A Comparison of Two Construct-a-Concept-Map Science Assessments: Created Linking Phrases and Selected Linking Phrases

CSE Report 624

Yue Yin, Jim Vanides, Maria Araceli Ruiz-Primo,
Carlos C. Ayala, and Richard Shavelson

National Center for Research on Evaluation, Standards,
and Student Testing (CRESST) / Stanford University

March 2004

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 2.2: Classroom and Teachers' Assessment. Strand 2: Curriculum Embedded Assessments—Studies 1 and 2
Project Director: Richard J. Shavelson, Stanford University / CRESST

The work reported herein was supported under the Educational Research and Development Centers Program, PR / Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

A COMPARISON OF TWO CONSTRUCT-A-CONCEPT-MAP SCIENCE ASSESSMENTS: CREATED LINKING PHRASES AND SELECTED LINKING PHRASES*

Yue Yin, Jim Vanides, Maria Araceli Ruiz-Primo, Carlos C. Ayala, and Richard Shavelson

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/Stanford University

Abstract

In this paper we examine the equivalence of two construct-a-concept-map techniques: construct-a-map with created linking phrases (C) and construct-a-map with selected linking phrases (S). The former places few constraints on the respondent and has been considered the gold standard; the latter is cost- and time-efficient. They are compared in terms of both concept map products and processes. Both quantitative and qualitative variables are used for comparison: total accuracy score, individual proposition scores, proposition choice, map structure complexity, proposition generation rate, and proposition generation procedures. We conclude that the two mapping techniques are not equivalent: The C mapping technique is better than S in capturing students' partial knowledge, even though the S mapping technique could be scored more efficiently than C. Based on the characteristics of the two techniques, if used as an assessment tool, the C mapping technique is suitable for formative assessment, and the S mapping technique is a better fit for large-scale assessments.

Knowledge structure is regarded as an important component of understanding in a subject domain, especially in science (e.g., Novak, 1990; Novak & Gowin, 1984). The knowledge structure of experts and successful learners is characterized by elaborate, highly integrated frameworks of related concepts (e.g., Chi, Glaser, & Farr, 1988; Mintzes, Wandersee, & Novak, 1997), which facilitate problem solving and other cognitive activities (e.g., Baxter, Elder, & Glaser, 1996). A knowledge structure, then, might well be considered an important but generally unmeasured aspect of science achievement. Concept mapping techniques are interpreted as representative of students' knowledge structures and so might provide one possible means of tapping into a student's conceptual knowledge structure (e.g., Mintzes et al., 1997; Novak & Gowin, 1984).

*A revised version of this report will appear in the *Journal of Research in Science Teaching*.

A concept map includes *nodes* (terms or concepts), *linking lines* (usually with a uni-directional arrow from one concept to another), and *linking phrases* which describe the relationship between nodes. Linking lines with linking phrases are called *labeled lines*. Two nodes connected with a labeled line are called a *proposition*. Moreover, concept arrangement and linking line orientation determine the *structure* of the map (e.g., hierarchical or nonhierarchical).

A concept map assessment is composed of a task, a response format, and a scoring system, and hundreds of concept map assessment permutations are possible (Ruiz-Primo & Shavelson, 1996b). Though the variation among maps provides practitioners with numerous options for use and interpretation, the diversity poses challenges and opportunities for the measurement of achievement: Do different concept mapping techniques measure the same or different constructs—for example, knowledge structures? Are the cognitive processes evoked when students construct different kinds of maps the same or different? Do different mapping techniques lead to different levels of performance? Which concept mapping techniques are best suited for which assessment purposes? Finally, how can different concept mapping techniques be scored effectively and efficiently?

Ruiz-Primo and Shavelson (1996b) raised questions about the reliability and validity of concept maps as assessment tools that have been followed by recent research (e.g., Kinchin, 2000; Klein, Chung, Osmundson, Herl, & O'Neil, 2002; McClure, Sonak, & Suen, 1999; Nicoll, Francisco, & Nakhleh, 2001; Ruiz-Primo, Schultz, Li, & Shavelson, 2001; Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Rye & Rubba, 2002). Among those studies, some compared different concept mapping tasks (e.g., Ruiz-Primo, Shavelson, et al., 2001), some explored different scoring systems (e.g., Kinchin, 2000; Klein et al., 2002; McClure et al., 1999), and others have examined the validity of concept map assessments by using think-aloud protocols and other measures (e.g., Herl, O'Neil, Chung, & Schacter, 1999; Ruiz-Primo, Schultz, et al., 2001; Ruiz-Primo, Shavelson, et al., 2001). Our study is an extension of this line of research. More specifically, we examined the equivalence of two concept map assessment tasks: (a) construct-a-map with created linking phrases (C) and (b) construct-a-map with selected linking phrases (S). In the C condition students are provided concepts and asked to construct a map using self-created linking phrases. In contrast, the S mapping technique supplies students with both linking phrases and the concept terms; students need to select and assemble the concepts and linking phrases.

The C mapping technique is characterized as the gold standard of concept maps (Ruiz-Primo, Schultz, et al., 2001; Ruiz-Primo, Shavelson, et al., 2001). Compared with the fill-in-a-map technique (where students fill in a predrawn map), the C mapping technique (a) more accurately reflects differences of students' knowledge structures; (b) provides greater latitude for demonstrating students' partial understanding and misconceptions; (c) supplies students with more opportunities to determine their conceptual understanding; and (d) elicits more higher order cognitive processes, such as explaining and planning. However, due to the range and diversity of students' self-created linking phrases, the C mapping technique is burdened with scoring difficulties.

A possible solution to these scoring difficulties is to ask students to construct a map selecting from predetermined linking phrases (i.e., the "S" condition). Researchers found that the advantage of this technique was that the scoring of these maps could be automated with computers (Klein et al., 2002; O'Neil, Chung, & Herl, 1999). Because the number of propositions was bounded, computers could easily compare students' maps with a criterion or expert map(s), typically created by science educators, teachers, and/or scientists. Klein et al. (2002) suggested that the computer made scoring straightforward and effective. This advantage is particularly appealing when considering the use of concept maps as a potential large-scale assessment tool.

Given the openness of the C mapping technique and the constraints of the S mapping technique the question remains: Are the two techniques equivalent? Our study aimed to supply conceptual and empirical evidence of their characteristics in order to make this comparison.

Framework for Comparing Two Construct-a-Map Techniques

We compare the two mapping techniques—C and S—based on six variables. Figure 1 summarizes these variables on two dimensions: (a) whether the variable is the product or the process of map construction, and (b) whether the variable is quantitative or qualitative. Later, we illustrate these variables in detail.

Concept map products are the result of the task demands imposed on the student, that is, the concept maps drawn on a piece of paper. In our study, the concept map product variables were derived from students' drawn concept maps and are both quantitative (*total proportion accuracy score* and *individual proportion accuracy score*) and qualitative (*proposition choice* and *structure complexity*).

		Comparison Targets	
		Concept Map Products	Concept Map Processes
Characteristics	Quantitative	Total Accuracy Score Individual Proportion Score	Rate of Proposition Generation
	Qualitative	Proposition Choice Structure Complexity	Procedure for Generating Propositions

Figure 1. Variables for comparing concept maps.

Concept map processes refer to a student's inferred cognitive activities elicited during the execution of the concept map task (Ruiz-Primo, Shavelson, et al., 2001). Concept map process variables were created from students' *think aloud* protocols while constructing their maps. Concept map processes include a quantitative variable (*proposition generation rate*) and a qualitative variable (*proposition generation procedure*). If two concept map assessments are equivalent, they should elicit the same cognitive activities as well as the same, or very similar, final products, leading to similar inferences about a student's knowledge structure (Ruiz-Primo, Shavelson, et al., 2001).

Six Comparison Variables

We used the six variables in the framework to compare the two concept mapping techniques. The variables will be elaborated as concept map products and concept map processes.

Concept map products. Construct-a-map assessments are challenging to score because students' products vary greatly. The total number of propositions is indeterminate and the structure of the map is unfixed. Therefore, to adequately characterize a student's concept map, we set up a multidimensional scoring system to evaluate the map from different perspectives.

Other researchers have proposed some of these dimensions. For example, Herl et al. (1999) used four types of scores: (a) semantic content score, (b) organizational structure score, (c) number of terms used, and (d) number of links. McClure et al. (1999) suggested scoring concept maps using holistic scoring, relational scoring, and

structural scoring. Kinchin (2000) suggested a two-tier analysis, scoring concept map links quantitatively and their structure qualitatively. Nicoll et al. (2001) suggested scoring concept map links with a three-tier system, coding links for their utility, stability, and complexity.

When scoring concept map products, we focused on the propositions and the structure of the concept map, the two features that have been of primary interest in most of these research studies. We applied three variables to describe propositions (total accuracy score, individual proposition score, and proposition choice) and one variable to describe structure (structure complexity).

The proposition, composed of two concepts and a linking phrase, acts as the building block of a concept map, supplying information about students' declarative knowledge on concept pairs. A proposition is relatively easy to score and is interpreted as revealing depth of understanding (McClure et al., 1999). Various methods have been used in previous research to score propositions. For example, criterion maps can either be applied or not when assessors score students' maps, and scores can focus on either quantity (number score), or quality (accuracy score), or both (proportion score). Table 1 presents variations in proposition scoring approaches: with a criterion map or without a criterion map.

Table 1

Summary of Scores Applied to Propositions

Type of score[a]	With a criterion map	Without a criterion map
Count score	Stringent semantic content score: based on exact link matches between student links and expert links (Herl et al., 1999)	Linkage: the total number of links (Astin & Shore, 1995; Herl et al., 1999; Lomask, Baron, Greig, & Harrison, 1992)
	Categorized semantic content score: based on students matching some set of possible links (Herl et al., 1999)	Good links: the total number of links showing good understanding (Astin & Shore, 1995; Herl et al., 1999; Lomask et al., 1992)
Accuracy score	Weighted relationship score: score is given to the individual relationship (proposition) based on its similarity to criterion one (Rye & Rubba, 2002)	Total proposition accuracy score: total sum of the quality scores obtained on all propositions (Ruiz-Primo & Shavelson, 1996a)
Proportion score	Congruence score: proportion of valid student links over all criterion links (Ruiz-Primo & Shavelson, 1996a)	Salience score: proportion of valid student links over all student links (Ruiz-Primo & Shavelson, 1996a)

[a]The authors of this paper named the three score types according to different scores' features.

We designed our assessment based on a criterion map. However, we scored the maps without using the criterion map in order to fully capture students' knowledge structures beyond what might have been in the criterion map. Moreover, research has shown that the total accuracy score is reliable, and also that it effectively shows differences in students' knowledge structures (Ruiz-Primo & Shavelson, 1996a). We chose to apply the total accuracy score from among the three score types without using a criterion map. Our assumption was that if the C and S techniques are equivalent, they should produce the same mean score, variance among scores, and high test-retest reliability.

In addition to the total accuracy score, we examined individual proposition scores for each student. For example, if a student constructed 10 propositions, we gathered the accuracy score for every proposition he or she constructed. In this way, we could compare the two mapping techniques' score distributions by proposition. Our expectation was that if the two mapping techniques were equivalent, C and S students' individual proposition score distributions should follow a similar pattern.

Both the total accuracy score and individual proposition scores are quantitative product measures. We also tracked qualitative characteristics using proposition choices. Even though the concepts are supplied in both map conditions, propositions are not fixed by the mapping task. That is, students decide which pairs of concepts have meaningful relationships. Assuming that students choose to establish relationships that they think are important or interesting, we examined the structures of students' declarative knowledge by analyzing the set of propositions they constructed. If two assessment techniques are equivalent, a given student should choose to build similar propositions in both assessments.

Besides scoring concept map propositions, we also examined concept map structure complexity. Novak and Gowin (1984) argued that concept maps should be hierarchically structured. However, other research has shown that hierarchical structures are not always necessary (e.g., Dansereau & Holley, 1982). In this study, we focused on the maps' graphic feature, rather than Novak's hierarchical structure one. Kinchin (2000) proposed three concept map structure types: spoke, chain, and net. He pointed out that a net structure is indicative of meaningful learning, and he suggested this qualitative scheme could be quickly and easily used, providing teachers with a simple starting point for concept map analysis.

In a preliminary review of our students' responses, we realized that the spoke, chain, and net structures proposed by Kinchin (2000) did not fully characterize the different structures students created in our study. Therefore, we added two new structure types—circle and line—to capture the different structures present in our students' maps. We defined the five structure types as follows (see Figure 2): (a) Linear—Propositions that are daisy-chained together; (b) Circular—Propositions that are daisy-chained with the ends joined; (c) Hub or Spokes—Propositions that emanate from a center concept; (d) Tree—A linear chain of propositions that has branches attached; and (e) Network or Net—A complex set of interconnected propositions. Among them, the network or net structure is considered the most complex, and the linear structure is the simplest. All the others fall in between. In our study, we refer to structure type as "structure complexity."

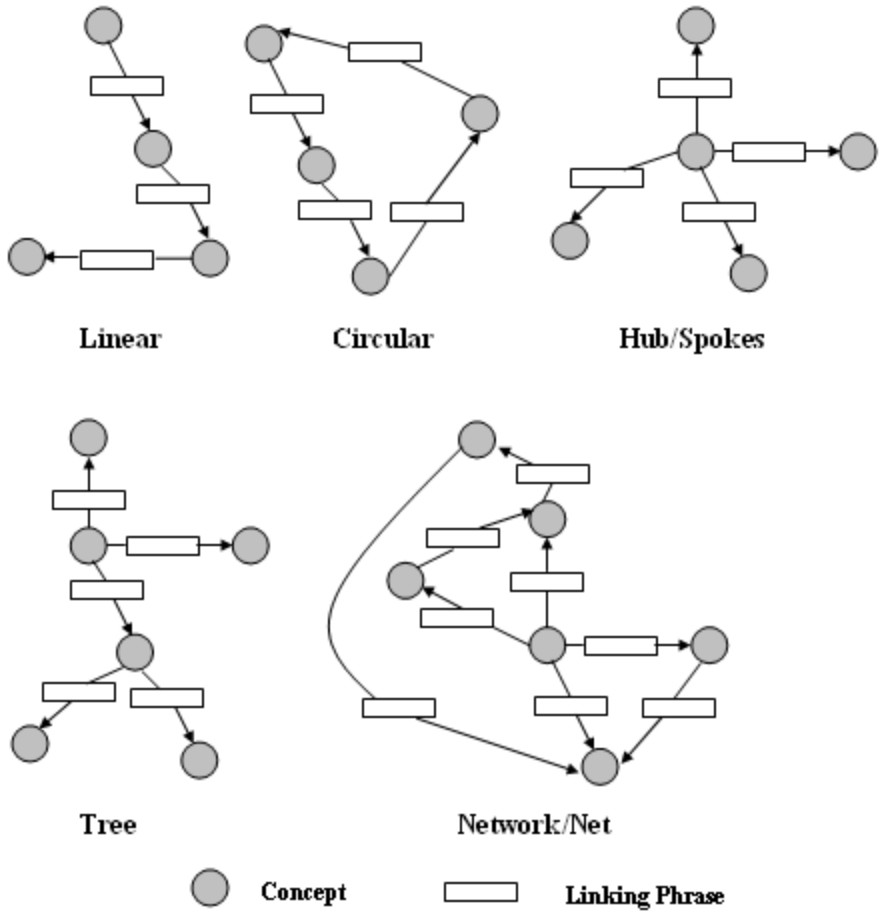

Figure 2. Structure complexity: Five key concept map structures.

Research has shown that experts, compared with novices, exhibit deeper and more connected and interrelated map structures (Mintzes et al., 1997); therefore, we expected that maps created by high performers would be more complex than those created by low performers. If the two concept mapping techniques, C and S, are equivalent, then a student responding to both techniques should construct concept maps with the same structure complexity.

Concept map processes. Previous studies have shown that students apply different strategies when completing fill-in-a-map and construct-a-map tasks (Ruiz-Primo, Shavelson, et al., 2001). Correspondingly, if the two construct-a-map techniques, C and S, are equivalent, we expected to find that students engaged in similar cognitive activities when completing the C and S maps.

To infer cognitive activities, we asked students to think aloud while mapping. The think-aloud technique has been used to reveal cognitive activities in performing a variety of tasks (Ericsson & Simon, 1993); for example, problem solving (Baxter & Glaser, 1998), multiple-choice test taking (Levine, 1998), concept map construction (Ruiz-Primo, Shavelson, et al., 2001), and performance assessment (Ayala, Yin, Shavelson, & Vanides, 2002; Yin, Ayala, & Shavelson, 2002). To collect think-aloud data, researchers ask participants to verbalize their thinking while they are performing the assigned activities. This verbal evidence is recorded and transcribed for analysis. However, previous studies suggested that some information might be lost when only talking was recorded because verbal evidence is sometimes incomplete or ambiguous (Yin et al., 2002). Therefore, in this study we also video-taped participants' actions. Our think-aloud approach focused on the overall pattern of proposition generation. In particular, we examined the proposition generation rate and proposition generation procedures. Proposition generation rate refers to the speed with which students constructed propositions; proposition generation procedures refer to what steps students take in proposition construction. We inferred students' cognitive activity from proposition generation rates and proposition generation procedures.

We suspected that the linking phrases supplied in the S technique might have two functions: On the one hand, a list of linking phrases could provide students with hints, reminding them of the scientific relationships between concepts. On the other hand, a linking phrase list might constrain the students' choices and prevent students from setting up relationships available and interesting to them, and finally slow down their map construction. Therefore, we attempted to infer the

assessments' influence on students' cognitive process by comparing their proposition generation rates in the two assessments. Moreover, we examined students' proposition generation procedures by analyzing students' verbalizations and actions during map construction, with the hope of identifying specific activities leading to the proposition generation rate differences, if there were any.

To summarize and clarify the goals and specific means used in our study, Table 2 presents the questions we asked in order to determine whether the two mapping techniques were equivalent. The questions align with two comparison targets of construct-a-map: concept map products and concept map processes. To answer those questions, we examined our six variables across the two concept mapping techniques.

Method

Participants

Ninety-four eighth graders from the California Bay Area participated in the study; 47 were girls and 47 boys. The students, drawn largely from upper middle class homes, belonged to six middle school science classes taught by the same teacher.

Prior to this study, the students had all previously studied a unit on density, mass, and matter. The science teacher was asked to indicate on her class rosters

Table 2

Research Questions in Comparing Two Construct-a-Map Techniques

Comparison targets	Comparison questions	Comparison variable
Concept map product	Do the different technique scores have the same mean, standard deviation, and test-retest reliability?	Total accuracy score
	Do the different technique scores reveal the same knowledge level for a given student?	Individual proposition score
	Do the different techniques lead students to construct similar propositions?	Proposition choice
	Do the different techniques elicit a similar concept map structure?	Structure complexity
Concept map process	Do they elicit with a similar cognitive process?	Proposition generation rate
		Proposition generation procedure

each student's science achievement level ranging from high to medium to low based on her observations of students in class and on student work in science. Of the 94 students, 17 students were ranked in the low level, 37 were ranked in the medium level, and 40 were considered to be in the top performing level. Previous research (Shavelson, 1987) suggests that teachers can accurately rank order their students' performance. Based on this finding, we regarded the teacher's classification of students as an outside standard in our study.

Mapping Techniques

In both the C and the S conditions, we gave students nine key concepts related to buoyancy and were instructed to connect pairs of concepts with a one-way arrow to indicate a directional relationship. Students then labeled the arrows with a linking phrase that described the relationship, creating a proposition, which could be read as a sentence (e.g., WATER has a property of DENSITY).

The selection of key concepts was a cooperative effort of the assessment design team working with the curriculum designers, content experts and a master teacher. The target curriculum was a unit on buoyancy from the Foundational Approaches to Science Teaching (FAST) curriculum developed at the Curriculum Research and Development Group at the University of Hawaii (Pottenger & Young, 1992). Previous experience in concept map design suggested that a manageable concept mapping activity should use only the most important science concepts and be limited to between 8 and 12 concepts. By using an iterative selection process involving ranking and voting by the team members, an initial list of 24 possible terms was reduced to a total of 9 concept terms—WATER, VOLUME, CUBIC CENTIMETER, WOOD, DENSITY, MASS, BUOYANCY, GRAM, and MATTER.

In the C condition, students wrote linking phrases of their own choosing. In the S condition, we provided students with a list of linking phrases that they had to use (or re-use) to describe the relationships between concepts. This list was based on a criterion map created by the assessment design team. This provided a starting point for identifying potential linking phrases, some of which were later modified to be age-appropriate. We supplied the following linking phrases in the S condition: "is a measure of…", "has a property of…", "depends on…", "is a form of…", "is mass divided by…", and "divided by volume equals…"

Scoring System

Total proposition accuracy scores were based on an evaluation of the quality of propositions that students constructed. A map's total accuracy score was the sum of individual proposition scores. Individual propositions were scored using a 4-point scale—0 for wrong or scientifically irrelevant propositions, 1 for partially incorrect propositions, 2 for correct but scientifically "thin" propositions, and 3 for scientifically correct and scientifically stated propositions. For example:

0—"GRAM is a form of MASS"

1—"GRAM is a symbol of MASS "

2—"GRAM measures MASS"

3—"GRAM is a unit of MASS"

To score individual maps, we created an Excel database that contained all of the propositions submitted by each student. All the unique student-generated propositions extracted from the database comprised a master list of propositions. Using the rubric previously described, two science education graduate students independently scored the master list. The interrater reliability for the scoring of this database was initially quite low. After discussions, we created detailed rules. (a) We considered scientific intent over grammar, that is, we mainly concentrated on students' conceptual understanding instead of their wording. For example, "GRAM is a measuring unit for MASS" is scored as "3" even though it is grammatically problematic. In both S and C maps, we gave students credit if they illustrated appropriate conceptual understanding even if they did not use the exact linking words expected. (b) We gave partial credit for wrong-way arrows that connect related terms. For example, "DENSITY divided by mass equals VOLUME" was scored as "1" because even though the relationship was wrong, the student was given credit for at least pointing out the existence of the relationship between the two terms. (c) We gave partial credit for "correct but not specific enough relationship." For example, "MATTER is related to MASS" was scored as "1" because the relationship was set up but not sufficiently clarified. (d) We did not give credit for "correct but meaningless relationship." For example, "MASS is different than GRAM" was scored as "0." With those guidelines established, the two well-trained raters' interrater reliability for 50 randomly selected propositions could reach 0.92 on S map propositions and 0.81 on C map propositions due to the great diversity of

propositions created in C map. After discussing all the disagreements, two raters finally agreed on the scoring of 95% propositions in the master list. A third expert, a science education professor, resolved the scores for the remaining propositions.

Having transferred each student's concept map propositions into the Excel database, the master scoring list was used to score each proposition. The two graduate students characterized each map according to its dominant structure: Linear, Circular, Hub, Tree, and Network (see Figure 2). Interrater agreement for assigning a map to a structure type was 100%.

Design

To examine the equivalence of the two mapping techniques we used a 4 X 2 (mapping sequence X occasion) design. We randomly assigned students to one of four mapping sequences across the two occasions. (a) CC—construct-a-map with created linking phrases, then construct-a-map again with created linking phrases (*n* = 27); (b) SS—construct-a-map with selected linking phrases, then with selected linking phrases again (*n* = 23); (c) SC—construct-a-map with selected linking phrases, then with created linking phrases (*n* = 22); or (d) CS—construct-a-map with created linking phrases, then with selected linking phrases (*n* = 22). The elapsed time between occasions was 7 weeks, with no content-relevant instructional intervention during that time.

To learn more about differences in cognition, if any, elicited by the different mapping techniques, we randomly selected four students who received different test formats on Occasion 1 and Occasion 2 and asked them to think aloud as they were constructing their maps. These think-aloud observations provided information regarding the cognitive processes involved in the two concept mapping approaches (Ruiz-Primo, Shavelson, et al., 2001).

Procedure

All students were trained on the creation of concept maps, using a training procedure designed in previous studies (Ruiz-Primo, Schultz, et al., 2001). We gave students in the C and S conditions different training exercises to match the students' assessment types. At the end of the 20-minute training period, remaining questions were answered and student work was checked to verify that students understood what was expected on a concept map. Students were then given the Buoyancy Concept Mapping Activity, of the type C or S, depending on their random assignment.

To facilitate the map creation and to allow students to easily organize and rearrange the layout of their maps, we pre-printed each of the nine concepts on separate sticky-notes. The students placed each note (concept) on the blank paper provided, drew their connecting arrows, and finally redrew the final draft to another blank page. The redrawing gave students one more reflection step and provided the evaluators with a more readable final product. The students were given 30 minutes to construct their maps, and 10 minutes to redraw and check their final maps.

Results and Discussion

To determine whether the two concept map assessments were equivalent, we compared (a) total accuracy scores, (b) individual proposition scores, (c) proposition choice, (d) map structure complexity, (e) proposition generation rate, and (f) proposition generation procedures. Among these measures, (a) to (d) accounted for the concept map product and (e) and (f) accounted for concept map process.

Total Accuracy Score

According to classical test theory, equivalent assessments should have equal means and standard deviations, and strong correlations among themselves and with an outside criterion. We tested these criteria based on the total accuracy scores, as well as by comparing test-retest reliabilities.

Means and standard deviations. The means, standard deviations, and retest correlation of the total accuracy scores across the four groups and two occasions are presented in Table 3. The Levene test indicates that the variances are homogeneous between the two groups, $F = 0.22$, $p > 0.20$. That is to say, the C and S groups have homogeneous variances but different means. Students did not perform close to the expert map score of 30. From Occasion 1 to Occasion 2, the mean score for all groups except the CS group increased. A split-plot ANOVA revealed a statistically significant interaction effect between occasion and group, $F(3, 90) = 5.66$, $p < 0.01$. We suspected that the mean differences came from two possible sources, a concept mapping task learning effect and a format effect. Task learning effect refers to the score increase when students perform the task again. Format effect refers to a difference in score due to the change of assessment format. For C and S to be equivalent, the format effect should not exist.

The CC group and the SS group took the same assessment on Occasion 1 and Occasion 2. A Tukey post-hoc test showed the CC group's mean score increased by 2.93 points ($p < .05$) and the SS group's mean score increased by 2.64 points ($p < .05$).

Table 3

Means and Standard Deviations by Concept Mapping Technique and Occasions and Corresponding Correlations

Type	*n*	Occasion 1		Occasion 2		Correlations between Occasions 1 & 2
		M	*SD*	*M*	*SD*	
CC	27	17.26	8.32	20.19	10.67	.806**
SS	22	11.91	9.48	14.55	11.80	.827**
SC	23	13.09	8.38	19.09	9.04	.618**
CS	22	19.36	9.61	16.09	6.88	.526**
Total	94	15.48	9.27	17.64	9.93	

Note. Total score of the expert map was 30. CC = created-linking-phrases technique, Occasions 1 and 2; SS = selected-linking-phrases technique, Occasions 1 and 2; SC = selected... / Occasion 1, created... / Occasion 2; CS = created... / Occasion 1, selected... / Occasion 2.

** $p < 0.01$.

The two groups' mean scores increased to a similar extent, and we considered the increase to be a task learning effect.

The SC and CS groups changed task format from Occasion 1 to Occasion 2. Unlike the CC and SS groups, the mean score of the CS group decreased 3.27 points ($p < .05$); students in this group received a lower score on Occasion 2 when they took the S-format assessment. In contrast, the mean score of the SC group increased 6 points from Occasion 1 to Occasion 2, which is much higher than the task learning effect increase occurring in the CC or SS groups. The decrease from C to S and the increase from S to C led to a disordinal interaction, indicating the existence of the format effect. This provides evidence that the two assessment techniques are not equivalent.

To further examine the format effect, we focused on Occasion 1, because no task learning effect existed on Occasion 1. The four groups differed significantly in means (see Table 3), $F(3, 90) = 3.47$, $p < .05$. Significant differences existed between SS and CS ($p < .05$), as well as between CC and SS ($p < .05$). However, no mean difference existed between CC and CS, or SS and SC. That is, the mean of the C group (CC and CS) on Occasion 1 significantly differed from the mean of the S group (SS and SC). Therefore, when analyzing the data on Occasion 1, we treat the CC and CS groups as one "C" group and treat the SC and SS groups as one "S" group. Similarly, in the following analyses, to avoid a task-learning effect, comparisons between C and S are based on the two "combined" groups on Occasion 1 unless otherwise noted.

Correlation and reliability. Equivalent assessments should also have high correlations when they are used as parallel tests. Table 3 also supplies the correlation between scores on two different occasions. To differentiate the reliabilities, we refer to the reliabilities of CC and SS as the coefficient of stability, because the same tests are administered on two separate occasions; we refer to the reliability of SC and CS as the coefficient of delayed equivalence, assuming that C and S are parallel forms of a test. We found the stabilities of CC and SS to be very high, but the equivalences of CS and SC to be less so. Further analysis reveals that the stabilities of CC and SS were not significantly different from each other ($p > .05$), and the equivalences of SC and CS were not significantly different from each other ($p > .05$). Therefore, we combined CC and SS, and found that the pooled stability coefficient was .766 ($n = 49$); the pooled equivalence coefficient of SC and CS was .460 ($n = 45$). They differed from each other significantly, $z = 2.39$, $p < .01$. This result suggested nonequivalence between C and S.

Correlation with outside criterion. Moreover, equivalent assessments should be similarly correlated with an outside criterion. To examine the two assessments' equivalence based on this standard, we further calculated the correlation between the two assessments' scores and an outside criterion, the teacher's rating of her students' performance in science. The correlation of the teacher's rating with C is only .243 ($p > .05$), whereas the correlation with S is .551 ($p < .05$). However, even though the correlation between teacher's rating and S seems to be higher than that of C, the two correlations do not differ significantly, $z = 1.44$, $p > 0.05$. More studies need to be done to decide whether the assessments are equivalent in terms of their relationship with an outside criterion.

In summary, even though the two assessments' variances are equal and they may be correlated with an outside criterion similarly, they do not satisfy other equivalence criteria in classical test theory: C has higher mean scores than S, and delayed equivalence coefficients for C and S are lower than C and S stability coefficients. Therefore, we tentatively concluded that the C and S methods were not "classically" equivalent for the total accuracy score.

Individual Proposition Scores

When scoring students' maps, we noticed that one potential shortcoming of the S technique might be the limited number of linking phrase options, which might lead to bipolar scores. That is, examinees either "got it" or "missed it" when

choosing linking phrases for two terms. To test this hypothesis, we compared the distribution of proposition scores from the two assessments on Occasion 1 (see Table 4).

The distribution of individual proposition scores in the C condition suggested that partial scores (1 and 2 points) exist but the S technique may not be as sensitive to them (see Table 4). As expected, S scores were largely bipolar—students generally obtained either no-credit (0 points) or full-credit (3 points). For example, 40.9% of all the C propositions were given midrange scores of 1 or 2. In contrast, only 15.1% of the S propositions received a score of 1 or 2. Apparently the less constrained C technique provided students more opportunities to reveal their partial knowledge than S.

The average number of propositions constructed in the C condition (10.2) was significantly greater than that in S condition (7.6), t (92) = 3.85, $p < .01$. We believe that the C technique gave students more freedom to construct concept maps in the way that they wanted than did the S condition. In contrast, the S technique prevented the students from fully expressing their knowledge, especially their partial knowledge. Our subsequent analyses of students' proposition choices, map structure complexity, and think-aloud protocols provided more information about this conjecture.

Table 4

Frequency of Individual Proposition Scores From Occasion 1

Proposition scores	Frequency	Percent (%)	Cumulative percent (%)
Group C (n = 49)			
0.00	105	20.1	20.1
1.00	117	22.4	42.4
2.00	97	18.5	61.0
3.00	204	39.0	100.0
Total	523	100.0	
Group S (n = 45)			
0.00	142	39.0	39.0
1.00	33	9.1	48.1
2.00	22	6.0	54.1
3.00	167	45.9	100.0
Total	364	100.0	

Note. C = created-linking-phrases technique; S = selected-linking-phrases technique.

Proposition Choice

Our assessment supplied students nine concepts with which to construct their maps. If we regard two concepts with one linking phrase as one proposition choice regardless of the direction of the relationship between the concepts (i.e., it doesn't matter which way the arrow is pointing), potentially, 36 proposition permutations can be constructed for a single map with nine concepts. Of course, not all possible propositions make sense scientifically. Figure 3 displays the propositions constructed by more than 50% of the examinees in either condition on Occasion 1. The "popular" propositions varied across the C and S groups. For example, propositions of "density-mass" and "density-volume" were quite popular in the S group, but were not frequently constructed in the C group. In contrast, "mass-wood" and "water-wood" were much more popular in the C group than in the S group.

A close examination of the students' maps revealed that many students in the C condition constructed the proposition "wood floats on water." Since our experts did not regard this proposition to be universally true (some wood sinks), we did not expect "water-wood" to be a scientifically important proposition when designing the assessment. Therefore, that corresponding linking phrase was not supplied in the

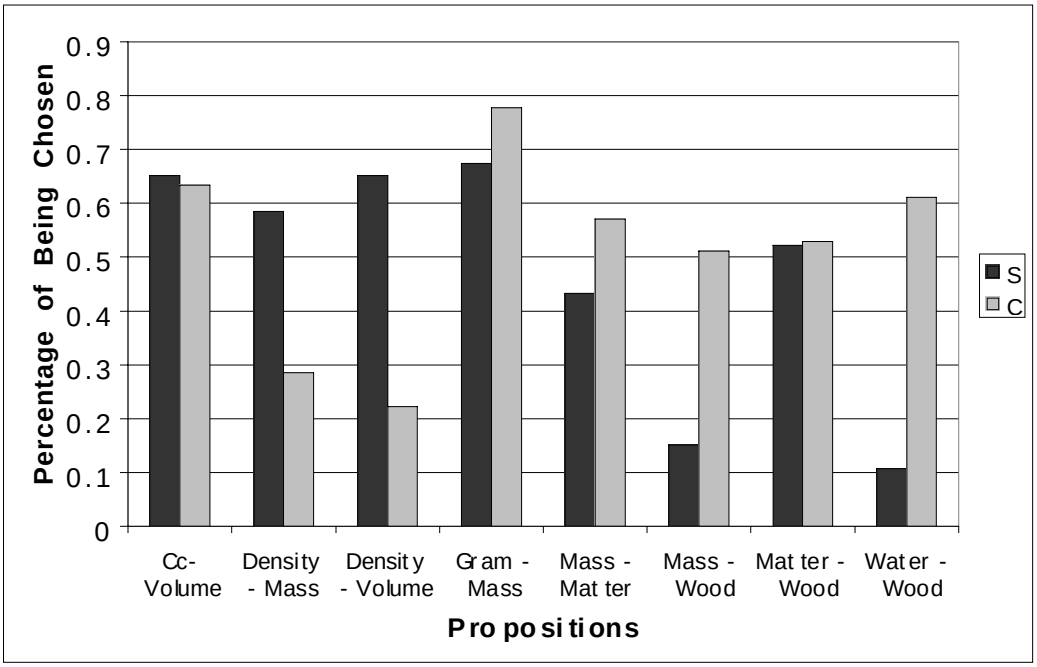

Figure 3. Popular propositions in two concept map techniques.

S condition. The lack of availability of this kind of linking phrases might have prevented the students from choosing their "favorite" or relevant propositions in the S condition. However, on the other hand, the relationships "density-mass" and "density-volume" are regarded as important scientific propositions and were included in the linking phrase list for the S condition. Given the differences between the C and S conditions with respect to these propositions, it appears that, in the S condition, students were prompted to choose them, even though they may not have done so spontaneously. Apparently, S students benefited from the linking phrases supplied when establishing the relationship among volume, density, and mass. This was as we had expected: that the S condition would be both limiting and prompting students. In conclusion, the C and S techniques elicited different propositions. From this perspective, C and S were not equivalent in eliciting students' knowledge structures.

Map Structure Complexity

We characterized maps according to their structure type. Table 5 provides the structure type distribution for the C and S techniques. For simplicity, we treated network as complex structure and all non-network structure as simple structure. Overall, more students in the C condition created a complex structure (55.1%) than in the S condition (26.7%), and fewer students applied a simple structure in the C condition than in the S condition. This pattern was statistically significant, χ^2 (1, N = 94) = 7.81, $p < 0.05$.

To further examine the mapping techniques' influence on structure types, we analyzed changes in structure type from Occasion 1 to 2 (see Table 6). For groups that repeated their assessment type across both occasions, the majority of students in CC (77.8%) and SS (77.3%) constructed maps with consistent structure complexity over time. A smaller number of students in the two groups changed their map structures either from simple to complex or from complex to simple.

The changes in map structure for the SC and CS groups, however, dramatically differed from that on the SS and CC groups. From S to C, students' map structures either remained the same (73.9%) or became more complex (26.1%), whereas from C to S, the structures either stayed the same (50%) or became simpler (50%). The trend is so overwhelming that no single exception exists. Students tended to construct concept maps with more complex structures in C than in S, providing evidence of the nonequivalence of the C and S techniques in eliciting students' knowledge

Table 5

Two Groups' Concept Map Structure Types on Occasion 1

Structure / Type	Group type C (n = 49) %	Group type S (n = 45) %
Simple		
Linear	4.1	17.8
Tree	28.6	31.1
Circle	4.1	17.8
Hub / Spoke	8.2	6.7
Complex		
Network / Net	55.1	26.7

Note. The values represent the percentages of the students using certain structures within groups. C = created-linking-phrases technique; S = selected-linking-phrases technique.

Table 6

Concept Map Structure Change From Occasion 1 to Occasion 2

Structure change type	Group type CC (n = 27) %	Group type SS (n = 22) %	Group type SC (n = 23) %	Group type CS (n = 22) %
Simpler	11.1	13.6	0.0	50.0
Same	77.8	77.3	73.9	50.0
More complex	11.1	9.1	26.1	0.0

Note. The values represent the percentages of the students making certain structure changes within groups. CC = created-linking-phrases technique, Occasions 1 and 2; SS = selected-linking-phrases technique, Occasions 1 and 2; SC = selected... / Occasion 1, created... / Occasion 2; CS = created... / Occasion 1, selected... / Occasion 2.

structures. And if, as claimed, the structure of the concept maps reveals differences between novices and experts (more structure, more knowledge), then our finding corroborates the total accuracy score findings—students show more of what they know in the C condition than the S condition.

Proposition Generation Rate

To understand the processes evoked during map construction, we videotaped four students concurrently verbalizing their thoughts while constructing their maps. When analyzing the think-aloud data, we focused on the overall pattern in the map construction processes.

We reviewed the video and recorded the elapsed time between proposition generations. Figure 4 represents the four students' proposition generation processes under the C and S conditions. Each point represents the generation of a proposition—the moment when a student recorded the proposition on the paper. Table 7 displays the average rate of proposition generation (propositions per minute).

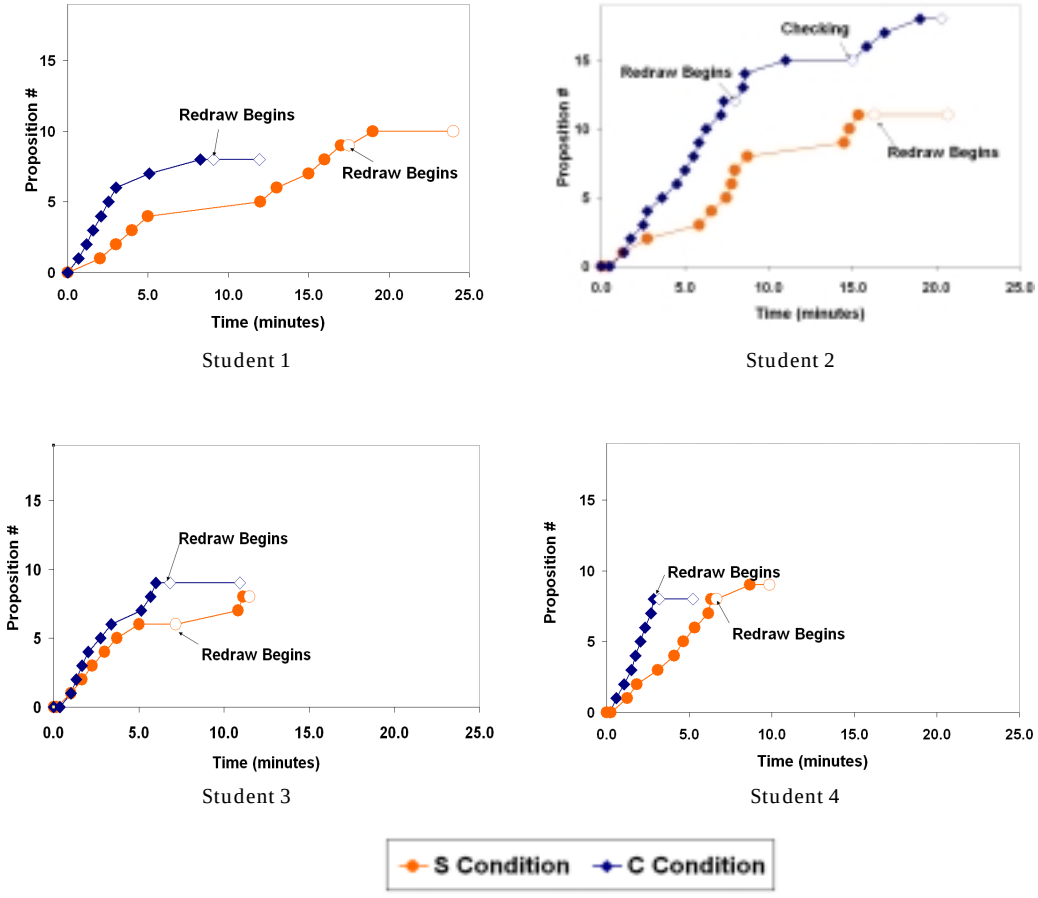

Figure 4. Cognitive process of four students.

Table 7

Comparison of Proposition Generation Rates

	Concept map type	
Student	C	S
1	0.67	0.42
2	0.89	0.53
3	0.92	0.70
4	1.54	0.91

Note. The values represent the average rate of proposition generation (proposition / min). C = created-linking-phrases technique; S = selected-linking-phrases technique.

Both Figure 4 and Table 7 show that the four students consistently constructed their propositions more slowly in S than in C. Recall that supplying linking phrases (a) may remind or prompt students while constructing their maps, and (b) may limit or delay students in map construction because of the mediating selection process. The generation rate comparison suggested that supplying linking phrases constrained most students more than helped them with constructing maps, slowing down students' map construction in S.

Proposition Generation Procedures

Students' think-aloud protocols can be used to illustrate the cognitive procedures leading to the proposition generation rate differences. In C, Student 4 verbalized that he "looked at the words and found the best solutions." In contrast, for the S condition he picked a pair of concepts, thought of phrases, and scanned the phrase list supplied to see if he could find a match. If he could not find what he needed, he tried to see if there was a close match. Matching mediated the map construction in the S condition. Figure 5 illustrates the processes applied by the small student sample in C and S. Compared to the C condition, the S condition has an extra "checking linking phrases" process either before the students thought of relationships or after, which may have slowed down map construction.

Additionally, students' comments after completing the maps shed light on the reason for the difference in proposition generation rates. When working in the S condition, Student 2 mentioned, "I used everything, but there could be some other relations." Student 3 had similar comments when she was asked to compare two

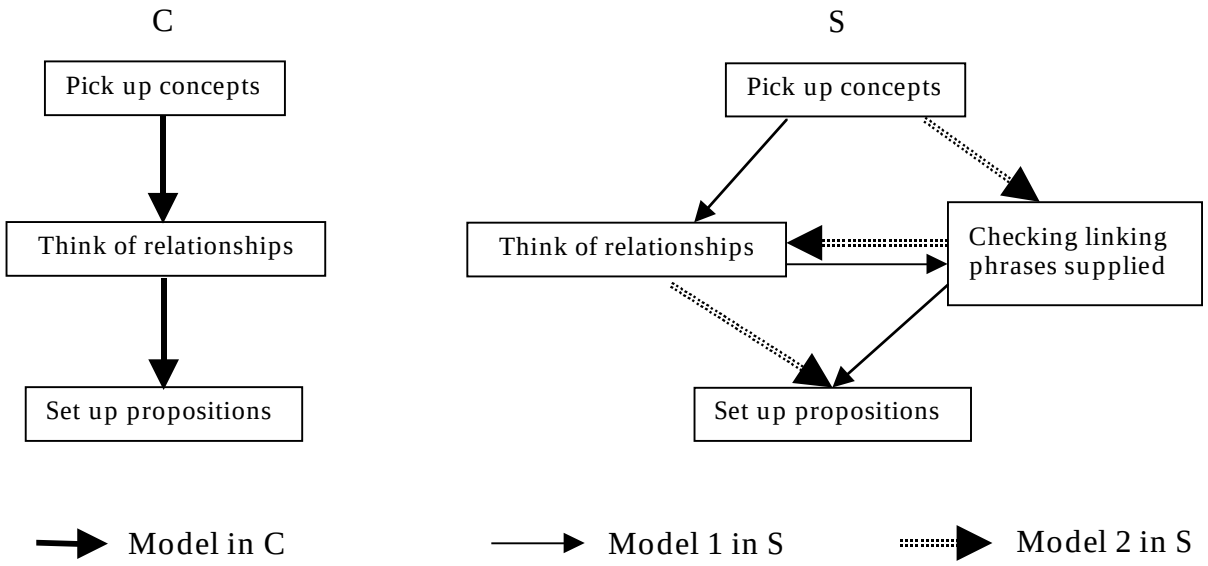

Figure 5. Procedural of concept map construction in C and S conditions.

techniques after she finished S: "The other one [C] is easier . . . you can make any link you want . . . I thought of a lot of things that I could put between them that weren't there . . ." Students were limited by the linking phrases supplied when they constructed their maps, which may explain the difference between C and S mapping techniques. Overall, our observation on the concept map construction process shows that the C and S techniques elicited different cognitive processes. This adds more evidence to our conclusion that techniques C and S were not equivalent.

Conclusions

Of all the various designs for concept map assessments, the construct-a-map with created linking phrases (C) is regarded as the benchmark of concept map assessment (Ruiz-Primo, Shavelson, et al., 2001; Ruiz-Primo, Schultz, et al., 2001). Construct-a-map with selected linking phrases (S) is considered an effective technique that addresses the challenge of scoring construct-a-map assessments (Klein et al., 2002). In this study, we examined the equivalence of these two concept mapping techniques in terms of their products and processes, both quantitatively and qualitatively.

When comparing concept map products, we found that C scores and S scores had similar variances and might be similarly correlated with an outside standard. However, compared with C, in the S condition, mean scores were lower, individual proposition scores were more polarized, map structures were simpler, and fewer and different propositions were generated. Furthermore, in stability and delayed equivalence–forms reliability, when C and S were used as parallel tests, their coefficients of equivalence were lower than for the CC's or SS's coefficients of stability.

A comparison of the cognitive processes evoked in C and S revealed that when students constructed the maps, students in C and S followed different procedures. Students in S spent extra effort and time matching what linking phrases they *wanted* to use with what they *could* use to construct their maps. Consequently, students in the S condition constructed their maps more slowly than students in the C condition.

We concluded that the two concept map assessments were not equivalent in both product and process. The two concept map task formats elicited different student responses and representations of the students' declarative knowledge structure.

The C condition might be better than S at capturing students' partial knowledge and misunderstandings due to its lack of constraints. However, we also noticed that C was much harder to score than S because of C's open-ended design. To express the same meaning, students used a variety of creative expressions, preventing us from scoring them automatically using a predetermined database of potential responses. To make it even more challenging, some students were not proficient in English. Consequently, some linking phrases in C suffered grammatical errors and were often difficult to understand. Since language skill was not our assessment target, we did not want students to lose credit due to their lack of language proficiency. Therefore, we had to make an informal judgment as to the intended meaning underneath their awkward wordings. As a result, numerous diverse linking phrases created by the students led to significant scoring challenges and a significant interrater disagreement. Also high interrater reliability is more difficult to get in C than S. Automatic scoring of such an open-ended task would require the development of a very large (and adaptive) database of possibilities, with rater intervention when new phrases emerged. The practicality of such an approach for large-scale assessments is doubtful, though not impossible.

Therefore, the S condition, if constructed properly, might be a promising approach for designing and implementing concept maps in large-scale assessments—if we still want to use a graphical approach. From this perspective, the S mapping technique may hold a position similar to that of multiple-choice tests. That is, multiple-choice tests still play an irreplaceable role in large-scale assessment, although they are criticized for missing important aspects of students' achievement. Trade off is always needed in reality. On the other hand, even though the C condition is difficult to score, it is superior to the S condition in capturing students' conceptual understanding and knowledge structure. Accordingly, C might be an effective tool for *formative assessment* in a classroom setting, where fully understanding a student's current thinking is more important than scores (Black & Wiliam, 1998). Recognizing that no assessment is perfect, assessment researchers must focus on characterizing each assessment's qualities, and thoughtfully recommend its appropriate use.

Limitations and Suggestions

Reflecting on our study, we realize it could be improved in several ways: (a) To examine the cognitive processes underlying map construction, a larger and more representative sample should be selected. For example, the distribution of gender, performance level, and group type (CC, CS, SC, SS) should be considered in sample selection. (b) To examine the relationship between concept map scores and an outside standard, a standardized multiple-choice or short answer test might be used in addition to teachers' overall ratings. (c) To examine the generalization of our conclusions, content other than density and buoyancy should be used.

Moreover, based on the findings of our study, we highlight some directions for improving construct-a-map as an assessment tool. One is related to the scoring system, and the other is related to task design.

First, how can constructed maps be scored fairly? The open-ended nature of a construct-a-map results in a superior approach for capturing students' knowledge structures. However, the openness leads to great uncertainty in map structure, the number of propositions, and the proposition choices. As a result, it is difficult to score maps fairly. Propositions can be scored in many ways, but no scoring approach is perfect. A total accuracy score was recommended by several studies and was also applied in our study. However, it suffers shortcomings when it is used to score constructed maps: When the propositions and number of proportions are

uncertain, students might reach the same score in different ways, which a total accuracy score cannot differentiate. For example, suppose in a construct-a-map test, student A is very conservative or concise—she only constructs 5 propositions, where each proposition is given a score of 3 (individual proposition score). In contrast, student B uses trial and error to construct 15 low-quality propositions, with each receiving a score of 1. As a result, they obtain the identical total accuracy score: 15. In this case, we cannot differentiate student A and B by their total accuracy scores.

Two potential solutions could be applied to solve this problem. First, set a maximum proposition number. For example, require examinees to construct at most 10 propositions that they think are the most important or meaningful as they "think like a scientist." We expect that a good performer should be able to tell important propositions from unimportant ones, essential ones from trivial ones. By constraining the maximum total proposition number, we might be able to better differentiate students according to their different structural knowledge levels. The second possible solution is to score only "key" propositions. Instead of regulating the maximum number of propositions, assessors encourage students to construct as many propositions as possible, but they only score the key propositions—ones that are most important in a domain. Here, key propositions refers to the propositions existing in expert or criterion maps. Some researchers have also called these *mandatory* propositions (Ruiz-Primo & Shavelson, 1996a). We suspect that the second way might work for younger children better than the first one, because younger students may not be able to differentiate importance from unimportance well. Moreover, it might be too demanding to have young children keep many requirements in mind simultaneously.

Our second suggestion is related to how to improve the linking phrase design in the S mapping technique. In our study, we derived linking phrases from the criterion maps constructed by several experts. This method led to two problems: (a) Some students did not understand the linking phrases well, and (b) students were very likely to obtain bipolar scores, which do not reflect partial knowledge. To solve these two problems, in addition to using experts' scientific knowledge, assessors might also select from students' C maps linking phrases that represent partial knowledge or misunderstanding. We also need to consider students' language expertise: linking phrases that are more understandable to children may need to be included. By taking these careful steps in the design of the linking phrase

list, we might be able to increase the utility of the S technique in eliciting partial understanding among students.

Concept mapping remains an exciting arena for assessment design research and application. Further studies to examine the nature of various approaches to concept mapping for assessment will highlight new possibilities, and the merging of open-ended assessment design with emerging computer automation capability will unlock the full potential of concept mapping to address the needs of classrooms for formative assessment and of large scale education accountability systems.

References

Astin, L. B., & Shore, B. M. (1995). Using concept mapping for assessment in physics. *Physics Education, 30*, 41-45.

Ayala, C. C., Yin, Y., Shavelson, R. J., & Vanides, J. (2002, April). *Investigating the cognitive validity of science performance assessment with think alouds: Technical aspects.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

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 Practice, 17*(3), 37-45.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139-148.

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

Dansereau, D. F., & Holley, C. D. (1982). Development and evaluation of a text mapping strategy. In A. Flammer & W. Kintsch (Eds.), *Discourse processing* (pp. 536-554). Amsterdam: North Holland.

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

Herl, H. E., O'Neil, H. F., Chung, G. K. W. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15*, 315-333.

Kinchin, I. M. (2000). Using concept maps to reveal understanding: A two-tier analysis. *School Science Review, 81*(296), 41-46.

Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., Herl, H. E., & O'Neil, H. F. (2002). *Examining the validity of knowledge mapping as a measure of elementary students' scientific understanding* (CSE Tech. Rep. No. 557). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

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

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36,* 475-492.

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

Nicoll, G., Francisco, J., & Nakhleh, M. (2001). A three-tier system for assessing concept map links: A methodological study. *International Journal of Science Education, 23,* 863-875.

Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching, 27,* 937-949.

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

O'Neil, H. F., Chung, G. K. W. K., & Herl, H. E. (1999). *Computer-based collaborative knowledge mapping to measure team processes and team outcomes* (CSE Tech. Rep. No. 502). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Teaching.

Pottenger, F. M., & Young, D. B. (1992). *The local environment: FAST 1 Foundational Approaches in Science Teaching*. Honolulu: University of Hawaii at Manoa, Curriculum Research and Development Group.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38,* 260-278.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996a, April). *Concept-map based assessment: on possible sources of sampling variability*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996b). Problem 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., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7,* 99-141.

Rye, J. A., & Rubba, P. A. (2002). Scoring concept maps: An expert map-based scheme weighted for relationships. *School Science and Mathematics, 102*(1), 33-44.

Shavelson, R. J. (1987). Teachers' judgments. In M. J. Dunkin (Ed.), *The international encyclopedia of teaching and teacher education* (pp. 486-490). New York: Pergamon.

Yin, Y., Ayala, C. C., & Shavelson, R. J. (2002, April). *Student's problem solving strategies in performance assessments: hands on minds on.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.