4] As with salience scores, we report regular convergence score means, although the analyses were carried out using both the proportions and the arcsine transformation of proportions. Results were also consistent across both types of data.

Table 1

Concept Map Components and Variations Identified.

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

Table 2

Five Examples of Different Types of Tasks, Response Format and Scoring Systems Used in Research of Concept Maps

Authors	Task	Response	Scoring System
• Barenholz & Tamir, 1992	Select 20 to 30 concepts considered key concepts for a course in microbiology and use them to construct a map.	Paper-and-pencil response. Students drew the concept map in their notebooks.	Score of map components: number of concepts and propositions, the hierarchy and the branching, and quality of the map based on overall impression.
• Fisher, 1990	Task 1. Enter concepts and relation names in the computer with as many links as desired. Task 2. Fill-in-the-blank when a central concept is masked and the other nodes are provided.	Computer response in both tasks. Students construct their maps on a blank screen for task 1, and filled-in the node(s) in a skeleton map for task 2.	The author only proposed the SemNet computer program as an assessment tool, but did not present any scoring system to evaluate the maps.
• Lomask, Baron, Greig, Harrison, 1992	Write an essay on two central topics on biology (i.e., growing plant and blood transfusion).	Paper-and-pencil response. Trained teachers construct a map from students' written essay. No effort was made to elicit any hierarchy.	Comparison with a criterion map. Two structural dimensions were identified for the comparison: the *size* and the *strength* of structure. The final scored was based on the combination of both dimensions.
• Nakhleh & Krajcik, 1991	Semi-structured interview about acids and bases.	Oral response. The interviewer drew three concepts maps--one for acids, one for bases, and one for pH--based on statements that revealed the student's propositional knowledge.	Score based on map components: Propositions and examples, cross-links, hierarchy. Experts' maps were used to identify critical nodes and relationships.
• Wallace & Mintzes, 1990	Construct a hierarchical concept map from ten given concepts on life zones.	Paper-and-pencil response. Students drew the concept map on a blank page.	Score based on map components: number of relationships, levels of hierarchy, branching, cross-links, and general-to-specific examples.

Table 3.

Accuracy of the Propositions.

Accuracy of Proposition	Definition
Excellent:	Outstanding proposition. Complete and correct. It shows a deep understanding of the relation between the two concepts. *acids-compounds: < that gives off H^+ when dissolved in water are*
Good:	Complete and correct proposition. It shows a good understanding of the relation between the two concepts. *acids-compounds: > are examples of*
Poor:	Incomplete but correct proposition. It shows partial understanding of the relation between the two concepts. *acids-compounds: < form*
Don't Care:	Although valid, the proposition does not show understanding between the two concepts. *acids-compounds: > is a different concept*
Inaccurate/ Invalid	Incorrect proposition. *acids-compound: >made of*

Table 4

Mean and Standard Deviations of the Propositions Accuracy, Convergence, and Salience Scores on Each Condition by Each Group.

	No-Concepts		Sample A		Sample B	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Proposition Accuracy	11.98[a]	(7.62)	11.76	(8.86)	12.52	(7.20)
	(Max=192)		(Max=108)		(Max=124)	
Convergence	b		.17	(.10)	.16	(.08)
			(Max. # of Prop=27)		(Max. # of Prop=31)	
Salience	.51	(.25)	.47	(.27)	.51	(.24)

a Maximum score was calculated based on the 48 excellent valid mandatory propositions students could provide if 10 concepts were used in a map.

b Proportions were not calculated since no criterion could be established to determine the expected number of propositions.

Table 5

Estimated Variance Components and Generalizability Coefficients for a Person x Rater x Condition G Study Design for No-Concept, Sample A, and Sample B Conditions using the Propositions Accuracy and Salience Scores.

Source of Variation	Estimated Variance Components	Percent of Total Variability
Proposition Accuracy Score (NC, A, B)		
Persons (p)	46.991	73.47
Rater (r)	0.074	0.12
Condition (c)	0.000*	0.00
pr	0.000*	0.00
pc	13.526	21.15
rc	0.000*	0.00
prc,e	3.365	5.26
$\hat{\rho}^2$ ($n_r = 2, n_c = 3$)	.90	
$\hat{\phi}$	.90	
Salience Score (NC, A, B)		
Persons (p)	.03682	52.48
Rater (r)	.00018	0.26
Condition (c)	.00007	0.10
pr	0.00000*	0.00
pc	.02469	35.19
rc	0.00000*	0.00
prc,e	.00840	11.97
$\hat{\rho}^2$ ($n_r = 2, n_s = 3$)	.79	
$\hat{\phi}$	.79	

* Negative variance components set to zero; in no case the variance component was more than -0.08910 for proposition accuracy and -0.0036 for salience score.

Table 6

Multiscore-Multitechnique Matrix*

	No-Concept		Sample A			Sample B		
	PA	S	PA	C	S	PA	C	S
1. No-Concept								
Prop Accuracy (PA)	(.91)							
Salience (S)	.81	(.79)						
2. Sample A								
Prop Accuracy (PA)	.73	.47	(.98)					
Convergence (C)	.68	.48	.96	(.97)				
Salience (S)	.62	.47	.89	.94	(.96)			
3. Sample B								
Prop Accuracy (PA)	.73	.53	.83	.83	.72	(.96)		
Convergence (C)	.62	.47	.70	.74	.61	.95	(.93)	
Salience (S)	.63	.50	.72	.76	.71	.90	.91	(.90)

* Interrater reliability on the diagonal.

Table 7

Correlation Between the Multiple-Choice Test and Proposition Accuracy, Convergence, and Salience Score.*

	No-Concepts	Sample A	Sample B
Proposition Accuracy	.57 .73	.63 .78	.64 .80
Convergence	a -	.64 .79	.56 .71
Salience	.43 .59	.61 .76	.50 .64

* The second line presents the coefficients corrected for attenuation.

[a] Not calculated (see note on Table 4).

Table 8

Estimated Variance Components and Generalizability Coefficients for a Person x Rater x Concept-Sample G Study Design for Sample A, and Sample B Across the Three Types of Scores.

Source of Variation	Estimated Variance Components	Percent of Total Variability
Proposition Accuracy		
Persons (p)	52.85513	79.82
Rater (r)	0.00000*	0.00
Sample (s)	.05353	0.08
p x r	.64103	0.97
p x s	11.37147	17.17
r x s	.05897	0.09
prs,e	1.24103	1.87
$\hat{\rho}^2$ $(n_r = 2, n_s = 2)$	.89	
$\hat{\phi}$	.89	
Convergence		
Persons (p)	.00668	70.02
Rater (r)	.00001	0.10
Sample (s)	.00000*	0.00
p x r	.00003	0.31
p x s	.00244	25.58
r x s	.00001	0.10
prs,e	.00037	3.88
$\hat{\rho}^2$ $(n_r = 2, n_s = 2)$	.83	
$\hat{\phi}$	.83	
Salience		
Persons (p)	.04660	67.08
Rater (r)	.00014	0.20
Sample (s)	.00077	1.11
p x r	.00000*	0.00
p x s	.01727	24.86
r x s	.00011	0.16
prs,e	.00458	6.59
$\hat{\rho}^2$ $(n_r = 2, n_s = 2)$	.83	
$\hat{\phi}$	.82	

* Negative variance components set to zero; in no case the variance component was more than -0.01603.

Table 9

Mean and Standard Deviation of Concepts Considered as Key Concepts, "Other" Concepts, and the Total Number of Concepts Used by Students Across the Three Conditions

Mapping Conditions	Key/Core Concepts Mean	Key/Core Concepts S.D.	Other Concepts Mean	Other Concepts S.D.	Total Number Mean	Total Number S.D.
No-Concepts	6.40	(2.05)	3.28	(2.54)	9.68	(2.20)
Sample A	9.88	(.33)	.10	(.50)	9.98	(.62)
Sample B	9.83	(.45)	.03	(.16)	9.85	(.43)

Appendix A

Compiling the List of Concepts

Teachers were asked to answer these two questions about the unit: (1) "Explain in a few words what you want your students to know when they finish this chapter. In answering this question, think about why this chapter is included in the curriculum and what is the most important thing you want the students to learn about the topic;" and (2) "Based on your answer to question 1, please review the chapter and list all the concepts you think students should know and understand after studying this topic."

Teachers' answers to question 1 involved two aspects of the unit: (a) conceptual understanding (e.g., "Students should have a good understanding of the formation of ions, the differences between molecules/compounds... molecular/ionic compounds and acids.") and (b) application (e.g., "They should be able to form ionic compounds, binary, ternary, and acids...be familiar with the periodic table to identify metals, non-metals...." "Students... should be able to write chemical/molecular formulas; name different substances..."). We focused on the conceptual understanding of the unit since concepts maps are about the interrelatedness of concepts. From teachers' responses we concluded that the conceptual understanding about "chemical compounds" (i.e., the other group of substances that are not elements) was the main focus of the unit--how compounds are formed, the types of compounds, and how they can be combined are issues discussed in the chapter.

Answers to question 2 lead to a list of 12 concepts considering both teachers' responses (see table below). We noticed, however, that even though teachers included in their answers to question 1 concepts such as binary ionic compounds or polyatomic ions, they did not include them in the list of concepts in question 2.

Our list included 23 key concepts selected from the chapter. We gave this list to the teachers with the following instructions: "This is a list of concepts that were selected from the chapter, Chemical Names and Formulas. Based on what you think are the most important ideas for students to understand about Chemical Names and Formulas, check (√) the concepts that are essential. Please feel free to add any concepts that are missing." Only one of the two teachers returned the list reviewed. Based on the concepts selected by the teacher we reduced the list to 20 concepts (see Appendix B). This list of 20 concepts was considered to represent the "key concepts" of the chapter.

List of Concepts Selected from the Revision of the Chapter, The Teachers, and The Expert

Original Key Concept List	Teachers' List	Expert's List
1. acids	1. acids	1. acids
2. anions	2. anions	2. anions
3. atoms	3. cations	3. atoms
4. bases	4. compounds	4. bases
5. binary ionic compounds	5. element	5. binary ionic compounds
6. binary molecular compounds	6. ionic charge	6. cations
7. cations	7. ionic compounds	7. compounds
8. compounds	8. molecules	8. electrons
9. electrons	9. molecular compounds	9. elements
10. elements	10. periodic table	10. ions
11. ions	11. chemical formulas	11. ionic compounds
12. ionic compounds	12. molecular formulas	12. metals
13. metals		13. molecules
14. metalloids		14. molecular compounds
15. molecules		15. negative charge
16. molecular compounds		16. neutral charge
17. negative charge		17. non-metals
18. neutral charge		18. representative elements
19. non-metals		19. polyatomic ions
20. polyatomic ions		19. positive charge
21. positive charge		20. ternary ionic compound
22. ternary ionic compound		21. transition elements
23. transition metals		

Appendix B

List of Concepts Considered for the Three Conditions

Key Concept List	List A	List B
1. acids	1. acids	1. acids
2. anions	2. anions	2. anions
3. atoms	3. cations	3. binary ionic compounds
4. bases	4. compounds	4. compounds
5. binary ionic compounds	5. electrons	5. electrons
6. binary molecular compounds	6. ions	6. ions
7. cations	7. metals	7. molecules
8. compounds	8. molecules	8. negative charge
9. electrons	9. molecular compounds	9. non-metals
10. ions	10. polyatomic ions	10. ternary ionic compound
11. ionic compounds		
12. metals		
13. molecules		
14. molecular compounds		
15. negative charge		
16. neutral charge		
17. non-metals		
18. polyatomic ions		
19. positive charge		
20. ternary ionic compound		

Appendix C

Sample of Instructions

Instructions for the Concept Mapping Technique 1--No-Concepts Are Provided to the Students.

Name ______________________________ Period __________

You recently studied the chapter on Chemical Names and Formulas.

Construct a concept map that reflects what you know about Ions, Molecules, and Compounds.

The concept map should have 10 concepts in it. We are providing you with 3 concepts: ions, molecules, and compounds.

Select another 7 concepts to construct your map. The 7 concepts should be the ones that you think are the most important in explaining ions, molecules, and compounds.

Organize the terms in relation to one another in any way you want. Draw an arrow between the terms you think are related. Label the arrow using phrases or only one or two linking words.

You can construct your map on the blank pages attached. When you finish your map check that: (1) all the arrows have labels; (2) your concept map has 10 concepts, and (3) your map shows what you know about ions, molecules, and compounds.

After checking your map redraw it so someone else can read it. Staple your final map to this page.

Instructions for the Concept Mapping Technique 2--List of 10 Concepts (Sample A) Are Provided to the Students.

Name ______________________________ Period__________

Examine the concepts listed below. They were selected from the chapter on Chemical Names and Formulas that you recently studied. The terms selected focus on the topic Ions, Molecules, and Compounds.

Construct a concept map using the terms provided below.

Organize the terms in relation to one another in any way you want. Draw an arrow between the terms you think are related. Label the arrow using phrases or only one or two linking words.

You can construct your map on the blank pages attached. When you finish your map check that: (1) all the arrows have labels; (2) your concept map has 10 concepts, and (3) your map shows what you know about ions, molecules, and compounds.

After checking your map redraw it so someone else can read it. Staple your final map to this page.

You have 30 minutes to construct the map.

LIST OF CONCEPTS-
acids
anions
cations
compounds
electrons
ions
metals
molecules
molecular compounds
polyatomic ions

Appendix D

Matrix of the Relations Between Pairs of Concepts

	acids	anions	atoms	bases	binary ionic compound	binary molecular compound	cations	compounds	electrons	ions	...	positive charge	ternary ionic compounds
acids	X	1	M* 2	3	M 4	5	M 6	M 7	8	9	...	18	M 19
anions		X	20	21	M 22	23	24	25	M 26	M 27	...	36	37
atoms			X	38	39	40	41	M 42	M 43	M 44	...	53	54
bases				X	55	56	M 57	M 58	59	60	...	69	M 70
binary ionic compound					X	71	M 72	73	74	75	...	84	85
binary molec. compound						X	86	87	88	89	...	98	99
cations							X	100	M 101	M 102	...	111	M 112
compounds								X	113	M 114	...	123	124
electrons									X	M 125	...	134	135
ions										X	...	144	145
⋮											...	⋮	⋮
positive charge												X	190
ternary ionic compounds													X

*M stands for mandatory propositions.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM028965

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: "Concept-Map Based Assessment: On Possible Sources of Sampling Variability."

Author(s): Maria Ruiz-Primo, Richard Shavelson

Corporate Source: Stanford University

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Maria Araceli Ruiz Primo	Printed Name/Position/Title: Research Associate	
Organization/Address: School of Education Stanford University. Stanford CA 94305-3096	Telephone: (650) 725-1253	FAX: (650) 725-7412
	E-Mail Address: aruiz@leland-stanford.edu	Date: 4/30/98

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.