On the Cognitive Validity of Interpretations of Scores From Alternative Concept Mapping Techniques

CSE Technical Report 503

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

June 1999

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 Richard J. Shavelson, Project Director CRESST/Stanford University

Copyright © 1999 The Regents of the University of California

The work reported herein was supported 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.

ON THE COGNITIVE VALIDITY OF INTERPRETATIONS OF SCORES FROM ALTERNATIVE CONCEPT MAPPING TECHNIQUES[1]

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

Abstract

The emergence of alternative forms of achievement assessments and the corresponding claims that they measure "higher order thinking" have significantly increased the need to examine their cognitive validity (Glaser & Baxter, 1997; Linn, Baker, & Dunbar, 1991). This study evaluates the validity of connected understanding interpretation of three mapping techniques. The study focused on the correspondence between mapping-intended task demands, inferred cognitive activities, and scores obtained. We analyzed subjects' concurrent and retrospective verbalizations at different levels of competency performing the mapping tasks and compared the directedness of the mapping tasks, the characteristics of verbalization, and the scores obtained across techniques. Our results led to the following general conclusions: (a) Consistent with a previous study, we found that the three mapping techniques provided different pictures of students' knowledge. (b) Inferred cognitive activities across assessment tasks were different and corresponded to the directedness of the assessment task. The low-directed technique seemed to provide students with more opportunities to reflect their actual conceptual understanding than the high-directed techniques.

The emergence of alternative forms of achievement assessment and the corresponding claims that they measure "higher order thinking" have significantly increased the need to examine their cognitive validity (Glaser & Baxter, 1997; Linn, Baker, & Dunbar, 1991). To address cognitive validity, evidence is sought about correspondence between intended task demands and the cognitive activity evoked, as well as the correspondence between quality of cognitive activity and performance scores (Glaser & Baxter, 1997). This study provides evidence bearing on the cognitive validity of an alternative assessment in science, concept maps. More specifically, this paper provides evidence about differences and similarities in the cognitive activity observed across various forms of assessments (i.e., three mapping techniques) with students of different levels of performance.

[1] The authors want to thank Tamara Danoyan for her help in transcribing subjects' verbal protocols.

Concept maps have been used to assess students' knowledge structures, especially in science education (e.g., Rice, Ryan, & Samson, 1998; Ruiz-Primo & Shavelson, 1996; White & Gunstone, 1992). This form of assessment is based on theory and research showing that understanding in a subject domain such as science is associated with a rich set of relations among important concepts in the domain (e.g., Baxter, Elder, & Glaser, 1996; Chi, Glaser, & Farr, 1988; Glaser, 1991; Mintzes, Wandersee, & Novak, 1997). With this form of assessment 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. The combination of two concepts and the explanation of the relationship between them is called a proposition—the fundamental unit of a concept map.

We have shown that mapping techniques vary widely (Ruiz-Primo & Shavelson, 1996), and we have evaluated the technical characteristics of some of these techniques (Ruiz-Primo, Schultz, Li, & Shavelson, 1998; Ruiz-Primo, Schultz, & Shavelson, 1996; Ruiz-Primo, Shavelson, & Schultz, 1997). We suspected (Ruiz-Primo & Shavelson, 1996) that different mapping techniques imposed different cognitive demands on students. This claim was based on the characteristics of the mapping tasks, that is, the constraints imposed on a student in eliciting a representation of her knowledge structure.

In a previous study (Ruiz-Primo et al., 1998), we tested this hypothesis by comparing students' performance across three different mapping techniques. The construct-a-map technique asked students to construct a map using 20 concepts provided by the assessors. The fill-in-the-nodes and fill-in-the-linking-lines techniques provided students with the structure of the map, including some blank nodes or linking lines, and students were asked to fill in the map using the provided concepts or words explaining the relations.

We found that the fill-in-the-map technique (i.e., fill-in-the-nodes and fill-in-the-linking-lines) and the construct-a-map-from-scratch technique led to different interpretations about students' knowledge of a topic (Ruiz-Primo et al., 1998). Whereas the scores obtained under the fill-in-the-map technique indicated that students' performance was close to the maximum criterion, the scores obtained with the construct-a-map technique revealed that students' knowledge was incomplete compared to a criterion map. Furthermore, the construct-a-map technique provided a symmetric distribution of scores, whereas scores from the fill-in-the-map technique were negatively skewed. We concluded that the construct-a-map technique better reflected differences in students' knowledge.

On the Validity of Mapping Scores

Our conclusion that fill-in mapping techniques lead to different interpretations about students' knowledge reflects the fact that some characteristics of the assessment tasks permit students to respond correctly or appropriately in ways that are irrelevant to the construct assessed (i.e., students' connected understanding).[2]

One way to evaluate this source of invalidity is to compare the different assessment tasks on a set of dimensions: (a) intended task demands, (b) inferred cognitive activities that underlie a task, and (c) scores obtained (e.g., Glaser & Baxter, 1997).

Task demands. One dimension that can be used to characterize concept map assessment task demands is *directedness*. In a previous paper (Ruiz-Primo et al., 1998) we proposed the use of "directedness" as a dimension that represents differences in the constraints imposed on students by different mapping techniques. We characterized concept map techniques as high- or low-directed according to the information provided to the students (Figure 1).

If the characteristics of the concept map assessment task fall on the left extreme (high-directed), a student's representations would probably be determined more by the mapping technique than by the student's own knowledge or connected

	Degree of Directedness	
	High ◄──────►	**Low**
Concepts	Provided by Assessor	Provided by Student
Linking Lines	Provided by Assessor	Provided by Student
Linking Words	Provided by Assessor	Provided by Student
Structure of the Map	Provided by Assessor	Provided by Student

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

[2] This has been called *construct-irrelevant variance,* a source of invalidity (e.g., Messick, 1995).

understanding.[3] In contrast, if the assessment task falls on the right extreme (low-directed), students are free to decide which and how many concepts to include in their maps, which concepts are related, and which words to use to explain the relation between the concepts.

We reasoned that the task demands imposed on students by high- and low-directed techniques are different since more informed decisions need to be made by the student in a low-directed technique. That is, the cognitive demands of a low-directed technique seem to be higher than those of a high-directed technique. Using Baxter and Glaser's (1998) terms, low-directed techniques involve *rich content process* and high-directed techniques involve *lean content process*.

Inferred cognitive activities. A second dimension on which to compare mapping techniques is the inferred cognitive activities and their correspondence with the intended task demands. If task demands for high- and low-directed mapping techniques are different, these differences should be reflected in the cognitive activities displayed by students while performing the tasks. Furthermore, inferred cognitive activity should differ across students of different levels of competence (e.g., Baxter et al., 1996; Chi et al., 1988; Glaser, 1991) if mapping assessment tasks differentiate among levels of connected understanding.

Research on expertise (e.g., Chi et al., 1988; Ericsson & Smith, 1991; Glaser, 1991) provides an informative framework for distinguishing important characteristics of experts and novices. In the context of education, expertise is translated as competence in a content domain. Baxter, Elder, and Glaser (1996) characterized competent students in science classrooms as those who (a) provide coherent, content-based explanations rather than descriptions of superficial features or single statements of facts, (b) generate a plan for a solution, (c) implement solution strategies that reflect relevant goals and subgoals, and (d) monitor their actions and flexibly adjust their approach. These characteristics can be used as a framework to guide the analysis of the cognitive activities displayed by students of different competence in a domain. Students' verbal reports while performing a task provide access to these cognitive activities.

[3] 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 such a 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.

Verbal reports have been recognized as a major source of evidence on subjects' cognitive processes in specific tasks (e.g., Ericsson & Simon, 1993). Collecting verbal reports has become a standard method in, for example, validating multiple-choice tests (e.g., Levine, 1998; Norris, 1991) or performance assessments (e.g., Baxter & Glaser, 1998; Glaser & Baxter, 1997). We used students' talk-aloud protocols to examine the cognitive activities students displayed while performing concept mapping assessment tasks.

Observed scores. The third dimension used to compare the three mapping assessment techniques is the correspondence of the score obtained with the cognitive activity observed and the intended task demand.

The way students were scored across techniques varied according to the characteristics of the assessment tasks (see footnote 2). For the construct-a-map technique—a low-directed technique—the nature of the students' responses (i.e., explanation of the relationships between concepts) allows the assessor to score the quality of the propositions provided (see Ruiz-Primo et al., 1996, 1997), whereas for the fill-in-the-map technique—a high-directed technique—students' responses can be scored only as correct or incorrect.[4]

In this study we compared the three concept mapping assessment techniques using these three dimensions. Evidence about the correspondence between intended task demands, inferred cognitive activity, and scores obtained is provided for the three mapping techniques. We used talk-aloud protocol analysis to examine students' performance to ascertain (a) whether these assessment techniques imposed different cognitive demands, and (b) whether cognitive activities varied qualitatively between more and less proficient students within a particular assessment.

Method

Subjects

Twelve students were selected from a larger sample (*N* = 152) in a previous study (Ruiz-Primo et al., 1998). Two chemistry teachers who participated in that study also participated in this study.

[4] For construct-a-map, it is also possible to score propositions as correct or incorrect (see Ruiz-Primo et al., 1996, 1997).

Students who participated in the previous study were first classified into one of three groups—*high* scorers (top students), *medium* scorers (students closest to mean), and *low* scorers (low students)—based on their low-directed map scores (i.e., construct-a-map-from-scratch) since scores from the high-directed maps (i.e., fill-in-the-nodes and fill-in-the-linking-lines) varied little. Three students from each group were selected. (If more than one student had the same score, one was randomly selected.)

Procedure

The three mapping techniques and the topic, Chemical Names and Formulas, were the same as those used in the previous study (Ruiz-Primo et al., 1998): (a) Construct-a-map-from-scratch—students were asked to construct a map using the 20 concepts provided; (b) Fill-in-the-nodes—students were asked to fill in a 12-blank-node skeleton map with the correct concepts provided; and (c) Fill-in-the-linking-lines—students were asked to fill in a 12-blank-line skeleton map with a description of the relationship provided of each pair of connected concepts. Also a 30-item multiple-choice test used in the previous study was re-administered in this study. Maps were scored using the same criteria as before. Students' constructed maps were scored for the quality of the propositions (0 for inaccurate/incorrect to 4 for excellent proposition). Students' responses on each skeleton map and the multiple-choice test were scored as correct or incorrect.

Students and teachers constructed three concept maps and answered multiple-choice questions. All participants were asked to "think aloud" (concurrent verbalization) as they engaged in the different assessments. Students were not asked to provide reasons or explanations for their answers or choices; they were instructed just to think aloud as they were performing the tasks. After they finished each assessment, they were asked to describe, retrospectively, the strategies used (retrospective verbalization). Instructions given to participants for concurrent and retrospective verbalizations were those recommended by Ericsson and Simon (1993).

Students were tested six months after the original study in two sessions. In the first session, students were trained in and practiced talking aloud, reminded about how to construct concept maps, and asked to do the first two assessment tasks: construct-a-map-from-scratch and fill-in-the-nodes map. In the second session, students were reminded to talk aloud and asked to do the last two assessment tasks: fill-in-the-linking-lines map and the multiple-choice test. Teachers did all assessments in one session. Unfortunately, two students from the medium-level group did not come to the second session. Data and verbal protocols for those students are available only for the construct-the-map and fill-in-the-nodes techniques.

Verbal Analysis Coding

To evaluate the nature and quality of cognitive activity, we developed a system that taps students' verbalizations at two levels. The *fine-grain level* includes a set of coding categories for classifying the propositions/sentences/ideas students provided while performing the mapping assessment tasks (detailed coding). The *high-level* categories focus on the entire protocol and were used to describe planning and strategies students used to address the assessment tasks based on the sequence of events reported on the protocols. The system does not attempt to deal with all the verbalizations in the protocol; the coding categories were developed considering two issues: aspects of cognitive activity that previous research had found to be useful in comparing experts and novices (e.g., Baxter et al., 1996; Baxter & Glaser, 1998), and aspects of cognitive activity that reflect a difference in the demands imposed by the assessment tasks.

Our fine-grain-level coding categories include four types of statements: *Explanation*—information representing coherent and correct construction or elaboration of a student's response, choice, or alternative (e.g., "It has to be acid, because acids have H^+ cation"); *Monitoring*—information representing evaluations of a student's own strategies and actions for adjusting purposes (e.g., "This doesn't make sense," "I need to cross out the concepts to know what I already used"); *Conceptual Errors*—information representing misconceptions/alternative conceptions (e.g., "Molecular compounds are formed by anions"); and *No Code Applicable*—information that does not provide any insight into a student's cognitive activity (e.g., mechanical information: "I will draw a circle, write the concept inside and draw a line to the other circle").

The high-level categories in the system are *Planning*—information representing a plan, a sequence of possible steps, to approach the task (e.g., "I will read the concepts and select the most important concept to put it in the center"); and *Strategy*—a description of the solution approaches used by students to address the assessment task. To account for the quality of the explanations and the forms of self-monitoring, we created subcategories within the categories Explanation and Monitoring. Table 1 presents the description of each subcategory.

For the fine-grain-level analysis, students' protocols from concurrent verbalizations across the three mapping techniques were segmented into units. Content of students' verbalizations determined the segment boundaries. For example, if the student was describing the drawing of a pair of concepts, all information related to the same chain-description was considered a verbal unit. Each verbal unit was given a number, and two independent coders coded every unit of the protocols on a coding form (Figure 2).

Table 1

Categories of Verbal Analysis

	Subcategories	Example
Explanation	E.1 Defining Information that provides more details about student's response, choices, or alternatives.	*"Ions are formed when atoms lose or gain electrons…"*
	E.2 Comparing / Contrasting Information that groups / combines OR points out similarities and / or differences between student's responses (concepts, propositions), choices, or alternatives.	*"Two types of ions, cations and anions, each have charges, but anions have negative charge, cations have a positive charge…"*
	E.3 Justifying Information that provides a reason for the student's response, choice, or alternative.	*"N_2O_4 is a molecular compunds because they are both nonmetals."*
Monitoring	M.1 Defining / Applying a Strategy Information that identifies / determines a strategy or its use for performing the task.	*"I need to check out the concepts to know what I already used."*
	M.2 Effective Reflecting Information describing self-checking of students. For example, questioning the meaning of words, relationships, verifying the accuracy of responses, choices, alternatives.	*"I think I am going to change my thing, and draw a line from polyatomic ions to ionic compounds and write 'can form' ionic compounds…"*
	M.3 Ineffective Reflecting Information describing self-checking without any effect on student's performance. For example, making statements about difficulty of task but no strategy to solve the problem, making ambiguous statements.	*"I can't remember exactly what this is…"*
	M.4 Reviewing Quality of Product Information representing statements about making decisions for improving the task product.	*"I'm looking over the map to make sure all my concepts are drawn so that people could see them…"*

Number of Verbal Units by Category and Assessment

Assessment	E.1	E.2	E.3	M.1	M.2	M.3	M.4	CE	NCA	Total UV	% Agr
Construct-a-map											
Fill-in-the-nodes											
Fill-in-the-linking-lines											
Multiple-choice											

Coding by Verbal Unit

VU#	Codes					VU#	Codes					VU#	Codes				
1						41						81					
2						42						82					
3						43						83					
4						44						84					
5						45						85					
.						.						.					
.						.						.					
.						.						.					

Figure 2. Coding form.

For analyzing planning and strategy, the high-level categories, students' protocols from the concurrent and retrospective verbalizations were not segmented. To interpret the results meaningfully, the student's entire verbalization during each assessment was considered the unit of analysis. The two coders independently described the plan and the strategy used by each student on each assessment based on the concurrent verbalization and synthesized each student's description based on the retrospective verbalizations.

Results

The main question that guided our initial analysis was whether there was a correspondence between the intended task demands, the inferred cognitive activities, and the scores obtained. To answer this question we compared mapping techniques according to their directedness, the characteristics of the cognitive activities displayed by students while performing the mapping assessment tasks, and the subjects' scores. We first provide a description of the students' performance across the mapping techniques and then present the analysis of the students' verbal protocols between assessment techniques and within students on the same assessment technique.[5]

Students' Scores Across Mapping Techniques

Students' mean scores and standard deviations across studies and assessments are presented in Table 2. We present the mean scores first for the complete sample in Study 1 and then for the nine students tested in Study 2 on both occasions.[6]

Except for the fill-in-the-nodes technique, mean scores for the sample used in this study were higher than those observed for the complete sample. One possible explanation for this difference may be that students were not selected randomly, except for those students who had exactly the same score. For example, the three top students in the complete sample were selected for this study, but it was not the case for the three low-score students. The lowest scoring students in the complete sample had scores that ranged from 0 to 5. We reasoned that these students probably did not have the necessary content knowledge to approach the assessment tasks, especially for the construct-a-map technique. Thus, the three low-score students selected for this study were the ones with the next lowest scores (range from 20 to 30).

Table 2

Subjects' Mean Scores and Standard Deviations Across the Two Occasions and the Four Assessments

		Study 1				Study 2		Teachers	
Mapping technique	Max	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
		(*n* = 152)		(*n* = 9)		(*n* = 9[a])		(*n* = 2)	
Construct-a-map	135	53.91	22.17	62.63	34.15	55.39	27.75	84.50	23.33
Fill-in-the- nodes	12	11.02	1.59	10.43	2.07	10.29	2.06	12.00	0
Fill-in-the-lines	12	9.39	2.93	10.14	2.97	8.86	2.85	11.00	1.41

[a] For the fill-in-the-map techniques only 7 students on Occasion 2.

[5] In this paper we focus only on the three mapping techniques. The comparison of the cognitive activities displayed on mapping techniques and the multiple-choice tests will be reported elsewhere.

[6] Reliability coefficients for each assessment were calculated in the previous study (Ruiz-Primo et al., 1998). The internal consistency for the fill-in-the-nodes technique was .70 (low coefficient due to restriction of range); for the fill-in-the-links, internal consistency was .84. We recalculated for this study the interrater reliability for the proposition accuracy score on the construct-a-map technique; results were the same as in the previous study.

Magnitude of the mean scores decreased on the second study. However, a statistically significant difference between occasions was observed only for the fill-in-the-lines technique ($t = 2.71$; $p = .03$).[7] As expected, teachers' mean scores were the highest across the three assessments.

Consistent with the results of the previous study (Ruiz-Primo et al., 1998), the three techniques did not provide the same picture of the level of students' knowledge. Whereas fill-in-the-map techniques indicated that students' performance was close to the maximum possible, the proposition accuracy scores indicated that students' knowledge was rather partial compared to a criterion map.

Correspondence Between Directedness, Inferred Cognitive Activities, and Assessment Score

In this section we provide evidence about the correspondence between the directedness of the mapping technique with the inferred cognitive activities and the students' scores. We first describe the procedure followed to transform, organize and synthesize the verbal codes; then we discuss the comparison between techniques using the three dimensions; finally, we compare students within a technique according to their level of competence.

Verbal analysis. Two coders independently coded subjects' protocols. Agreement between coders was based on the frequency observed on each subcategory by assessment on each subject. Agreement was defined as when the frequency in the subcategories did not differ by more than one unit (±1) between coders. For simplicity, agreement was averaged across students for each type of assessment (Table 3). Results indicated that coders agreed on the frequency observed for each subcategory across assessments and students at least 85% of the time. Percent of agreement varied according to assessment technique; the lowest percent was observed for the construct-a-map technique. A possible explanation for this difference is that the number of verbal units coded in this type of assessment was much greater than the number coded for the other two techniques, which increases the possibility of disagreement.

Since each student had a different number of verbal units, a proportion score was created for each subcategory on each assessment for each coder. Proportions were used to calculate intercoder reliability for each subcategory of each assessment.

[7] We acknowledge that the number of students limits the power of the statistical test.

Table 3

Agreement,[a] Intercoder Reliability,[a] and Proportion Scores by Category and Assessment

				Proportion scores			
Mapping technique	*n*	Percent of agreement	Intercoder reliability	Explanation	Monitoring	Conceptual errors	No code applicable
Construct-a-map	11	85	.91	.35	.28	.09	.27
Fill-in-nodes	10	88	.78	.04	.48	.007	.47
Fill-in-lines	9	89	.78	.02	.39	.001	.58

[a] Averaged across subcategories.

Intercoder reliabilities were averaged across subcategories for each assessment (Table 3). Magnitude of the reliability coefficients varied across assessments, but in general, coefficients indicate that coders similarly ranked subjects based on their proportion scores.

We averaged coders' proportion scores for each subcategory. Table 3 provides the average of proportions across subjects for the same assessment. For simplicity, subcategory proportions were added within each category (e.g., E.1 + E.2 + E.3); proportions by category are provided in the Appendix. Weighted means were calculated for each category since the number of subjects per group varied.

Directedness of the assessment tasks and correspondence with inferred cognitive activities. If different mapping techniques did impose different demands on students (e.g., content knowledge required to approach the task), we then expected the characteristics of students' cognitive activities to vary across techniques. That is, we expected the patterns of verbalizations to vary from one technique to the next. Overall, results indicated that this was the case. When students constructed a map, the verbalizations were mainly distributed across three types—Explanation, Monitoring, and No Code Applicable; for students using the fill-in-the-map techniques verbalizations were mainly distributed across only two types—Monitoring and No Code Applicable. For the fill-in-the-lines technique, the higher proportion of verbalizations was for No Code Applicable.

Verbal units that reflected Explanation were more frequent in the construct-a-map technique than in the fill-in-the-map techniques. We interpreted this result as reflecting that low-directed techniques demanded more content knowledge than high-directed techniques. We concluded that the low-directed technique provided subjects with more opportunities to display what they knew about the topic at hand than the high-directed techniques. It is important to note that most of the explanations verbalized across assessments were of type E.1 (defining; see Appendix). Few explanations of type E.2 (comparing/contrasting) or E.3 (justifying) were found across assessments.

We expected a greater proportion of Monitoring verbalizations in the low-directed technique than in the high-directed technique due to the “openness” of the task. We thought students would need to check themselves more in the low-directed technique since they needed to make more decisions on their own, for example, on the connections between concepts they were establishing. However, results indicated that subjects monitored themselves more on the high-directed techniques than on the low-directed technique. A possible explanation for this result is that because subjects made more decisions on their own in the construct-a-map technique, those decisions were considered as “final” since not much was left to question students’ own knowledge. In the fill-in-the-map techniques students were more aware of the accuracy of their responses (assessment format focuses more clearly on correct/incorrect choices), leading them to monitor themselves more frequently. Proportions by subcategory (see Appendix) indicated that a higher number of M.2 monitoring verbalizations (checking/reviewing content accuracy) were provided in the fill-in-the-map techniques than in the construct-a-map technique, whereas a higher number of M.1 monitoring verbalizations (defining/applying a strategy to approach the task) were found in the construct-a-map technique.

Another difference in the pattern of proportions observed across assessments was in the “conceptual error” category. Results indicated that more conceptual errors arose in the construct-a-map technique than in the fill-in-the-map technique. We interpreted this result as indicating that the low-directed technique may allow students to more accurately show their actual conceptual understanding. Students revealed their misconceptions more frequently while developing connections between concepts than when they were only selecting a response.

Based on these results, we concluded that the inferred cognitive activities across assessment tasks were different and corresponded with the directedness of the assessment task. The low-directed technique seemed to provide students with more opportunities to reflect their actual conceptual understanding than the high-directed techniques. In contrast, the high-directed techniques encouraged students to more closely monitor the accuracy of their responses in the map.

Directedness of the assessment tasks and correspondence with scores. The third dimension proposed to compare the mapping techniques is the subjects' obtained scores across the techniques. In this section, we first present the comparison of scores across the assessment techniques, and then we compare the scores with the inferred cognitive activities.

In our previous study (Ruiz-Primo et al., 1998) we compared scores across mapping techniques using a correlational approach since score scales were different across mapping techniques. The magnitude of the correlations between construct-a-map and fill-in-the-map scores indicated that the techniques—low- and high-directed—ranked students somewhat differently (averaged *r* = .51).[8] The magnitude of the correlations was higher between construct-a-map and fill-in-the-lines scores than between construct-a-map and fill-in-the-nodes scores. The correlations in the subsample of the students that participated in this study (Table 4, Occasion 1) showed a higher averaged magnitude (averaged *r* = .60), although the pattern was the same.

Correlations between studies (scores obtained across assessments in the first study and this study) indicated that construct-a-map and fill-in-the-lines techniques ranked students more similarly than the fill-in-the-nodes technique. The low

Table 4

Correlations Between Scores Obtained Within and Between Studies 1 and 2 and Assessments

	Study 1			Study 2			Studies 1 and 2		
Mapping technique	CM	NOD	LIN	CM	NOD	LIN	CM	NOD	LIN
Construct-a-map-CM	—			—			.82**		
Fill-in-the-nodes-NOD	.53	—		.65	—		.51	.63	
Fill-in-the-lines-LIN	.66	.41	—	.94**	.69	—	.83*	.46	.91**

Note. For the fill-in-the-map techniques only 7 students for Study 2.

* Correlation is significant at the .05 level. ** Correlation is significant at the .01 level.

[8] 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.

correlation between occasions for the fill-in-the-nodes technique may be due to restriction of range.

Notice, however, that the magnitude of the correlations changed substantially between construct-a-map and fill-in-the-lines scores. Whereas the magnitude of these correlations in Study 1 for the complete sample was .51 and for this sub-sample was .66, the correlation jumped to .83 across occasions and to .94 for Occasion 2. A possible explanation of these results may be that for the fill-in-the-lines technique, only one medium-score student was included in the correlation, leaving the group with basically high- and low-score students, which made the ranking more straightforward across techniques. For the fill-in-the-nodes technique, however, high scores were observed across students independently of the groups they belonged to (high, medium, or low). Remember also that the mean score for the fill-in-the-lines technique was significantly lower on Occasion 2 (Table 2). A closer look into the scores of the nine students participating in this study indicated that scores of the low-score students were lower in Study 2 than in Study 1.

To conclude that construct-a-map and fill-in-the-lines can be considered similar techniques, a different conclusion from the one we arrived at in our previous study, will require further investigation since the sample used in this study is small. Possibly the only clear conclusion so far is that fill-in-the-nodes, a high-directed technique, is tapping a different aspect of students' connected understanding than the construct-a-map or fill-in-the-lines techniques.

A second piece of evidence related to students' scores obtained across mapping techniques varying in directedness is the inferred cognitive activity displayed during concurrent verbalization. Below we present correlations between the proportion scores by type of verbalization and students' scores according to the mapping technique. Then we provide a description of the characteristics of the students' cognitive activities by their level of competence.

We correlated the students' Explanation, Monitoring, and Conceptual Errors proportion scores by techniques with scores obtained with the correspondent map technique. Our expectations for pattern of these correlations were based on the characteristics of competent science students (Baxter et al., 1996). Competent students tend to give content-based explanations as they perform. Therefore, we expected that competent students in our study would provide more explanations and obtain higher scores than less competent students. In sum, we expected a correspondence between the proportion of explanations provided and the scores obtained (Table 5).

All correlations with Explanation proportions were positive, indicating that the higher the proportion of verbalizations on Explanation, the higher was the score obtained. The low correlations for the fill-in-the-map techniques were due to the low Explanation proportions observed across those techniques. Remember that, in general, few explanations were provided by students while performing those assessment tasks (Table 3). We concluded that construct-a-map, a low-directed technique, better reflected the correspondence between students' content-based explanations and the scores obtained.

Correlations between students' Monitoring verbalizations and map scores were all negative and nonsignificant. This result was expected based on the obtained proportions by level and type of monitoring (see Appendix). The pattern of proportions for Monitoring by level and type showed that low-score students had proportions of Monitoring verbalizations similar to or higher than those of both high-score students and teachers. This pattern was unexpected based on theory and research that says that self-monitoring is a characteristic of competent students (e.g., Baxter et al., 1996; Baxter & Glaser, 1998).

Why, then, did competent students and teachers not provide higher levels of Monitoring verbalizations than medium- and low-score students? We concluded that possibly the interaction between the characteristics of the assessment tasks and the student's level of competence determined the amount and type of monitoring required by each subject. For example, high-score students in the construct-a-map task checked off the concepts that they already used in the map (a M.1 form of monitoring), and looked at the list of concepts trying to make more connections (a M.2 form of monitoring). Low-score students also checked off the concepts on the

Table 5

Correlations Between Scores Obtained Within Assessments and Type of Verbalization

Mapping technique	Explanation	Monitoring	Conceptual errors
Construct-a-map	.86**	-.32	-.55
Fill-in-nodes	.38	-.30	-.85
Fill-in-lines	.07	-.08	-.13

** Correlation is significant at the .01 level.

list and asked themselves many questions, but the questions were about the terms' meanings and their connections (a M.2 form of monitoring) or about the terms they did not remember (a M.3 form of monitoring). In summary, competent subjects monitored themselves on key steps in the tasks but did not frequently ask themselves about their progress, since they probably knew they were doing "fine." In contrast, low-competent students frequently asked themselves about what and how they were doing since they probably did not know the content (see Salthouse, 1991). To draw a general conclusion about the correspondence between students' scores, monitoring, and directedness of assessment tasks will require a more detailed analysis of the monitoring category.

We found negative correlations between students' scores and conceptual errors. Although correlations were nonsignificant, they were in the expected direction: As the proportion of errors decreased, students' map scores increased. The magnitude of the correlations varied by mapping technique. The highest correlation was found for the fill-in-the-nodes technique and the lowest for the fill-in-the-lines. The latter correlation was due to restriction of range since conceptual errors were identified only in the low-score students. As mentioned before, the highest number of conceptual errors across subjects was observed in the construct-a-map technique; however, the correlation was not as high as we expected.

Final conclusions about the correspondence of scores with the directedness of the assessment tasks will require more analysis. It is possible that combining the monitoring proportions was not the most appropriate way to deal with the subcategories for this form of verbalization. Also, we plan to review the conceptual errors and probably define subcategories, since we found that the conceptual errors provided by the high-score students and teachers differed from the conceptual errors of medium- and low-score students. For example, low-score students' conceptual errors were very basic (e.g., they could not recognize the difference between cations and anions) compared to the errors made by the high-score students (e.g., "…electrons electrons, cations lose electrons, anions gain electrons…").

Comparing students' cognitive activities by performance level and within assessment techniques. To compare patterns of verbalizations across groups of subjects we averaged proportion scores within the same group (i.e., teacher, high-, medium-, and low-score students) by assessment (Table 6). As expected, patterns in each category within the same assessment differed according to students' level of

Table 6

Percentage of Verbal Units by Category, Assessment, and Level of Subjects

Mapping technique	*n*	Explanation	Monitoring	Conceptual errors	No code applicable
Construct-a-map					
Teachers	2	.44	.21	.06	.27
High	3	.42	.28	.07	.22
Medium	3	.32	.23	.13	.31
Low	3	.23	.40	.10	
Weighted mean		.35	.28	.09	.27
Fill-in-nodes					
Teachers	2	0	.42	0	.59
High	3	.09	.37	0	.54
Medium	2	.04	.75	.03	.19
Low	3	.03	.45	.008	.52
Weighted Mean		.04	.48	.007	.47
Fill-in-lines					
Teachers	2	.01	.27	0	.72
High	3	.03	.38	0	.58
Medium	1	0	.57	0	.42
Low	3	.02	.42	.003	.55
Weighted mean		.02	.39	.001	.58

competency. In general, high-score students provided more explanations than low-score students. This shows that effective learning of content knowledge enables students to explain their reasoning underlying their responses or choices (e.g., high-score student: “binary ionic compounds are formed by two ions, or cation and anion, binary because they have two ions”). Only in the construct-a-map technique did teachers provide more explanations than high-score students; on the other techniques subjects provided few explanations. As mentioned previously, subjects had to construct and elaborate each of the relationships between concepts in the construct-a-map technique, whereas they only needed to recognize the correct node or description for the fill-in-the-map techniques.

Differences in the patterns of the proportion of conceptual errors were also expected: High-score students and teachers provided fewer conceptual errors than medium- and low-score students. Notice that the proportion of conceptual errors was higher in the construct-a-map technique than in the fill-in-the-map technique.

Although for the Monitoring category we expected a pattern similar to the one observed for Explanation, this was not the case. The pattern between groups varied according to the assessment. As noted previously, frequency and type of monitoring varied according to the characteristics of the task and the competency of the subjects.

Based on the pattern differences between level of competency groups for Explanation and Conceptual Errors, we concluded that the construct-a-map technique better reflected differences in subjects' cognitive activities according to their level of competence. However, more analysis will be done to define which characteristics of this mapping technique are shared with the fill-in-the-lines technique since both ranked students similarly (see the previous section).

Finally, we provide information about the higher order categories, Planning and Strategy, by group and assessment. Subjects' verbalizations at the beginning of each protocol were the only ones considered as Planning; the rest of the verbalization were analyzed as characteristics of the strategy(ies) used by the subjects while performing the mapping assessment tasks. To avoid lengthy quotes comparing students, we first provide a summary of the characteristics observed in Planning and Strategy across groups, and then two prototypical subjects' statements.

To characterize planning and strategy we used the attributes provided by Baxter et al. (1996) and Baxter and Glaser (1998). Planning was characterized according to Presence and Type of Plan. Presence was defined as when subjects' verbalizations within each assessment showed a plan. If some sign of planning was observed, it was categorized either as a plan providing *Procedures and Outcomes* (the type of planning provided by competent students) or as providing only a *Single Decision* with no justification. Strategy was characterized as *Efficient*—strategy used by the student reflected relevant goals and subgoals, or *Trial-and-Error*—strategy did not reflect any systematic way for performing the assessment task. Presence of Strategy was not considered because subjects needed to use a strategy, either an efficient one or one based on trial and error. Table 7 presents the findings.

Table 7

Summary of Planning and Strategy by Assessment and Level of Subjects

		Planning			Strategy	
Mapping technique	*n*	Presence	Presence & outcomes	Single decision	Efficient	Trial-and-error
Construct-a-map						
Teachers	2	1	1	—	2	—
High	3	2	1	1	3	—
Medium	3	2	2	—	3	—
Low	3	3	—	3	3	—
Fill-in-nodes						
Teachers	2	0	—	—	2	—
High	3	1	1	—	3	—
Medium	2	2	1	1	2	—
Low	3	0	—	—	—	3
Fill-in-lines						
Teachers	2	0	—	—	1	1
High	3	1	—	1	1	2
Medium	1	1	—	1	—	1
Low	3	0	—	—	—	3

Although we expected teachers and high-score students to have “a plan” that provided procedures and outcomes for performing each assessment task, this was not the case. Providing a plan was not a generalized characteristic of competent subjects, in contrast to the literature (e.g., Baxter et al., 1996). A possible explanation may be, again, the interaction between the characteristics of the tasks and level of competence. Notice that in construct-a-map, a low-directed technique, more subjects stated a plan than in the high-directed techniques (Table 7). It is possible that competent subjects do not formulate a plan when the characteristics of the task do not lead to many different options for performing it. Nevertheless, low-score students in the construct-a-map technique provided plans with only single statements; their plans did not provide justification or possible procedures or outcomes.

A similar situation was found for Strategy. We found teachers and high-score students using a trial-and-error strategy for one of the high-directed techniques, fill-in-the-lines. Once again, it seems that the interaction hypothesis can help explain these results.

As an example of the verbalization used to analyze Planning and Strategy, we present here the analysis of two students’ protocols, one high- and one low-score, for the construct-a-map technique. Although both students, high-score and low-score, provided a plan, the planning of the high-score student showed a clear procedure: Before starting the concept map, the student provided a definition of almost all the concepts in the list and grouped the concepts:

> The charge has negative, electrons are negative charges, anions have negative charge, so cations are positive charge, so they are related [student goes on and defines and groups the concepts] . . . so periodic table is the most general one, so the periodic table is in the middle.

The low-score student read all the concepts, selected one, with no justification for doing so, and started drawing the map:

> I am just going to go over the list of concepts . . . [reads each one of the concepts aloud] . . . I am going to use the periodic table as my starter.

The high-score student's plan was composed of actions (defined the concepts and grouped them) that helped him anticipate the map—a sort of trial run through the solution strategy. The low-score student lacked a plan. Anticipating a solution strategy that would help to generate and anticipate steps has been defined as a characteristic of competent students (e.g., Baxter et al., 1996; Gentner & Stevens, 1983).

It seems that for the high-directed technique, subjects used a trial-and-error strategy, which led to correct responses. High-directed techniques allowed low-score students to guess and obtain correct responses, despite their partial knowledge. Further refinement to characterize planning and strategy are needed since what we used did not provide a complete picture of the students' inferred cognitive activities. It is possible that the quality of cognitive activities by level of knowledge described by Baxter et al. (1996) better fits performance assessments than concept maps.

Conclusions

This paper evaluated the validity of connected understanding interpretation of three mapping techniques. The study focused on the correspondence between mapping-intended task demands, inferred cognitive activities, and scores obtained. We analyzed subjects' concurrent and retrospective verbalizations at different levels of competency performing the mapping tasks and compared the directedness of the mapping tasks, the characteristics of verbalization and the scores obtained across techniques.

Our results led to the following general conclusions. (a) Consistent with a previous study, we found that the three mapping techniques provided different pictures of students' knowledge. High-directed techniques, such as fill-in-the-map, indicated that students' performance was close to the maximum criterion, whereas construct-a-map, a low-directed technique, indicated that students' knowledge was rather partial compared to a criterion map. (b) Inferred cognitive activities across assessment tasks were different and corresponded to the directedness of the assessment task. The low-directed technique seemed to provide students with more opportunities to reflect their actual conceptual understanding than the high-directed techniques. (c) Results on the convergence of mapping scores were not consistent with the results of our previous study. The magnitude of the correlation between construct-a-map and fill-in-the-lines scores was much higher than the one observed before. We suggest a more detailed investigation of the characteristics of these two mapping techniques before considering them as equivalent.

Evidence about the correspondence between the three dimensions used to compare the techniques indicated that the construct-a-map technique seems to be the most congruent in terms of the directedness of the task, the cognitive activities inferred, and the scores obtained.

References

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

Baxter, G. P., & Glaser, R., (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practices, 17*(3), 37-45.

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal reports as data*. Cambridge, MA: The MIT Press.

Ericsson K. A., & Smith, J. (Eds.). (1991). *Toward a general theory of expertise. Prospects and limits*. New York: Cambridge University Press.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice Hall.

Glaser, R., & Baxter, G. P. (1997, February). *Improving the theory and practice of achievement testing*. Paper presented at the BOTA meeting, National Academy of Science / National Research Council, Washington, DC.

Levine, R. (1998). *Cognitive lab report* (Report prepared for the National Assessment Governing Board). Palo Alto, CA: American Institutes for Research.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 5-21.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

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

Rice, D. C., Ryan, J. M., & Samson, S. M. (1998). Using concept maps to assess student learning in the science classroom: Must different methods compete? *Journal of Research in Science Teaching, 35,* 1103-1127.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (1998, April). *Comparison of the reliability and validity of scores from two concept-mapping techniques*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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 achievement. *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, IL.

Salthouse, T. A. (1991). Expertise as the circumvention of human processing limitations. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise. Prospects and limits* (pp. 286-300). New York: Cambridge University Press.

White, R., & Gunstone, R. (1992). *Probing understanding*. London: The Falmer Press.

Appendix

Proportion of Verbal Units by Subcategory, Group, and Assessment

Assessment and group	E.1	E.2	E.3	M.1	M.2	M.3	M.4	CE	NCA
Construct-a-map									
Teachers	.44	.004	0	.12	.04	.05	.002	.07	.27
High	.39	.02	.01	.11	.12	.04	.008	.07	.21
Medium	.30	.006	.01	.17	.05	.01	.003	.013	.31
Low	.22	.003	.003	.21	.12	.06	.01	.10	.27
Weighted mean	.33	.009	.007	.15	.08	.04	.006	.09	.27
Fill-in-the-nodes									
Teachers	0	0	0	.17	.21	.04	0	0	.59
High	.04	.01	.05	.19	.18	0	0	0	.54
Medium	.04	0	0	.30	.31	.13	.02	.025	.19
Low	.01	0	.02	.18	.22	.05	.007	.008	.52
Weighted mean	.02	.003	.02	.20	.22	.05	.006	.007	.47
Fill-in-the-lines									
Teachers	0	0	.01	.08	.16	.03	0	0	.72
High	.02	.003	.01	.18	.19	.01	0	0	.58
Medium	0	0	0	.32	.21	.02	.02	0	.41
Low	.01	0	.01	.17	.18	.07	0	.003	.55
Weighted mean	.01	.001	.008	.17	.18	.04	.002	.001	.58

Note. E.1 = Defining. E.2 = Comparing / Contrasting. E.3 = Justifying. M.1 = Defining / Applying a Strategy. M.2 = Effective Reflecting. M.3 = Ineffective Reflecting. M.4 = Reviewing Quality of Product. CE = Conceptual Error. NCA = No Code Applicable.