Comparison of the Reliability and Validity of Scores From Two Concept-Mapping Techniques

Concept-Map Representation of Knowledge Structures: Report of Year 2 Activities

CSE Technical Report 492

Maria Araceli Ruiz-Primo, Susan E. Schultz, Min Li, and Richard J. Shavelson
CRESST/Stanford University

November 1998

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Models-Based Assessment Design: Individual and Group Problem Solving—Individual Problem Solving in Science Richard J. Shavelson, Project Director, CRESST/Stanford University

The work reported herein was supported in part 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 do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

COMPARISON OF THE RELIABILITY AND VALIDITY OF SCORES FROM TWO CONCEPT-MAPPING TECHNIQUES

Maria Araceli Ruiz-Primo, Susan E. Schultz, Min Li, and Richard J. Shavelson

Stanford University/CRESST

Abstract

This paper reports the results of a study that compares two mapping techniques, one high-directed, "fill-in-a-skeleton map," and one low-directed, "construct-a-map-from-scratch." We examined whether (a) skeleton map scores were sensitive to the sample of nodes or linking lines to be filled in, (b) the two forms of skeleton maps were equivalent, and (c) the two mapping techniques provided similar information about students' connected understanding. Results indicated that high-directed (i.e., fill-in-the-map) and low-directed (i.e., constructing-a-map) maps lead to different interpretations about students' knowledge structure. Whereas scores obtained under the high-directed technique indicated that students' performance was close to the maximum criterion, the scores obtained with the low-directed technique revealed that students' knowledge was incomplete compared to a criterion map. Furthermore, the low-directed technique provided a symmetric distribution of scores, whereas the high-directed technique scores were negatively skewed. We concluded that construct-a-map technique better reflected differences among students' knowledge structures.

Concept maps have been used to assess students' knowledge structures, especially in science education. The justification for assessing student's knowledge structures is based on theory and research showing that understanding a subject domain such as science is associated with a rich set of relations among important concepts in the domain. We know, for example, that successful learners develop elaborate and highly integrated frameworks of related concepts (Mintzes, Wandersee, & Novak, 1997), just as experts do (Chi, Glaser, & Farr, 1988; Glaser, 1991). Furthermore, we know that highly organized structures facilitate problem solving and other cognitive activities (e.g., generating explanations or recognizing rapidly meaningful patterns; Baxter, Elder, & Glaser, 1996; Mintzes, Wandersee, & Novak, 1997). Research has shown

that differences in the performance of experts and novices is due, largely, to how knowledge is structured in their memories (Chi et al., 1988; Glaser, 1991).

Concept maps are interpreted as providing a "picture" of how key concepts in a domain are mentally organized/structured by students. With this assessment technique, students are asked to link pairs of concepts in a science domain and label the links with a brief explanation of how the two concepts go together.

Although concept maps have been used in large-scale as well as classroom assessment, a wide variety of techniques are called concept maps and little is know about the reliability and validity of scores produced by these varying mapping techniques (e.g., Ruiz-Primo & Shavelson, 1996). We suspect that the observed characteristics of the representation of a student's knowledge structure depend to a large extent on how the representation is elicited. Simply put, the method used to ask students to represent their knowledge can affect the representation they provide as well as the score they obtain (Ruiz-Primo & Shavelson, 1996; Ruiz-Primo, Schultz, & Shavelson, 1996; Ruiz-Primo, Shavelson, & Schultz, 1997). Through a series of studies we seek to increase our understanding of how different mapping techniques affect the representation and interpretation of a student's knowledge structure. In this paper, we provide reliability and validity evidence on the effects of two mapping techniques, "fill-in-the-map" and "construct-a-map."

Concept-Map Assessment

We define a concept map as a graph in which the nodes represent concepts, the lines between nodes represent relations, and the labels on the lines represent the nature of the relations. The combination of two nodes and a labeled line is called a proposition—the fundamental unit of the map. Our characterization of a concept map assessment as based on its three components—a task, its response format, and a scoring system—has revealed the enormity of variations in mapping techniques used in research and practice (see Ruiz-Primo & Shavelson, 1996).

The characteristics of the task, the response format, and the scoring system hold the key for tapping what concept-map based assessments are intended to evaluate: knowledge structure (or "connected understanding," for some

authors). The assessment task, for example, can vary in the constraints (directedness) it imposes on a student in eliciting her representation of structural knowledge. One dimension in which directedness varies lies in what is provided for use in the concept map (Figure 1).

If the characteristics of the assessment task fall on the left extreme, the student's representation is probably determined more by the mapping technique (or the assessor, if you will) than by the student's own knowledge or connected understanding.[1] If the assessment task falls on the right extreme, the student is free to decide which and how many concepts to include in her map, which concepts are related, and which words to use for explaining the relation. This openness may also be undesirable because of practical issues. For example, asking the student to generate the concepts to construct her map provides a good piece of information about the student's knowledge in a particular domain (e.g., are the concepts selected by the student relevant/essential to the topic?). However, scoring issues may make this option impractical—for example, each concept map has a unique scoring system. In one of our studies (Ruiz-Primo et al., 1996) we compared two mapping techniques that differed on whether the concept sample was student-generated or assessor-generated. The student-generated sample technique presented more challenges in scoring students' representations.

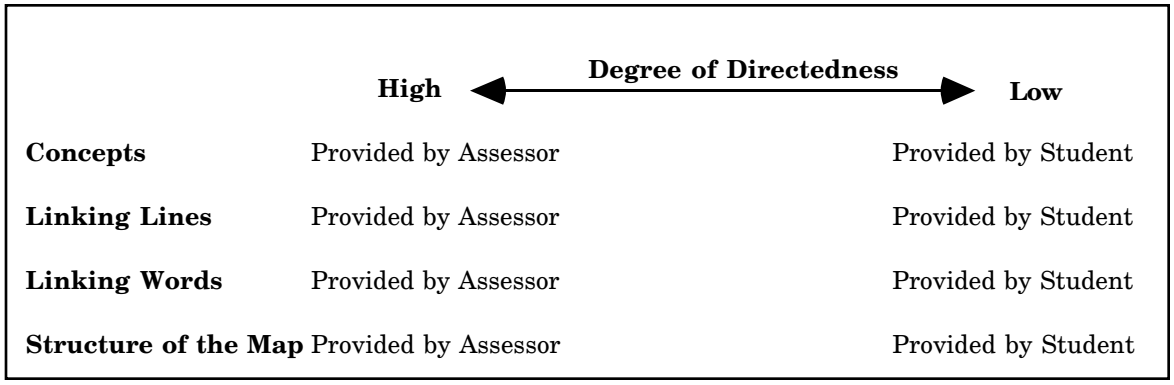

Figure 1. Degree of directedness in a concept assessment task.

[1] The characteristics of the assessment task have an impact on the response format and the scoring system. For example, a task that provides the structure of the map will probably provide that structure in the student's response format. If the task provides the concepts to be used, the scoring system will not focus on the "appropriateness of the concepts" used in a map. The combination of the task, the response format, and the scoring system is what determines a mapping technique.

The cognitive demands imposed on students by high-directedness techniques are different from those imposed by low-directedness techniques. Furthermore, high-directed techniques are more likely to misrepresent the student's knowledge structure by imposing a structure on their responses. In this study we examined the reliability and validity of two mapping techniques, one that can be considered high-directed and the other low-directed.

Defining the Two Mapping Techniques

Some researchers (e.g., Schau & Mattern, 1997) have argued that asking students to draw a map from scratch imposes too high a cognitive demand on students to produce a meaningful representation of their knowledge. An alternative technique is the fill-in-the-map method. Below we describe both techniques.

Fill-in-the-map. The "fill-in-the-map" technique provides students with a concept map where some of the concepts and/or the linking words have been left out. Students fill in the blank nodes or blank linking lines (e.g., Anderson & Huang, 1989; McClure & Bell, 1990; Schau, Mattern, Weber, Minnick, & Witt, 1997). The response format is straightforward; students fill in the blanks and their responses are scored correct or incorrect. Arguments can be made for (e.g., ease of administration, scoring, and retrieval of propositions from long-term memory) and against (e.g., imposes a structure on a student's knowledge) the technique. We posit that as students' subject matter knowledge increases, the structure of their maps should increasingly reflect the structure of the domain as held by experts (see Glaser, in press; Shavelson, 1972, 1974). By imposing a structure on the relations between concepts, it is difficult to know whether or not students' knowledge structures are becoming increasingly similar to experts'. Structure of representation, however, is not the only issue to consider. With "fill-in," students are usually provided with linking words in the skeleton map, and they only select the concepts from a list of concepts. Yet, in our research using the construct-a-map technique, we found that the linking words students used to relate two concepts provide insight into students' understanding in a particular content domain (e.g., Ruiz-Primo et al., 1996).

Construct-a-map from scratch. The "construct-a-map" technique varies as to how much information is provided by the assessor (Figure 1). The assessor may provide the concepts and/or linking words or may ask students to construct a

hierarchical or nonhierarchical map. The response format is simply a piece of paper provided on which students construct the map. Scoring systems vary from counting the number of nodes and linking lines (not recommended) to evaluating the accuracy of propositions (see Ruiz-Primo & Shavelson, 1996).

This mapping technique, however, has been considered problematic for large-scale assessment because students need to be trained to use maps, and scoring is difficult and time consuming (e.g., Schau et al., 1997). Our research has tried to overcome these two problems (see Ruiz-Primo et al., 1996, 1997). We designed a 50-minute program to teach students how to construct concept maps. The program proved to be effective in achieving this goal with more than 100 high school students. Moreover, to find an efficient scoring system, we have explored different types of scores, some based only on the propositions, others using a criterion map. Map propositions can be scored as to degree of their accuracy and comprehensiveness or simply as correct or incorrect. Based on this differentiation we have studied three types of scores: *proposition accuracy score*—the sum of individual proposition scores obtained on a student's map; *convergence score*—the proportion of accurate propositions in the student's map out of all possible propositions in the criterion map; and *salience score*—the proportion of correct propositions out of all propositions in the student's map. (The scoring system we have used has yielded high interrater reliability coefficients, above .90, even when the quality of the propositions is judged.)

Purpose

This study explored the technical characteristics of the "fill-in-the-map" and "construct-a-map" techniques. More specifically, we examined whether (a) the two mapping techniques can be considered equivalent, (b) fill-in-the-map scores are sensitive to the nodes (concepts) selected to be filled in (construct-a-map scores have proven not to be sensitive to the sample of concepts used; Ruiz-Primo et al., 1996), and (c) fill-in-the-map scores are sensitive to the linking lines selected to be filled in (linking words).

Method

Participants. One hundred fifty-two high school chemistry students and two chemistry teachers participated in the study. Students were in one of seven chemistry classes. Four of the classes were considered advanced; the remainder

(56 students) were regular chemistry classes. Two of the four advanced classes were taught by Teacher 1 (six years of teaching experience) and the other two by Teacher 2 (one year of teaching experience). The three regular classes were taught by Teacher 1. All participants were drawn from the Palo Alto, CA, area.

Students and teachers were trained to construct concept maps, including the fill-in-the-map technique, with the same 50-minute training program used in previous studies (see Ruiz-Primo et al., 1996, 1997). To evaluate the training, 25% of the maps constructed by students at the end of the training session were randomly sampled and analyzed. The analysis focused on whether students used the concepts provided on the list, labeled the lines, and provided accurate propositions. Results indicated that 92% of the students used all the concepts provided in the list; all used labeled lines; and all provided four or more accurate propositions. We concluded that the program succeeded in teaching students to construct concept maps.

Design. To evaluate whether the fill-in-the-map scores were sensitive to the sample of nodes or linking lines to be filled in, we used a 2 x 2, concept sample by linking-line sample design. Four 20-node skeleton maps were constructed. In two of the maps 12 nodes (60% of the nodes) were left blank. In the other two skeleton maps, 12 linking lines (31.5% of the linking lines in the criterion map) were left blank (i.e., no linking words). Concepts and linking lines to be left blank were randomly selected from the list of key concepts and the list of propositions in a criterion map. The four skeleton maps were as follows: A—skeleton map with Sample 1 of nodes left blank; B—skeleton map with Sample 2 of nodes left blank; C—skeleton map with Sample 1 of linking lines left blank; and D—skeleton map with Sample 2 of linking lines left blank.

Students were tested on three occasions. On Occasion 1, all students constructed a concept map from scratch using all 20 concepts provided by the assessor. On Occasion 2, half the students filled in skeleton map A and half filled in skeleton map B. On Occasion 3 half the students filled in skeleton map C and half filled in skeleton map D.

Within each of the 7 classes (groups) students were randomly assigned to one of four sequences of skeleton maps: Sequence 1—skeleton map A followed by skeleton map C; Sequence 2—skeleton map A followed by skeleton map D;

Sequence 3—skeleton map B followed by skeleton map C; and Sequence 4—skeleton map B followed by skeleton map D.

Selection of concepts and development of the criterion/skeleton map. To identify the structure of the skeleton map for the fill-in mapping technique, we assumed that (a) there is some "agreed-upon organization" that best reflects the structure of a content domain, (b) "experts" in that domain (in this context, teachers) have a high degree of agreement, and (c) experts' concept maps provide a reasonable representation of the subject domain (e.g., Glaser, in press). Therefore, the skeleton maps were based on the criterion map.

We used the topic "Chemical Names and Formulas" as the domain for sampling the concepts used in the study.[2] Teachers and researchers (the second author was a high school chemistry teacher for 10 years) were involved in the process of selecting the concepts and creating the criterion map. Teachers were asked to identify the concepts they considered to be the most important in the unit. Researchers also selected the most important concepts by carefully reviewing the text used to teach the topic. Figure 2 describes briefly the procedure followed to select the concepts and to define the criterion map (see Ruiz-Primo et al., 1996).

Procedure Used To Construct a Criterion Map

1. Ask each participant to provide a list of the 20 most important concepts in the subject domain.
2. Have participants compare and discuss their lists of selected concepts until consensus is reached about which are the most important concepts. This will be considered the "Key-Concept List."
3. Ask each participant to construct a concept map with the key concepts.
4. Construct a concept map with relations that appear in at least 80% of the participants' concept maps.
5. Discuss and modify the resulting concept map with participants until consensus is reached about which relations should be present in the map.
6. Use the resulting concept map as the "Criterion Map."

Figure 2. Procedure followed to define the Key-Concept List and the Criterion Map.

[2] Although we used this topic in previous studies, the selection of concepts for mapping was carried out again since different teachers participated on this occasion.

The "agreed"-upon links across teachers' and researchers' maps were represented in the *criterion map* and considered the "substantial" links that students were expected to know after instruction on the topic. The criterion map was used as the master map for the purpose of constructing the four skeleton maps. The concepts selected for the blank nodes on the skeleton maps were randomly sampled from the key-concept list. The linking lines selected to be filled in on the skeleton maps were sampled from the linking lines on the criterion map. The propositions provided in the skeleton maps were taken from the criterion map. The concepts for the construct-a-map technique were all those on the key-concept list.

Instrumentation. The two mapping techniques varied in their task demands and constraints imposed on students. Figure 3 provides a profile of the directedness of the assessment tasks for both techniques. The construct-a-map technique asked students to construct a map using the 20 concepts provided by the assessor. Students were encouraged to provide propositions (linking words) as specific as they wanted in order to explain the relationship between the two concepts they were linking. No restriction was imposed on the type of structure students could use in the map (e.g., students were not instructed to create a hierarchical structure).

The fill-in-the-map technique asked students to fill in two skeleton maps, one with blank nodes and the other with blank linking lines. After randomly selecting nodes, only six nodes were different between skeleton map A and skeleton map B. For the blank-linking-line maps, only one proposition was the same across skeleton map C and skeleton map D. Students' responses on each skeleton map were scored as correct or incorrect. A maximum of 12 points could be awarded to each student on each skeleton map.

Technique	Concepts	Linking lines	Linking words	Structure of map
Construct-a-map	Provided	Not provided	Not provided	Not provided
Fill-in-the-map	Provided	Provided	Provided	Provided

Figure 3. Directedness profile of two mapping techniques: *Construct-a-map* and *Fill-in-the-map*.

As in previous studies, to score students' constructed maps we developed a *proposition inventory* to account for variation in the quality of the students' propositions. This inventory contained the 190 possible relations between a specific pair of concepts in the key-concept list. Based on this inventory, each proposition was scored on a 5-point scale, from 0 for inaccurate/incorrect to 4 for excellent/outstanding (complete proposition that showed deep understanding of the relation between two concepts; see Ruiz-Primo et al., 1996, for a definition of each category). The maximum score for a map constructed by a student was based on the criterion map: the number of links (38) in the criterion map was multiplied by 4 (all propositions were scored as excellent).

After constructing the concept maps, all classes received a 30-item multiple-choice test on "Chemical Names and Formulas" designed by the teachers and the researchers. The internal consistency of the test was .74.

Results

In this study we asked the following questions: (a) Are scores based on fill-in-the-node skeleton maps equivalent to scores on the fill-in-the-linking line skeleton maps? (b) Are fill-in-the-map scores sensitive to the sample of nodes or the linking lines to be filled in? (c) Does the fill-in-the-map technique provide the same picture of a student's knowledge structure as the construct-a-map technique?

Before focusing on these questions, one preliminary issue needs to be addressed, the planned contrast between advanced and regular chemistry students. When we compared the multiple-choice test scores for the seven classes, only Class 6 differed significantly the other classes (viz. Classes 2 and 4). Consequently, we decided to collapse the seven classes and present overall results for simplicity and brevity.

Equivalence of Types of Fill-In Maps

We planned the following steps in examining the equivalence of scores from the two types of fill-in maps. First, we would examine the equivalence of the scores from the two forms of each type of map, each form being created by randomly leaving blank a set of nodes or lines. If the forms were found to be equivalent, we would then examine the equivalence of the fill-in-the node map scores with the fill-in-the-line map scores. For scores from two forms of a

skeleton map, or from two different types of skeleton maps to be considered equivalent, their means, variances, and covariances (correlations) with each other and an outside criterion (e.g., multiple-choice scores) should be equal (within sampling error).

Equivalence of forms: Node maps. Means, standard deviations, and correlations between scores on the node map and the multiple-choice test are presented in Table 1.[3] The level of students' performance across the two samples was high, and close to the maximum possible score. (Indeed, this may give rise to range restriction, a topic addressed later in the paper.) Independent-sample *t*-tests indicated no significant difference between the means on the two node-map forms ($t = 1.57$, $p = .12$). An F_{Max} test also indicated no significant difference between the variances of the two forms ($F_{Max} = 1.50$, $p > .05$). The magnitude of the correlations with multiple-choice test scores across the two samples was virtually the same. We concluded that the two forms were parallel.

Equivalence of forms: Line maps. Descriptive statistics for the two forms of the linking-line maps are also presented in Table 1. Results were similar to those found for the two forms of the node map. No significant differences were found between means of the two forms ($t = 1.65$, $p = .10$) or the variances ($F_{Max} = 1.50$, $p > .05$). The magnitude of the correlations with multiple-choice scores was also the same. We concluded that the two forms of the linking-line map were parallel.

Table 1

Means and Standard Deviations by Type of Skeleton Map and Sample

Type of skeleton map	*n*	Mean (max. = 12)	*SD*	Correlation with multiple-choice test
Fill-in-the-nodes				
Sample 1	80	11.21	1.43	.37
Sample 2	72	10.80	1.74	.38
Fill-in-the-linking-lines				
Sample 1	78	9.77	2.74	.65
Sample 2	73	8.99	3.09	.66

[3] Note that students were randomly assigned to complete node map A (Sample 1) or B (Sample 2). Consequently, we cannot correlate scores between the two forms of the same map. We can correlate scores from each node map with scores from the multiple-choice test.

Equivalence of skeleton-map types: Nodes and lines. To examine the equivalence of the node- and line-map types, we carried out a 2 x 4, Type x Sequence, split-plot ANOVA. This analysis provides information on the equivalence of means across the two types, as well as on whether the particular sequence of maps (e.g., node 1-line 1 vs. node 2-line 2) influenced the pattern of mean scores. We also examined differences in variance between the node and the equivalence of the covariances across the four sequences (i.e., the covariance of the node map scores with line map scores).

Table 2 provides the means, standard deviations, and correlations with multiple-choice scores for each type and sequence. Overall, students' performance across the two types of skeleton maps was high. However, students performed higher on node maps than on the linking-line maps ($F_T = 65.95$, $p = .000$). The sequence effect was not statistically significant ($F_S = .63$; $p = .599$).

ANOVA results also indicated a significant interaction between technique (T) and sequence (S) ($F_{TxS} = 2.73$, $p = .046$). An examination of the interaction showed that it was ordinal. The mean difference in scores between node and linking-line maps was not significant for those students in Sequence 3 ($F_{S_3} = 3.73$, $p = .055$), whereas the mean difference was statistically significant for those students in the other three sequences ($F_{S_1} = 13.49$, $p = .000$; $F_{S_2} = 24.66$, $p = .000$; $F_{S_4} = 32.53$, $p = .000$). Filling-in-the-nodes using sample 2 for the skeleton map somehow facilitated the fill-in-the-linking-lines when sample 1 was used for the skeleton map. A closer look into the skeleton maps revealed that the number of

Table 2

Means, Standard Deviations and Correlations With Multiple-Choice Test by Type of Skeleton Map and Sequence

			Fill-in-the-node		Fill-in-the-linking line		Correlations with multiple-choice	
	Sequence	*n*	Mean	*SD*	Mean	*SD*	Mean	*SD*
1	Nodes 1-Lines 1	43	11.09	1.52	9.72	2.84	.40	.69
2	Nodes 1-Lines 2	36	11.03	1.33	9.31	3.06	.35	.66
3	Nodes 2-Lines 1	35	10.63	1.81	9.83	2.65	.52	.61
4	Nodes 2-Lines 2	37	10.97	1.67	8.68	3.13	.29	.69
Total		152	11.02	1.59	9.39	2.93	.37	.65

propositions students needed to read to fill in the nodes in skeleton map B that overlapped with the linking lines they needed to filled in on skeleton map C was higher than the number observed in any other sequence. We conclude, then, that the two types, generally, were not equivalent in their mean scores; the node technique producing systematically higher scores than the line type.

The observed difference between the variances of the node and line technique scores was sizable: 2.54 and 8.60, respectively. Clearly, the variances from the two techniques are not equivalent. Box's test of equality for covariance matrices across the four sequences indicated equal covariances between the node and linking line scores across the four sequences (Box's M = 13.02, p = .177; F_{Max} = .63; F_{Max} = .68; F_{Max} = .50; F_{Max} = .43). Ignoring sequence, the overall correlation between fill-in-the-nodes and fill-in-the-linking-lines map scores was .52 (p = .01).

Finally, the pattern of correlations between map scores with multiple-choice scores differs by technique (see Table 1). Higher correlations are observed between the line and multiple-choice scores than between node and multiple-choice scores. The moderate correlation suggests that students were ranked somewhat differently across the two types of maps. It is important to mention that the magnitude of the correlations that involve the node scores may be lowered due to the restriction of range observed in the fill-in-the-nodes maps.

In summary, we concluded that the two randomly constructed forms of each skeleton map type (node and line) were equivalent. However, the types themselves did not produce equivalent scores. More specifically, the node map produced higher scores with lower variances than did the line map; the correlation between node scores and multiple-choice scores was lower than between line scores and multiple-choice scores.

Comparing Mapping Techniques

In this section we compare the two mapping techniques, fill-in-the-map and construct-a-map. First, we examine the consistency of scores across raters for the construct-a-map technique. Then we characterize students' constructed maps, and compare the two techniques. Finally, we compare scores from the two mapping techniques with multiple-choice scores.

Interrater reliability. All construct-a-maps were scored for accuracy and comprehensiveness. For each student we calculated a *propositional accuracy*

score—the sum of the (0-4) scores obtained on all propositions; a *convergence* score—the proportion of accurate propositions in a student's maps out of all possible propositions in the criterion map; and a *salience* score—the proportion of valid propositions out of all the propositions in the student's map.

A sample of 55 student maps (more than a third of the total sample) were scored by three raters. To examine the generalizability of scores across raters, three person (p) by rater (r) G studies were carried out, one for each type of score (Table 3).

Results indicated that the error introduced by raters was negligible. Both relative ($\hat{\rho}^2$) and absolute ($\hat{\phi}$) coefficients were very high across types of scores. Based on these results, the remaining 97 concept maps were randomly distributed among the three raters and only one rater scored each map. The randomization was done within each of the seven classes. Thus, all three raters scored a sample of students' maps across the seven classes.

Students' maps. Table 4 provides information about the characteristics of students' constructed maps. Two thirds of the students used all 20 concepts provided in the list to construct their maps. Another fifth used 18-19 concepts, and only one student used just 14 concepts.

Table 3

Estimated Variance Components and Generalizability Coefficients for a Person by Rater G Study Across Types of Scores

	Proposition accuracy		Score type convergence		Salience	
Source of variation	Estimated variance component	Percent of total variability	Estimated variance component	Percent of total variability	Estimated variance component	Percent of total variability
Persons (p)	290.54	96.26	0.03114	97.65	0.02863	95.15
Raters (r)	0.36	0.12	0.00011	0.34	0.00020	.66
pr,e	10.92	3.62	0.00064	2.00	0.00126	4.19
$\hat{\rho}^2$	.99		.99		.98	
$\hat{\phi}$	.99		.99		.98	

Table 4

Means and Standard Deviations of Students' Concept Map Components

Map components	*n*	Mean	*SD*	Min	Max
Nodes in the map	152	19.34	1.23	14	20
Linking lines	152	25.41	6.60	14	43
Accurate propositions	152	18.88	7.44	0	42

A surprising finding was that 6.6% of the students provided more than 38 links in their maps, which is the number of links on the criterion map.[3] Furthermore, 40% of these students provided more than 38 accurate propositions.

It is important to mention that a few of the students provided better propositions than those in the criterion map! This led us to re-score the criterion map using the same criteria applied for students. Therefore, some propositions in the criterion map became "Good" instead of "Excellent," and one proposition became "Poor." The original maximum was 158 and was corrected to 135.

Students' scores across assessment techniques. Table 5 provides the descriptive statistics for the three types of assessments administered to the students: construct-a-map, fill-in-the-map, and multiple-choice test.

Table 5

Means and Standard Deviations Across the Three Types of Assessments Administered to Students

Assessment	*n*	Max	Mean	*SD*
Construct-a-map				
Proposition accuracy	152	135	53.91	22.17
Convergence	152	1	.50	.19
Salience	152	1	.73	.17
Fill-in				
Fill-in-the-nodes	152	12	11.02	1.59
Fill-in-the-linking-lines	151	12	9.39	2.93
Multiple-choice test	150	30	24.05	3.74

[3] In fact, 18% of students provided between 25 and 38 links.

Mean scores across the forms of assessments do not provide the same picture of students' knowledge. Whereas salience, fill-in-the-map and multiple-choice scores indicated that students' performance was close to the maximum criterion, the proposition accuracy and convergence scores indicate that students' knowledge was rather partial compared to the criterion map.

All types of scores, except proposition accuracy and convergence, showed negatively skewed distributions (skewness value ranged from -.755 for fill-in-the-linking-lines, to -1.538 for fill-in-the-nodes) indicating that most of the students obtained high scores. Furthermore, the Kolmogorov-Smirnov normality test confirmed that only proposition accuracy and convergence scores were normally distributed (p = .200). It seems that proposition accuracy and convergence scores better reflect the differences in students' knowledge than the other scores.

A correlational approach was used to compare techniques because of the different score scales across techniques. Table 6 provides a multiscore-multitechnique matrix. We first focus on comparing scores within each mapping

Table 6

Multiscore-Multitechnique Matrix

	Construct-a-map			Fill-in-the-map		
	PA	CON	SAL	NOD	LIN	MC
Construct-a-map						
Proposition-accuracy (PA)	(.99)[a]					
Convergence (CON)	.95	(.99)[a]				
Salience (SAL)	.73	.75	(.98)[a]			
Fill-in-the-map						
Fill-in-the-nodes (NOD)						
Observed	.50	.47	.45	(.70)[b]		
Corrected	.61	.56	.54			
Fill-in-the-lines (LIN)						
Observed	.51	.44	.40	.53	(.84)[b]	
Corrected	.56	.49	.44	.69		
Multiple-choice (MC)						
Observed	.51	.44	.46	.37	.65	(.74)[c]
Corrected	.60	.51	.54	.51	.83	

[a] Interrater reliability.
[b] Internal consistency averaged between the two skeleton maps.
[c] Internal consistency.

technique. Then, we evaluate the extent to which the scores on the two mapping techniques converge, and finally, we evaluate the extent to which the two mapping technique scores converge with multiple-choice scores.

In the matrix, reliability coefficients are enclosed in parentheses on the diagonal. Along with the observed correlations, we present correlations corrected for unreliability.[4] However, because different reliability estimates are used in the matrix, and hence error measurement is defined differently, some of these corrections may not be accurate and must be interpreted cautiously. Therefore, we focus on the observed correlations.

Construct-a-map scores. The correlation between proposition accuracy and convergence scores is very similar to correlations we have found in other studies (e.g., Ruiz-Primo et al., 1996, 1997). This very high correlation suggests that both scores rank students similarly. Furthermore, when G theory has been used to evaluate the dependability of these measures (see Ruiz-Primo et al., 1996, 1997), we found that the percent of variability among persons is higher for proposition accuracy and convergence scores than for salience scores. This indicates that these two measures better reflect the differences in students' knowledge structures than do salience scores.

The correlations between proposition accuracy and convergence scores with salience scores (.73 and .75 respectively), however, are lower than the ones we have observed before (~.85). A possible reason for this lower correlation may be students' knowledge level. In this study, students clearly had better knowledge about the topic than students we tested before. The means obtained in this study were impressively higher when compared with those we obtained before (Proposition Accuracy = ~11; Convergence = ~.17; Salience = ~.50). Students in this study provided more accurate propositions in their maps, thereby improving their salience scores. In our previous studies, students' scores were low across types of scores so their ranking did not differ across scores.

The general conclusion about construct-a-map scores is consistent with our previous research. Proposition accuracy and convergence scores reflect the differences in students' knowledge structure better than do salience scores. Based on practical (e.g., scoring time) and technical (e.g., instability of scores) arguments, we conclude that the convergence score is the most efficient.

[4] No correction needed for the construct-a-map technique.

Mapping technique scores. If the construct-a-map and fill-in techniques measure the same construct, we should expect a high correlation among these scores. Yet, correlations were lower than expected (r = 46) averaged across types of scores). Restriction of range observed in both types of fill-in-the-map scores may contribute to the magnitude of the correlations; interpretation of the low coefficients should be considered with caution.

Although correlations between fill-in-the-nodes and fill-in-the-linking-lines with construct-a-map scores were not of the same magnitude (correlations are higher between fill-in-the-node and construct-a-map), no significant difference ($p > .05$) was found among the correlations ($\chi^2_{(5)} = 3.23$; Meng, Rosenthal & Rubin, 1992). The pattern of correlations, however, is the same: The highest correlation is with proposition accuracy and the lowest with salience scores. The magnitude of the correlations across columns in Table 6 indicates that students are ranked differently according to the technique used. It seems that different aspects of the students' connected understanding are being tapped with the construct-a-map technique and the fill-in-the-map technique.

Mapping and multiple-choice scores. Correlations between multiple-choice test scores and each type of construct-a-map scores are similar to the ones observed between the fill-in-the-map and construct-a-map scores. No significant difference was observed among the nine correlations ($\chi^2_{(8)} = .161$; $p > .05$). We concluded that construct-a-map scores correlated similarly with fill-in-the-map and multiple-choice scores, on average of r = .37

The correlations between fill-in-the-map scores with multiple-choice scores were quite surprising. The magnitude of the correlations between fill-in-the-nodes and multiple-choice test reported by Schau et al. (1997) is higher (r = .75 on average) than the one we found in this study (r = .37).

Two issues may explain these differences: restriction of range observed in the fill-in-the-nodes skeleton map scores (i.e., skeleton map was very easy for students in our study) and differences between the characteristics of the fill-in-the-nodes maps used in both studies (e.g., Schau et al., 1997, used 37 nodes; 50% were left blank; we used 20, and 60% were left blank.) Also, the propositions in the skeleton map used by Schau et al. were less complex than the ones used in ours. Whether the characteristics of the maps can affect students' scores is a topic that deserves to be studied more carefully. For example, what number of nodes

in a skeleton map is optimum? How many nodes need to be left blank? What is the best way to select the nodes left blank?

The correlation between scores on the fill-in-the-linking-line and the multiple-choice tests is the highest among all the correlations with mapping indicating that about 43% of the variance on these measures is shared, whereas only 14% is shared with the fill-in-the-node scores, due to restriction of range. When the correlation is corrected for restriction of range, the magnitude of the correlation is estimated to be .59 (36% shared variance). Hence, the apparent difference in node and line score correlation with multiple-choice may be due largely to range restriction in the node scores.

We think that the construct-a-map technique better reflects the state of students' knowledge structure. We based this conclusion on the fact that this technique is the only one that accurately reflects the differences among students' scores. But, what is the fill-in-the-map technique tapping? What aspect of the students' knowledge is being measured with this form of assessment? A closer look at the cognitive activities displayed in this technique is needed. Talk-aloud protocols may help to better define the cognitive activities reflected by both techniques.

An overall conclusion is that we need to invest time and resources in finding out more about what aspects of students' knowledge are tapped by different forms of assessment. What makes those assessments share variance? What is the unique variance? Which technique should be considered the most appropriate for large-scale assessment? Practical issues, though, cannot be the only criteria for selection. Students' partial knowledge may be hidden more easily on some forms of assessment than on others. To resolve the issue of what is being measured with these different techniques, we need information about the cognitive activity displayed on each of them.

Conclusion

In this study we explored the equivalence of two mapping techniques, fill-in-the-map and construct-a-map from scratch. We examined whether (a) skeleton map scores were sensitive to the sample of nodes or linking lines left blank, (b) the two forms of skeleton maps were equivalent, and (c) the two

mapping techniques provided similar information about students' connected understanding.

Our results led to the following tentative conclusions: (a) Skeleton map scores are not sensitive to the sample of concepts or linking lines to be filled in. Probably the list of concepts and propositions was cohesive enough so that any combination of concepts or propositions could provide similar information about students' knowledge. (b) Fill-in-the-node and fill-in-the-linking-line techniques are not equivalent forms of fill-in-the-map. Further research is needed to define which of these two forms provides the most accurate information about students' knowledge or connected understanding. (c) The relationship between the two mapping techniques studied suggests that both mapping techniques are tapping somewhat similar, but not identical, aspects of students' connected understanding. As we previously suggested, talk-aloud protocols may provide insight about the cognitive activities involved in constructing and filling in a map. (d) Construct-a-map scores most accurately reflected the differences across students' knowledge structure. (e) The relationship between scores from the multiple-choice test and both mapping techniques confirmed that mapping techniques were not equivalent. The pattern of correlation coefficients was different across mapping techniques. (f) Convergence scores—the proportion of accurate propositions in the students' maps to the number of all possible propositions in the criterion map—is the most efficient indicator when scoring construct-a-map concept maps.

References

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

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist*, *31*, 133-140.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-39). Englewood Cliffs, NJ: Prentice Hall.

Glaser, R. (in press). Changing the agency for learning: Acquiring expert performance. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the art, sciences, sports, and games*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (1997). *Teaching science for understanding*. San Diego: Academic Press.

Ruiz-Primo, M. A., Schultz, S. E., & Shavelson, R. J. (1996, April). *Concept-map based assessment in science: An exploratory study*. Paper presented at the annual meeting of the American Educational Research Association, New York.

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

Ruiz-Primo, M. A., Shavelson, R. J., & Schultz, S. E. (1997, March). *On the validity of concept map based assessment interpretations: An experiment testing the assumption of hierarchical concept-maps in science*. Paper

presented at the annual meeting of the American Educational Research Association, Chicago.

Schau, C., & Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *The American Statistician*, *51*, 171-175.

Schau, C., Mattern, N., Weber, R., Minnick, K., & Witt (1997, March). *Use of fill-in concept maps to assess middle school students' connected understanding of science*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Shavelson, R. J. (1974). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching*, *11*, 231-249.