Comparison of Two Concept-Mapping Techniques:
Implications for Scoring, Interpretation, and Use

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

School of Education, 485 Lasuen Mall, Stanford University, Stanford, California 94305-3096

Received 12 May 2003; Accepted 8 June 2004

Abstract: 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. Both their products and processes are compared quantitatively and qualitatively as to 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 technique is better than the S technique in capturing students’ partial knowledge, even though the S can be scored more efficiently than C. Based on their characteristics, if used as an assessment tool, the C technique is more suitable for formative assessment while the S technique is a better fit for large-scale assessments. © 2005 Wiley Periodicals, Inc. J Res Sci Teach 42: 166–184, 2005

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, & Far, 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 concept map includes nodes (terms or concepts), linking lines (usually with a unidirectional 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, 1996). While 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, e.g., 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 mapping techniques are best suited for which assessment purposes? Finally, can different mapping techniques be scored more or less effectively and efficiently?

Ruiz-Primo and Shavelson (1996) raised questions about the reliability and validity of concept maps as assessment tools that have been explored by recent research (e.g., Kinchin, 2000; Klein, Chung, Osmundson, Herl, & O’Neil, 2001; McClure, Sonak, & Suen, 1999; Nicoll, Francisco, & Nakhleh, 2001; Ruiz-Primo, Schultz, Li, & Shavelson, 2001a; Ruiz-Primo, Shavelson, Li, & Schultz, 2001b; Rye & Rubba, 2002). Among those studies, some compared different concept-mapping tasks (e.g., Ruiz-Primo et al., 2001b), some explored different scoring systems (e.g., Kinchin, 2000; Klein et al., 2001; 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 et al., 2001a, 2001b). Our study is an extension of this line of research. More specifically, we examined the equivalence of two concept-map assessment tasks: (1) construct-a-map with created linking phrases (C), and (2) 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 concept terms; students need to select and assemble the concepts and linking phrases.

The C mapping technique has been characterized as the gold standard of concept maps (Ruiz-Primo et al., 2001a, 2001b). Compared with the fill-in-a-map technique (in which students fill in a pre-drawn map), the C technique (1) more accurately reflected differences of students’ knowledge structures; (2) provided greater latitude for demonstrating students’ partial understanding and misconceptions; (3) supplied students with more opportunities to reveal their conceptual understanding; and (4) elicited more high-order cognitive processes, such as explaining and planning. However, because of the range and diversity of students’ self-created linking phrases, the C 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., 2001; 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. (2001) 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 technique and the constraints of the S technique, the question remains: Are the two techniques equivalent? The goal of our study is 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: (1) product vs. process; and (2) quantitative vs. qualitative.
Concept-map products are the result of the task demands imposed on the student. In our study, the products were the concept maps drawn on a piece of paper. The concept map product variables, derived from students’ drawn concept maps, can be both quantitative (total proportion accuracy score and individual proportion accuracy score) and qualitative (proposition choice and structure complexity). Accordingly, quantitative variables have numerical values, while qualitative variables have categorical values.

Concept map process refers to a student’s inferred cognitive activities elicited during the execution of the concept-map task (Ruiz-Primo et al., 2001b). A concept map process variable can be created from students’ think aloud while constructing their maps. Concept map process includes 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 et al., 2001b).

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 Product Variables

Construct-a-map assessments are challenging to score because students’ products vary greatly. The total number of propositions constructed by a student is uncertain and the structure of the map is unfixed. Therefore, to characterize a student’s concept map adequately, more than a one-dimensional scoring system is needed to evaluate the map comprehensively. To this end, number and accuracy of propositions and map structure have been of primary focus in most research studies (Herl et al., 1999; McClure et al., 1999; Kinchin, 2000; Nicoll et al. 2001). We also scored concept-map products based on their propositions and the structure of the concept map. We apply three variables to describe propositions (total accuracy score, individual proposition score, and proposition choice) and one variable to describe structure (structure complexity).

A proposition, defined as 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 could either be applied or not when assessors score students’ maps; the score 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.

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 it effectively shows differences in students’ knowledge structures (e.g., Ruiz-Primo & Shavelson, 1996, April). 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 were equivalent, they should produce the same mean score, variance among scores, and high test–retest reliability.

In addition to the total accuracy score, we also 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 and provide a linking phrase for their relationship. Assuming that students choose to establish relationships that they think are important or interesting, we examined students’ declarative knowledge by analyzing the set of propositions they constructed. If two assessment techniques are equivalent, a given student should choose to build up 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; Ruiz-Primo & Shavelson, 1996). In this study, we focused on the maps’ graphic feature, rather than Novak’s hierarchical structure. Kinchin (2000) proposed three concept-map structure types: spoke, chain, and net. He pointed out that a net structure is indicative of meaningful learning. He suggested this qualitative scheme could be quickly and easily used, providing teachers with a simple starting point for concept-map analysis.

### Table 1

*Summary of scores applied to propositions*

<table>
<thead>
<tr>
<th>Type of Score</th>
<th>With a Criterion Map</th>
<th>Without a Criterion Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>Stringent semantic content: based on exact link matches between student links and expert links (Herl, O’Neil, Chung, &amp; Schacter, 1999). Categorized semantic content score: based on students matching some set of possible links (Herl et al., 1999).</td>
<td>Linkage: the total number of links (Astin &amp; Shore, 1995; Herl et al., 1999; Lomask, Baron, Greig, &amp; Harrison, 1992, March). Good links: the total number of links showing good understanding (Astin &amp; Shore, 1995; Herl et al., 1999; Lomask et al., 1992, March).</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Weighted relationship: score is given to the individual relationship (proposition) based on its similarity to criterion one (Rye &amp; Rubba, 2002).</td>
<td>Total proposition accuracy: total sum of the quality scores obtained on all propositions (Ruiz-Primo &amp; Shavelson, 1996, April).</td>
</tr>
<tr>
<td>Proportion</td>
<td>Congruence: proportion of valid student link over all criterion links (Ruiz-Primo &amp; Shavelson, 1996, April).</td>
<td>Salience: proportion of valid student link over all student links (Ruiz-Primo &amp; Shavelson, 1996, April).</td>
</tr>
</tbody>
</table>

Authors of this paper named the three score types according to different scores’ features.
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’ map. We defined the five structure types as follows (see Figure 2): (1) linear propositions, which are chained together; (2) circular propositions, which are daisy-chained with the ends joined; (3) hub or spoke propositions, which emanate from a center concept; (4) tree propositions, a linear chain that has branches attached; and (5) network or net propositions, a complex set of interconnected propositions. Among them, the network or net structure is considered the most complex, while the linear structure is considered the simplest. All others fall in between. We also called the structure typology used in our study “structure complexity.”

As research has shown that experts, compared with novices, exhibit deeper, more connected and interrelated map structures (Mintzes et al., 1997), maps created by high performers should 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 Process Variables**

Previous studies have shown that students apply different strategies when completing fill-in-a-map and construct-a-map tasks (Ruiz-Primo et al., 2001b). Correspondingly, if the two construct-a-map, C and S, techniques 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 et al., 2001b), and performance

![Figure 2. Structure complexity. Five key concept-map structures.](image-url)
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 talking 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, while proposition generation procedures refer to the 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: (1) a list of linking phrases could provide students with hints, reminding them of the scientific relationships between concepts; and (2) 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 processes by comparing their proposition generation rates with the two techniques. 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 any were there.

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

A total of 92 eighth-graders from the California Bay Area participated in the study; 46 were girls and 46 boys. The students drawn largely from upper middle class homes, belonged to six middle-school science classes taught by the same teacher.

### Table 2

<table>
<thead>
<tr>
<th>Comparison Targets</th>
<th>Comparison Questions</th>
<th>Comparison Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Map Product</td>
<td>Do the different technique scores have the same mean, standard deviation, and test–retest reliability?</td>
<td>Total Accuracy Score</td>
</tr>
<tr>
<td></td>
<td>Do the different technique scores reveal the same knowledge level of a given student?</td>
<td>Individual Proposition Score</td>
</tr>
<tr>
<td></td>
<td>Do the different techniques lead students to construct similar propositions?</td>
<td>Proposition Choice</td>
</tr>
<tr>
<td></td>
<td>Do the different techniques elicit a similar concept map structure?</td>
<td>Structure Complexity</td>
</tr>
<tr>
<td>Concept Map Process</td>
<td>Do they elicit a similar cognitive process?</td>
<td>Proposition Generation Rate</td>
</tr>
<tr>
<td></td>
<td>Proposition Generation Procedure</td>
<td>Proposition Generation Procedure</td>
</tr>
</tbody>
</table>
Before 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 each student’s science achievement level as high, medium, or low based on student work and her observations in science class. Of the 92 students, 17 students were ranked low, 36 were ranked medium, and 39 were considered to be top performing. 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 concepts related to buoyancy and instructed them 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, 1996). 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 nine 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. Finally, 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 four-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 inter-rater reliability for the scoring of this database was initially quite low, <.70 on 50 randomly selected propositions.

After discussions, we created detailed rules. First, 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 demonstrated appropriate conceptual understanding, even if they did not use the exact linking words expected. Second, 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. Third, 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. Fourth, 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’ inter-rater reliability for 50 randomly selected propositions was .92 on S map propositions and .81 on C map propositions because of the great diversity of propositions created in C map. Three raters, consisting of two graduate students and one science education professor, reached agreement on the scores for all the unique propositions, and built up our master scoring list. Having transferred each student’s concept map propositions into the Excel database, we used the master scoring-list 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). Inter-rater agreement for assigning a map to a structure type was 100%.

Research Design

To examine the equivalence of the two mapping techniques we used a 4\times2 (mapping sequence \times occasion) design. We randomly assigned students to one of four mapping sequences across the two occasions. (1) CC: construct-a-map with created linking phrases then construct-a-map again with created linking phrases (n = 26); (2) SS: construct-a-map with selected linking phrases and then with selected linking phrases again (n = 21); (3) SC: construct-a-map with selected linking phrases then with created linking phrases (n = 23); or (4) CS: construct-a-map with created linking phrases then 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 et al., 2001b).

Procedure

All students were trained on the creation of concept maps, using a training procedure designed in previous studies (Ruiz-Primo & Shavelson, 1996, April; Ruiz-Primo et al., 2001a). We gave students in the C and S condition 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 they 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, and drew their connecting arrows. As in previous studies, students were asked to redraw their final maps on 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 map.

Results and Discussion

To determine whether the two concept-map assessments were equivalent, we followed our framework and compared the two mapping techniques on the six variables: total accuracy scores, individual proposition scores, proposition choice, map structure complexity, proposition generation rate proposition, and generation procedures.

**Total Accuracy Score**

According to classical test theory, equivalent assessments should have equal means, 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 mean score increased from occasion 1 to occasion 2 except for the CS group. A split-plot analysis of variance (ANOVA) revealed a statistically significant interaction effect between occasion and group, F (3, 88) = 6.55, p < .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 a score increase when students perform the task again. Format effect refers to a difference in score

<table>
<thead>
<tr>
<th>Type</th>
<th>n</th>
<th>Occasion 1 M</th>
<th>Occasion 1 SD</th>
<th>Occasion 2 M</th>
<th>Occasion 2 SD</th>
<th>Correlations Between Occasion 1 &amp; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>26</td>
<td>17.42</td>
<td>8.44</td>
<td>20.23</td>
<td>10.87</td>
<td>.808*&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>SS</td>
<td>21</td>
<td>12.00</td>
<td>9.70</td>
<td>13.29</td>
<td>10.47</td>
<td>.827*&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>SC</td>
<td>23</td>
<td>13.09</td>
<td>8.38</td>
<td>19.09</td>
<td>9.05</td>
<td>.618*&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>CS</td>
<td>22</td>
<td>19.36</td>
<td>9.61</td>
<td>16.09</td>
<td>6.88</td>
<td>.526*&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Total</td>
<td>92</td>
<td>15.48</td>
<td>9.27</td>
<td>17.64</td>
<td>9.93</td>
<td></td>
</tr>
</tbody>
</table>

Note. Total score based on expert map was 30.

*<sup>p</sup> < .01.
<sup>a</sup>Stability.
<sup>b</sup>Equivalence coefficient.
because of 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. We considered the increase in the CC and SS groups to be a task learning effect. A Tukey’s post hoc test showed the mean score of the CC group significantly increased from occasion 1 to occasion 2 ($p < .05$).

The SC and CS groups changed task format from occasion 1 to occasion 2. In contrast to the CC and SS groups, the mean score of the CS group decreased ($p < .05$); 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 greatly from occasion 1 to occasion 2 ($p < .05$), which is greater than task learning effect observed in the CC and SS groups. The decrease from C to S and the increase from S to C led to a disordinal interaction, indicating the existence of a format effect. This provides evidence that the two assessment techniques were not equivalent.

Because there was no task learning effect on occasion 1, we focused on occasion 1 to further examine the format effect without the interference of learning effect. We grouped all the students who took C on occasion 1 (CC and CS) as “C” ($n = 48$) and all the students who took S on occasion 1 (SC and SS) as “S” ($n = 44$). Levene’s test indicates that the variances were homogeneous between the C and S groups, $F = 0.037$, $p > .20$. That is, two techniques’ score standard deviations were very close. However, the two techniques had different means, $t(90) = 3.08$, $p < .01$.

Similarly, to avoid the learning effect, in the following analyses (i.e., correlation analyses with an outside standard, individual accuracy score distribution, proposition choice, and structure complexity), we compare C and S based on the two combined groups on occasion 1.

**Stability and Equivalence Coefficient.** 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 coefficients of stability, because they are the same tests 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 groups to be less so. Further analysis reveals that the stability 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 $.82$ ($n = 47$); the pooled equivalence coefficient of SC and CS is $.58$ ($n = 45$). They differed from each other significantly, $z = 2.28$, $p < .05$. 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 the teacher’s rating of her students’ performance in science. The correlation of the teacher’s rating with C was only $.259$ ($p > .05$), while the correlation with S is $.551$ ($p < .05$). 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, $z = 1.63$, $p > .05$. More studies need to be done to decide whether they are equivalent in terms of their relationship with 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 criterion in classical test
theory: C has higher mean scores than S; 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) existed but the S technique may not be as sensitive to them (see Table 4). As expected, S scores were largely bipolar, that is, students generally obtained either no-credit (0 points) or full-credit (3 points). For example, 41.3% of all the C propositions were given mid-range scores of 1 or 2. In contrast, only 14.8% 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 (10.1) was significantly greater than that in S condition (7.6), \( t(90) = 3.68, p < .01 \). 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.

**Proposition Choice**

Our assessment supplied students with 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 was 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 the C or S condition on occasion 1. For simplicity, we defined those propositions as “popular”

<table>
<thead>
<tr>
<th>Group</th>
<th>Proposition Scores</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C ((n = 48))</td>
<td>.00</td>
<td>98</td>
<td>19.2</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>115</td>
<td>22.5</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>96</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>201</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>510</td>
<td>100.0</td>
</tr>
<tr>
<td>S ((n = 44))</td>
<td>.00</td>
<td>139</td>
<td>38.9</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>33</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>2.00</td>
<td>20</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>165</td>
<td>46.2</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>364</td>
<td>100.0</td>
</tr>
</tbody>
</table>
propositions. Nine propositions were popular in the C condition, while only five propositions were popular in S. That is, the linking phrases supplied in S might, in general, have prevented students from choosing students’ favored propositions in C. Moreover, 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, “buoyancy-wood,” “density-water,” “density-wood,” “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 can help us better understand the different patterns. For example, 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 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, 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 student’s knowledge structure.

**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 treat network as complex structure and all non-network structure as simple structure. Overall, more students in the C condition created a complex structure (54.2%) than in the S condition (28.9%) and fewer students applied a simple...
structure in the C than S condition. This pattern was statistically significant, $\chi^2 (1, N = 92) = 6.85, \ p < .05$.

To examine further the influence of mapping techniques on structure types, we analyzed changes in structure type from occasion 1 to 2 (see Figure 4). For the groups with same type of assessment across both occasions, CC and SS, the map complexity changes were similar. The majority of students constructed maps with consistent structure complexity over time (CC = 80.8%; SS = 80.9%). A smaller number of students in the two groups changed their map structures either from simple to complex or from complex to simple.

<table>
<thead>
<tr>
<th>Structure Type</th>
<th>Group Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C ($n = 48$)</td>
</tr>
<tr>
<td>Simple Linear</td>
<td>4.2</td>
</tr>
<tr>
<td>Tree</td>
<td>29.2</td>
</tr>
<tr>
<td>Circle</td>
<td>4.2</td>
</tr>
<tr>
<td>Hub/Spoke</td>
<td>8.3</td>
</tr>
<tr>
<td>Complex Network/Net</td>
<td>54.2</td>
</tr>
</tbody>
</table>

Note. The values represent the percentages of the students using certain structures within groups.

Figure 4. Percentage of change in map structure complexity across occasions.
The changes in map structure for the SC and CS groups, however, dramatically differed from that of the SS and CC groups. From S to C, students’ map structures either remained the same (73.9%) or became more complex (26.1%), while 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. We considered this as another piece of evidence about the nonequivalence of the C and S techniques in eliciting students’ knowledge structures. We concluded that the C techniques allows students to show more of what they know than the S technique. These findings triangulated with the results obtained with 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 5 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

![Graphs showing proposition generation rate](image-url)

**Figure 5.** Rate of proposition generation across four students under the two mapping technique conditions.
recorded the proposition on the paper. Table 6 displays the average rate of proposition generation (proposition per minute).

Both Figure 5 and Table 6 show that the four students consistently constructed their propositions more slowly in S than in C. Supplying the linking phrases may remind or prompt students while constructing their maps and/or 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 Procedure**

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 whether he could find a match. If he could not find what he needed, he tried to see whether there was a close match. Matching mediated the map construction in the S condition. Figure 6 illustrates the processes applied by the small student sample in C and S. Compared with 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.

![Diagram](image_url)

**Table 6**

*Comparison of proposition generation rate by technique*

<table>
<thead>
<tr>
<th>Student</th>
<th>Concept Map Type</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>.67</td>
<td>.42</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>.89</td>
<td>.53</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>.92</td>
<td>.70</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>1.54</td>
<td>.91</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>1.01</td>
<td>.64</td>
</tr>
</tbody>
</table>

*Note.* The values represent the average rate of proposition generation (proposition/min).
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 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 of 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 assessments (Ruiz-Primo et al., 2001a, 2001b). 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., 2001). 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, when C and S were used as parallel tests, their coefficients of equivalence were lower than for the CC 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. However, because of the small sample size (only four students), we cannot draw a decisive conclusion about the processes of C and S yet. More studies need to be done to decide whether the patterns found in our study can hold for a larger sample size.

We concluded that the two concept-map assessments were not equivalent in product and possibly not equivalent in process either. The two concept-map task techniques elicited different student responses and representation of the students’ declarative knowledge structure.

The C condition might be better than S at capturing students’ partial knowledge and misunderstandings because of its lack of constraints. However, we also found 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 because of 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. High inter-rater reliability is more difficult to attain 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, although 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 want to use concept maps as a tool to assess students’ knowledge. From this perspective, the S mapping technique may hold a similar position as multiple-choice tests—multiple-choice tests still play an irreplaceable role in large-scale assessment, although they are criticized for missing important aspects of students’ achievement. We may have to use the S technique in large-scale summative testing programs. In contrast, 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 & William, 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. First, to examine the cognitive processes underlying map construction, a larger and more representative sample of students should be selected. For example, the distribution of gender, performance level, and group type (CC, CS, SC, SS) should be considered in sample selection. Second, 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. Third, 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 structure. However, the openness leads to great uncertainty in its 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 accurate score cannot differentiate. For example, suppose in a construct-a-map test, student A is very conservative or concise; she only constructed 5 propositions, where each proposition was given a score of 3 (individual proposition score). In contrast, student B used trial and error to construct 15 low-quality propositions, with each receiving a score of 1. As a result, they obtained the identical total accuracy score: 30. 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, which 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 only scoring “key propositions.” Instead of regulating the maximum proposition numbers, assessors encourage students to construct as many propositions as possible, but they only score the key propositions that are most important in a domain. Here key propositions...
refer to the propositions existing in an expert or criterion map. Some researchers have also called them mandatory propositions (Ruiz-Primo & Shavelson, 1996, April). We suspect that the second way might fit 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: (1) some students do not understand the linking phrases well; and (2) students were very likely to obtain bipolar scores, which do not reflect partial knowledge. To solve these two problems, in addition to 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 will unlock the full potential of concept mapping to address the needs of classrooms for formative assessment and of large-scale summative accountability systems.

References


