Application of Generalizability Theory to Concept Map Assessment Research

Yue Yin \textsuperscript{a}, Richard J. Shavelson \textsuperscript{b}

\textsuperscript{a} Department of Educational Psychology, College of Education, University of Hawaii at Manoa.
\textsuperscript{b} School of Education, Stanford University,

Online Publication Date: 01 July 2008


To link to this article: DOI: 10.1080/08957340802161840

URL: http://dx.doi.org/10.1080/08957340802161840
Application of Generalizability Theory to Concept Map Assessment Research

Yue Yin
Department of Educational Psychology, College of Education
University of Hawaii at Manoa

Richard J. Shavelson
School of Education
Stanford University

In the first part of this article, the use of Generalizability (G) theory in examining the dependability of concept map assessment scores and designing a concept map assessment for a particular practical application is discussed. In the second part, the application of G theory is demonstrated by comparing the technical qualities of two frequently used mapping techniques: construct-a-map with created linking phrases (C) and construct-a-map with selected linking phrases (S). Some measurement facets that influence concept-map scores are explored and how to optimize different concept mapping techniques by varying the conditions for different facets is shown. It is found that C and S are not technically equivalent. The G coefficients for S are larger than those for C under the same condition. Furthermore, a decision study shows that fewer items (propositions) would be needed for S than C to reach desired level of G coefficients if only one occasion could be afforded. On the other hand, C seems to reveal students’ understanding about different concepts than S better. For practical purposes, one might prefer S because it is easier to score and produces higher reliability. However, this efficiency comes at the cost of validity. We would trade off validity and reliability for efficiency by including more propositions in C map.

Knowledge structure is regarded as an important component of understanding in a subject domain (Novak, 1990; Novak & Gowin, 1984). The knowledge structure of
experts and successful learners is characterized by an elaborate and highly integrated organization of concepts (Chi, Glaser, & Far, 1988; Mintzes, Wandersee, & Novak, 1997), which facilitate problem solving and other cognitive activities (Baxter, Elder, & Glaser, 1996). Consequently, the structure, not just the quantity, of a student’s knowledge, is considered as an important component of achievement to measure (e.g., Jonassen, Beissner, & Yacci, 1993; White & Gunstone, 1992).

Concept mapping is considered as a powerful tool for representing important aspects of students’ knowledge structures and helping teachers identify misconceptions (McClure, Sonak, & Suen, 1999; Ruiz-Primo & Shavelson, 1996). Once concept mapping is used as an assessment tool, often its technical quality becomes critical. However, it is not always clear how to evaluate reliability and validity of concept map assessment scores. In this article, Generalizability theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) is applied in examining the dependability of concept map assessment scores.

**CONCEPT MAPS**

A concept map is a network that includes nodes (terms or concepts), linking lines (usually with a uni-directional arrow from one concept to another), and linking phrases that 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 non-hierarchical).

Concept maps were originally proposed to be used as an instructional tool (e.g., Novak & Gowin, 1984) and later as an assessment tool as well (e.g., Jonassen et al., 1993; White et al., 1992). Concept maps hold promise in tapping students’ declarative knowledge structures of which traditional assessments are presumably not capable. This feature of concept maps attracted assessment researchers’ attention. Ruiz-Primo and Shavelson (1996) characterized the variation among concept map assessments in a framework with three components: a task that invites students to provide evidence for their knowledge structure in a content domain, a response form that students use to do the task, and a scoring system that the raters can use to evaluate students’ responses. For a comprehensive review of the variations, readers can refer to Ruiz-Primo and Shavelson (1996).

Even though thousands of concept map assessment permutations are possible, not all alternatives are suited for assessment. Ruiz-Primo and Shavelson (1996) pointed out that reliability and validity information about different mapping techniques should be supplied before concept maps are used for assessment. This study is one such effort. In particular, the feasibility of using G theory to evaluate
the dependability of concept map scores is first discussed. Then how G theory can be applied in this kind of research is illustrated by comparing two frequently used concept-mapping tasks: construct-a-map by creating linking phrases (C) and construct-a-map by selecting linking phrases (S).

FEASIBILITY OF USING G THEORY IN CONCEPT MAP ASSESSMENT RESEARCH

Technical Properties of Concept Map Assessments

Concept maps vary greatly from one another for both instruction and assessment. When the concept maps are used as an assessment, it becomes critical to narrow down options by finding reliable, valid, and efficient mapping techniques. Ruiz-Primo, Shavelson, and Schultz (1997, March) suggested four criteria for eliminating alternatives: “(a) appropriateness of the cognitive demands required by the task; (b) appropriateness of a structural representation in a content domain; (c) appropriateness of the scoring system used to evaluate the accuracy of the representation; and (d) practicality of the technique” (p. 7). Even though criterion (c) only talked about the scoring system, it was found that the accuracy of the scores was not only related to the scoring systems, but also related to task format. For example, using the same scoring form, some task formats might be scored more reliably and accurately than others (Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, 2005).

This article focuses on criteria (b) and (c), which have typically been gauged by traditional statistical analyses and classical test theory (CTT). Researchers have examined scores for inter-rater reliability/agreement (Herl, O’Neil, Chung, & Schacter, 1999; Lay-Dopyera & Beyerbach, 1983; Lomask, Baron, Greig, & Harrison, 1992, March; McClure, et al., 1999; Nakhleh & Krajcik, 1991); stability (Lay-Dopyera & Beyerbach, 1983); convergent validity—the correlation between concept map score and other assessment score in the same content domain (Anderson & Huang, 1989; Baker, Niemi, Novak, & Herl, 1991, July; Markham, Mintzes, & Jones, 1994; Novak, Gowin, & Johansen, 1983; Rice, Ryan, & Samson, 1998; Schreiber & Abegg, 1991); predictive validity (Acton, Johnson, & Goldsmith, 1994); equivalence of different scoring methods (McClure et al., 1999; Rice et al., 1998); and equivalence of different concept-map tasks (Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Yin et al., 2005).

Those studies have supplied important information about technical properties of different concept map tasks, response formats, and scoring systems, which can undoubtedly help to eliminate improper alternatives. However, the sources of variation for concept map assessments are so many that CTT cannot handle them simultaneously and efficiently.
Examining Concept Map Assessment Scores’ Technical Properties with G Theory

If an examinee’s concept map assessment score is viewed as a sample from a universe of the examinee’s concept map scores under all kinds of varying conditions—for example, tasks, response formats, and scoring systems—concept map assessments can be examined in the framework of G theory. Compared with CTT, G theory (a) can integrate conceptually and evaluate simultaneously test-retest reliability, internal-consistency reliability, convergent validity, and inter-rater reliability; (b) can estimate not only the influence of individual measurement facets, but also interaction effects; (c) permits us to optimize an assessment’s dependability (“reliability”) within given dollar and time cost constraints—for example, concept map assessment designers can obtain information about how many occasions, how many concepts, and how many raters are needed to reach a dependable result; and (d) can, when assessing students’ performance, supply dependability information about students’ absolute level of knowledge structure quality as well as rank order information.

Applying G theory in concept map assessment research is particularly suitable because concept map assessments involve a large number of sources of variation, such as concepts, propositions, task types, response formats, occasions, raters, and scoring systems (Ruiz-Primo & Shavelson, 1996). All these sources of variation can be considered as sources of error variation, or “facets,” in G theory.

Several researchers have applied G theory to examine the technical properties of concept map scores and so have added to our knowledge of concept map assessments. Ruiz-Primo, Schultz, & Shavelson (1996, April) used G study to compare three concept map assessment tasks: a concept map task without concepts supplied, a concept map task with one concept sample, and a concept map task with another concept sample. Two facets were examined in their analysis: rater and condition (concept sample). They compared three scoring systems’ generalizability over raters and conditions: total proposition accuracy—total sum of the quality scores obtained on all propositions; convergence—proportion of valid student links over all criterion links; and salience—proportion of valid student links over all student links. They found that across the three scoring systems: (a) Raters introduced negligible error variability into the measurement. (b) Students’ relative standing varied under different conditions. (c) The G coefficients were quite high for both relative decisions (i.e., norm-referenced decisions) and absolute decisions (i.e., criterion-referenced decisions). That is, concept map tasks with the scoring methods used can consistently rank students relative and absolute performance levels. (d) Proposition accuracy scores had the highest relative and absolute coefficients and the salience score had the lowest.

Besides concept sampling and raters, mapping task type, proposition, and occasion are important facets involved in concept map assessment. None have been examined in the framework of G theory.
Concept-map tasks vary greatly. A concept map task may supply nothing but a topic and ask students to construct the map from scratch. The task may supply concepts only and require students to construct the map with the concepts supplied, a.k.a., construct-a-map with created linking phrases. It may supply both concepts and linking phrases, and ask students to construct the map by assembling the concepts and linking phrases supplied, a.k.a., construct-a-map with selected linking phrases. It may supply a partially complete map and ask the students to fill in the nodes (concepts) or fill in the lines (relationships). Finally, a concept map task may require a certain structure for the map—for example, hierarchical or linear—or may leave the student to decide how to structure the map (Ruiz-Primo & Shavelson, 1996). The list of the variations in concept-map tasks is lengthy. To make a long story short, because different map tasks vary in their difficulty levels and features, task variation may lead to variability in the evaluation of a student’s declarative knowledge structure. Therefore, tasks can be regarded as a facet of the concept map assessment procedure and captured by G theory.

Each proposition in a concept map can be regarded as an item in a test, sampled from a subject domain. Different propositions may vary in difficulty level. Proposition sampling, then, can cause variation in the measures of the proficiency of students’ declarative knowledge structures. Notice that proposition sampling is similar to concept sampling in that they are both related to the variation due to concept sampling; however, concept sampling focuses on sampling at a macro level, analogous to alternate-form reliability in the classic test theory, whereas proposition sampling, analogous to internal consistency, focuses on sampling at a micro level. The two facets’ similarity and differences again show the strength of G theory in that it allows researchers to flexibly focus on the error type of interest in the analysis to meet specific needs in a single analysis.

When a concept map assessment is re-administered to a group of students, as in a test-retest design, occasions are sampled. Students may perform inconsistently when taking concept map assessments on different occasions (Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999). CTT treats the consistency over time as the stability of a test. G theory simply regards occasion sampling as another source of error and estimates it systematically.

The following section illustrates the application of G theory to address the dependability issues related to the three facets and the interaction among them in the comparison of two commonly used concept map assessments.

**COMPARISON OF TWO CONCEPT MAP TECHNIQUES BY G THEORY**

Construct-a-map with created linking phrases (C) and construct-a-map with selected linking phrases (S) are two commonly used concept map
techniques. In C students are given concept terms and asked to construct a map. Students have to create their own phrase for each concept pair that they have linked. In S students are given both linking phrases and concept terms to construct a map. That is, in addition to the concepts provided in a C map, students are also provided linking phrases from which to select in constructing their map.

C and S concept-mapping tasks are two frequently used techniques. The C mapping technique has been characterized as the gold standard of concept maps (Ruiz-Primo, Schultz, Li, & Shavelson, 2001; Ruiz-Primo, Shavelson et al., 2001). Compared with the fill-in-a-map technique (where students fill in a pre-drawn map), the C technique (a) more accurately reflected differences of students’ knowledge structures; (b) provided greater latitude for demonstrating students’ partial understanding and misconceptions; (c) supplied students with more opportunities to reveal their conceptual understanding; and (d) elicited more high-order cognitive processes, such as explaining and planning. However, due to the range and diversity of students’ self-created linking phrases, the C technique presents 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, Chung, Osmundson, Herl, & O’Neil, 2001). Because the number of propositions was bounded, computers could easily compare students’ maps with a criterion or expert map(s), typically created by educators, teachers, and/or domain experts (e.g., scientists). Klein et al. (2001) suggested that the computer made scoring straightforward and effective. This advantage is particularly appealing if concept maps are considered to be used as a potential large-scale assessment tool, for example, as part of the 2009 NAEP Science Framework.

Given the openness of the C mapping technique and the constraints of the S, the following questions were raised: Do the two techniques vary in their technical characteristics? For example, do they vary in stability and internal consistency? Would S mapping be more reliable than C mapping? What can be done with assessment design if certain reliability levels are needed? For different techniques, does the manner in which they are optimized vary? How can validity, reliability, and efficiency be balanced in the concept map assessment design? These questions are addressed in the framework of G theory. Factors influencing concept-map scores’ generalizability in a G study are explored and how to optimize different concept mapping techniques by varying the conditions for different measurement facets in alternative decision (D) studies are estimated.
METHOD

Participants

Ninety-two eighth-graders from the California Bay Area participated in the study. Forty-six were female. The students were drawn largely from upper middle class homes. They 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.

Research Design

To compare the two mapping techniques a 4×2 (mapping sequence × occasion) design was used. Students were randomly assigned to one of four mapping sequences across two occasions: (a) CS—construct-a-map with created linking phrases first then with selected linking phrases (N = 22); (b) SC—construct-a-map with selected linking phrases then with created linking phrases (N = 23); (c) CC—construct-a-map with created linking phrases twice (N = 26); or (d) SS—construct-a-map with selected linking phrases twice (N = 21). The science teacher of the participant students was asked to rate the students’ performance on a scale of high, medium, and low. Kruskal-Wallis test indicated that no significant difference existed among the four groups in terms of teachers’ ratings.

The elapsed time between occasions was seven weeks, with no instructional intervention related to the content assessed—in this case mass, volume, density, and buoyancy. Because all the students were taught by the same science teacher and the four groups were randomly assigned, the four groups were assumed to have similar science knowledge.

Mapping Techniques

In both the C and the S conditions, students were given nine concepts related to buoyancy and 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 an assessment design team working with curriculum designers, content experts, and a master teacher. The target curriculum was a unit on Density and Buoyancy from the Foundational Approaches to Science Teaching curriculum developed by the Curriculum Research and Development Group at the University of Hawaii (Pottenger & Young, 1996). By using an iterative selection process involving ranking and voting by the team members, nine concept terms were
YIN AND SHAVELSON

selected—WATER, VOLUME, CUBIC CENTIMETER (CC), WOOD, DENSITY, MASS, BUOYANCY, GRAM, and MATTER.

In the C condition, students created linking phrases. In the S condition, students selected linking phrases from a list that they could use (or re-use) to describe the relationships between concepts. This list was based on a criterion map created by the assessment design team (Figure 1). The criterion map provided a starting point for identifying potential linking phrases, some of which were later modified to be age-appropriate. The following linking phrases were provided 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

Any two of the nine concepts supplied can be connected with two possible unidirectional arrows, for example, the relationship from “density” to “matter” can be stated as “density is the property of matter” or “matter has the property of density.” The relationships described by the two propositions are quite similar; two-way arrows are not allowed. Therefore, the direction of the relationship was considered when the adequacy of the proposition was evaluated. Nevertheless, both “Matter → Density” and “Matter → Density” scores was treated as the same proposition for “Density – Matter.”

FIGURE 1

Criterion concept map. (Solid lines in the criterion map are propositions constructed by experts. Dash lines are propositions that originally were not in the expert map but used by more than 50% of students.)
Mathematically, all combinations of the nine terms produce 36 (9*8/2) concept pairs. However, not all the concept pairs are scientifically relevant. For example, “volume” has no scientifically relevant relationship with “gram.” Based on experts’ and students’ map, a criterion concept map was constructed with sixteen concept pairs with scientific relationships (Figure 1 and Table 1). Following the terminology proposed by Ruiz-Primo et al. (1996), those concept pairs with their corresponding relationships appearing in the criterion map were called “mandatory” propositions (Table 1). This study focused only on the mandatory propositions as a sample from the subject-matter universe.

Mandatory propositions were scored using a four-point scale: 0 for missing/wrong/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, an Excel database was created that contained all of the propositions constructed by each student. All the unique student-generated propositions extracted from the database comprised a “master list” of propositions. Three raters, two graduate students, and one professor, all in science education, reached agreement on the scores for all the unique propositions, and built up the master scoring list. After each student’s concept map propositions were transferred into the Excel database, the master scoring-list was used to score each proposition.

Facets

A concept-map’s proposition scores are a sample representative of a student’s declarative knowledge structure drawn from a universe defined by a combination of all possible propositions, mapping techniques (e.g., C and S), and occasions (1 and 2). Because students’ map scores were the consensus score of three raters in the automated scoring with Excel, rater was not regarded as a facet in the design. Student was the object of measurement. Proposition, occasion, and format (task types) were the measurement facets.

Proposition, a facet of the concept map assessment, is analogous to an item in a multiple-choice test. The linked concept pairs seem to be dependent, because linking one pair constrains other options. However, there is no limit on how many links students can establish among those terms; also, misconnected terms receive a score of 0 rather than negative points. Therefore, it seemed reasonable to assume
TABLE 1
Scientifically Relevant Concept Pairs

<table>
<thead>
<tr>
<th></th>
<th>Buoyancy</th>
<th>CC</th>
<th>Density</th>
<th>Volume</th>
<th>Mass</th>
<th>Matter</th>
<th>Gram</th>
<th>Water</th>
<th>Wood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Buoyancy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Density</td>
<td>Depends on</td>
<td></td>
<td>Measure of</td>
<td>Mass per</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mass</td>
<td></td>
<td></td>
<td>/ Volume</td>
<td>* Density</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Matter</td>
<td></td>
<td></td>
<td>Property</td>
<td>Property</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gram</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Measure of</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>Has</td>
<td>Has</td>
<td>Has</td>
<td>Form of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wood</td>
<td>Has</td>
<td>Has</td>
<td>Has</td>
<td>Form of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Even though we could also construct a scientific proposition between “volume” and “mass,” we did not include this proposition as a mandatory one because we did not supply the corresponding linking phrase in the S condition, which may have constrained the students in S to construct this proposition.
that propositions in this study were independent of each other. Propositions sampled in this study could be considered exchangeable with any other possible proposition in the topic, therefore, proposition was treated as a random facet.

Two concept-mapping techniques that varied in format, C and S, in four sequences were examined: from C to S, from S to C, from C to C, and from S to S. Accordingly, besides proposition, the second facet in sequences 1 and 2 is format, and the second facet in sequences 3 and 4 is occasion. Format is a fixed facet because the two task types were purposively selected and one cannot generalize the conclusions drawn about these types to other task types. For instance, these two formats differ greatly from a format that provides a concept map structure. However, according to Shavelson and Webb (1991), one should first “run an analysis of variance treating all sources of variance as random” in treating fixed facets (p. 67). If the fixed facet, here format, contributes negligible variation in scores, one can average over the fixed facet and conduct a G study with the remaining sources of variation. However, if conceptually and statistically, the fixed levels of the facet are distinct, a separate G study should be run for each level. Therefore, a fully random analysis in CS and SC was performed before further steps were taken. In CC and SS, occasion was treated as a random facet, which is exchangeable with any other occasion drawn from the universe. That is, theoretically the pattern found from CC and SS can be generalized to other repeated measures using C map and S map, respectively.

RESULTS AND DISCUSSION

G Studies

Pre liminary matters

Four different sequences were involved. A preliminary generalizability study showed a big difference in the variance components associated with CS and SC. The difference might be due to the confounding effects of format and occasion, a phenomenon about which Cronbach et al. (1997) warned, because CS and SC groups only differed in the order of constructing S and C maps. Consequently given the greater interest in comparing the two task formats than averaging their properties over occasions, CS and SC results with C and S were not treated as fixed effects. Rather, G studies (Student $\times$ Proposition $\times$ Occasion) for CC and SS are presented separately.

Comparison of C and S mapping techniques

Overall, variance component patterns in CC and SS were similar (Table 2): Student $\times$ Proposition $\times$ Occasion and error term accounted for the largest proportion
of variance, Student × Proposition the second largest, Student the third, and Proposition the fourth. The magnitude of each variation component, however, differed for CC and SS.

The Student × Proposition × Occasion interaction confounded with random error was the major source of measurement error, suggesting that a substantial proportion of the variability was due to facets not included in the study and/or random error. This variance component accounted for more of the total variability in C (62.4%) than in S (50.3%), which indicates that the factors contributing to student’s generating their own linking lines were not adequately captured or that the students with partial knowledge of buoyancy responded in inconsistent ways. The C technique’s openness and flexibility might have introduced more randomness from one occasion to the other than did the S technique’s constraints. This result is consistent with previous study (Ruiz-Primo, et al., 1996).

The Student × Proposition interaction was the second largest source of error, and it was larger in S (25.3%) than in C (16.2%). It suggests that students performed differently on different propositions, for example, some students did better on certain propositions than others whereas other students did better in an opposite way. This pattern was stronger in S than in C. This pattern fit the experience of the researchers in scoring students’ maps. It seemed that in S, due to the

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Estimate</th>
<th>Percentage of Total Variance</th>
<th>df</th>
<th>Estimate</th>
<th>Percentage of Total Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students (S)</td>
<td>25</td>
<td>.1756</td>
<td>10.4%</td>
<td>20</td>
<td>.3015</td>
<td>18.6%</td>
</tr>
<tr>
<td>Occasions (O)</td>
<td>1</td>
<td>.0038</td>
<td>0.2%</td>
<td>1</td>
<td>.0036</td>
<td>0.2%</td>
</tr>
<tr>
<td>Proposition (P)</td>
<td>15</td>
<td>.1547</td>
<td>9.1%</td>
<td>15</td>
<td>.0822</td>
<td>5.1%</td>
</tr>
<tr>
<td>SO</td>
<td>25</td>
<td>(0.0)</td>
<td>0.0%</td>
<td>20</td>
<td>.0080</td>
<td>0.5%</td>
</tr>
<tr>
<td>SP</td>
<td>375</td>
<td>.2740</td>
<td>16.2%</td>
<td>300</td>
<td>.4101</td>
<td>25.3%</td>
</tr>
<tr>
<td>OP</td>
<td>15</td>
<td>.0282</td>
<td>1.7%</td>
<td>15</td>
<td>(.0)</td>
<td>0.0%</td>
</tr>
<tr>
<td>SOP, e</td>
<td>375</td>
<td>1.0581</td>
<td>62.4%</td>
<td>300</td>
<td>.8146</td>
<td>50.3%</td>
</tr>
</tbody>
</table>

\[ \sigma_0^2 \]

1.3321

\[ \sigma_A^2 \]

1.5188

Relative G Coefficient

\[ \hat{\rho}^2 \]

.1164

Absolute G Coefficient \( \phi \)

.1036
constraints of the linking phrases, students either got a proposition perfectly correct (score = 3) by connecting the right linking words with concept pairs or completely missed it (score = 0) by selecting the wrong linking words (Yin et al., 2005).

Proposition sampling created a larger portion of variability in C (9.1%) than in S (5.1%). This finding is not surprising, considering that the linking phrases in the S condition were taken from the criterion map and they were also the components of mandatory propositions. Some relationships between concepts might be more difficult than others for students if no hints (linking phrases) were available. For example, very few students in C constructed the relationship between “density and volume” and “density and mass.” But this was not the case in the S condition, where students obtained hints from the linking phrases: “is mass divided by . . . ,” and “divided by volume equals . . .” (Yin et al., 2005). Consequently, students in S were more likely than students in C to construct proper relationships for those concept pairs as a function of task and not necessarily of their knowledge structures.

Finally, variance created by students was larger in S (18.6%) than in C (10.4%). That was consistent with the larger G coefficients in S than C if one proposition and one occasion were involved: G coefficients for relative decision—S (0.1966) versus C (0.1164); and G coefficient for absolute decision—S (0.1861) versus C (0.1036).

The difference in variance components for SS and CC might be due to three reasons: (a) The two groups of students who took the SS and CC might not be equivalent. In particular, the group taking SS might be more diverse in their knowledge of buoyancy than the group taking CC. Although all four groups were formed randomly and their teacher’s overall ratings did not indicate group difference, imbalance among groups on this particular topic may still exist due to the small sample sizes. (b) The S mapping task might differentiate students’ performance better than C. That is, students’ performance on S varied more than students’ performance on C. (c) Artifacts of the S method produced more consistent measure of students’ performance than did C but a less valid measure.

Due to the lack of information about students’ specific knowledge of density and buoyancy, reason (a) could not be investigated in the current study. To examine reason (b) the student score variance on C maps and S maps on occasion 1 were calculated, when no testing effect was involved. When both CC and CS groups are included as C and SC and SS as group S, the variance of student scores on C was 56.04 and on S was 76.09 ($F = 0.74$, $p = .15$). When only CC and SS groups were included, the variance of student scores on C was 44.90 and on S was 89.43 ($F = 1.50$, $p = .052$). That is, students’ performance varied more greatly on S than C. When only CC and SS groups were included, the pattern was stronger, but not statistically significant. Finally, reason (c) cannot be tested directly, but studies by Ruiz-Primo et al. (1996) and Yin et al. (2005) suggested
that the S technique scaffolds and influences the knowledge structure to a greater extent than does the C technique, raising questions of validity of the S technique.

In this case, none of the possible explanations can be eliminated. The first two may contribute to the greater student variance in SS than CC to some degree while the third contributed to greater residual variance in the C condition than the S condition. That is, students in the SS and CC groups may not be equivalent in their ability; meanwhile, the S map may differentiate students’ ability level better than the C. However, this differentiation may be an artifact of the method itself. The phrases provided in S may help some students to get the most points possible for a proposition (once they correctly determine several phrases). The students can do so either by hinting at the relationship or by our scoring method and/or make other students lose points (once they misuse several phrases). In contrast, the openness of C leaves more flexibility for students to get partial credit overall and meanwhile lose points on some difficult propositions.

Overall, the G study suggested that C and S conditions were similar but not equivalent: The patterns of variance components were similar, but the magnitude of error was greater in C than in S. Then how would the overall G coefficients change when the numbers of propositions and occasions vary? Do S and C measure the same thing with equivalent reliability? When financial and time costs become important concerns in practice, it is necessary to find a less “costly” technique and the proper combination of propositions and occasions, but not at the cost of interpretability. Let us now turn to the question of assessment design and improvement in the reliability of both measurement techniques.

D Studies

Based on information from the G studies, the effects of increasing the numbers of propositions and occasions in a series of D studies were examined. Because the D study showed similar patterns in relative and absolute G coefficients, for simplicity, only D studies with relative G coefficients are discussed in this article.

The D studies indicated that more occasions or items are needed in C than S to reach the same generalizability coefficient. To reach generalizability of approximately .80 in evaluating students’ relative rank, if only one occasion is applied, about 18 propositions would be needed in S. The same number of propositions and one occasion would only lead to relative generalizability coefficient of .70 in C. To reach .80 generalizability coefficient in C, either the combination of two occasions and 18 propositions would be needed or 30 propositions would be needed if only one occasion could be afforded.

Based on the data obtained from this study, S has higher reliability than C under the same conditions. Considering that C is more valid than S in representing students’ knowledge structure and diagnosing students’ misconceptions, we would still prefer C mapping to S mapping. The tradeoff is that more propositions
are needed in C than S to obtain equivalent reliability, when data are collected on one occasion only.

CONCLUSIONS AND IMPLICATIONS

This article presents the application of G theory in the technical evaluation of concept map assessment scores with the hope of widening the theory’s use in concept map assessment research. Due to G theory’s power, convenience, and flexibility, it is preferable to CTT in examining measurement error and reliability for concept map assessments, which involve great number of potential sources of measurement error.

The following are some sources of variation characteristic of concept map assessments: (a) Concept/Term—Different concept terms may be chosen from the same domain, analogous to parallel form reliability in CTT; (b) Proposition—Each proposition in a concept map can be regarded as an independent item in a test, sampled from some domain and different propositions may vary in difficulty level—internal consistency in CTT; (c) Occasion—When a concept map assessment is administered at different times—test-retest reliability in CTT; (d) Concept-Map Task—A concept map task may vary from supplying nothing but a topic to asking examinees to filling in missing concepts or linking phrases; (e) Response-Format—Concept map assessments can be administered as a paper-and-pencil test or a computer interactive test; (f) Rater—When raters are involved and introduce unintended variation into the measurement—inter-rater reliability in CTT; (g) Scoring-system—Different scoring systems can be used to evaluate propositions, structures, or the whole map: semantic content score (Herl et al., 1999), total accuracy score, convergence score, salience score (Ruiz-Primo, et al., 1996), link score (Astin & Shore, 1995), structure score (Kinchin, 2000), holistic score, relational scoring (McClure et al., 1999) can all be used to judge students’ performance in concept map assessments and infer students’ declarative knowledge structure quality.

Besides the variation created by each facet above, the interactions among the individual facets, and with the object of measurement, usually students, can lead to variation in students’ concept-map scores. What matters for assessment purpose is that different inferences may be drawn about a students’ declarative knowledge structure based on variation in all these facets.

The flexibility of G theory allows researchers to examine individual facets or the combination of them to answer practical questions. Many technical properties of concept map assessments might be examined, for example, test-retest stability, inter-rater reliability, internal consistency, equivalence of different concept map tasks/response formats/scoring systems. In G theory, one can simultaneously estimate the impact of each of these sources of score variation individually and in
interaction. With information obtained from the G study, one can answer questions related to the design of concept mapping techniques for a particular application, say in large-scale assessment in a D study. For example, how many propositions are needed to obtain a dependable score for a student? How many raters are needed to score a concept map reliably? What scoring system has better technical properties, say more likely to provide reliable measures? Answers to these (and other) questions can help to narrow the large number of options involved in concept map assessments used in practice.

In addition to describing the application of G theory to concept maps, the application of G theory is illustrated by examining two concept-mapping techniques: construct-a-map with created linking phrases (C) and construct-a-map with selected linking phrases (S). C and S were similar but not equivalent in their measurement errors and reliability. The variance component analysis showed (a) Student × Proposition × Occasion interaction confounded with error accounted for the largest proportion of variation with either technique, larger in C than in S; (b) Student × Proposition interaction accounted for a larger proportion of variability in S than in C; and (c) the G coefficients for S for one proposition and one occasion were larger than those for C, but by adding a few concept terms producing more propositions, reliability (generalizability) of C could exceed .80.

Based on the current study, S is a more efficient and reliable mapping technique than C. However, before concluding that S might be a better assessment candidate than C, this analysis and other research raised concerns about the validity of the S technique. The S technique’s superior performance, in terms of reliability, was due in part to its scaffolding and constraining student responses. Other research shows that the S technique may overestimate the knowledge structures for some students because of these artifacts. The C technique, based on “think aloud” protocols, captures a much greater cognitive search of knowledge structure than does the S technique (Yin, et al., 2005).

That is, G theory mainly provides reliability information about these assessments. A reliable assessment may not be a valid one. To investigate the validity of different assessments, one could employ other qualitative methods, such as think aloud, to triangulate or add to the quantitative information. Previous research has shown that C better reflected students’ partial knowledge and misunderstandings (Ruiz-Primo, Schultz et al., 2001; Ruiz-Primo, Shavelson et al., 2001; Yin et al., 2005). This study also confirmed previous findings. For example, a larger proportion of variability in C than S is due to variation in proposition difficulty. This pattern indicates that C is more powerful than S in diagnosing the propositions with which students tended to have problems. Therefore, C might be an especially effective assessment tool in classroom settings, where fully understanding a student’s current thinking is important; while larger number of propositions must be included in C than S when they are used in high-stake tests.
This study was a small demonstration of how G theory could contribute to the study of concept map assessment. The study has limitations: (a) Students sampled were mainly from upper middle-class homes. (b) Concept terms sampled were only related to one science topic. (c) Occasion was confounded with format in the CS and SC G studies, and the problem could not be addressed in the current design and analysis. (d) Sample size is rather small. (e) Most importantly, the small sample size might lead to nonequivalence among different groups, which may create potential alternative explanations for some of the results. In this study, although students’ general science performance was found equivalent among the four groups according to their teacher’s evaluation, the evaluation method was rather rough—with three levels only and not specific for the topic of buoyancy. Further similar studies should use a more precise instrument than teacher rating to better capture the variance of student performance and a larger sample size or a stratified random sampling to ensure the equivalence of different groups.

This study just scratches the surface of the many psychometric unknowns in this field that could be investigated with G theory. With more technical information about concept map assessments, assessment researchers would be in a better position to utilize concept mapping to help understand what students know and finally improve students’ learning. This study is hoped to help expand the application of G theory in concept map assessment research.

ACKNOWLEDGMENTS

The authors thank Carlos C. Ayala and Jim Vanides for their assistance in data collection and scoring. The authors also owe thanks for valuable comments and suggestions to Maria Araceli Ruiz-Primo, Edward Heartel, Andrew Dean Ho, and the editor and anonymous referees of Applied Measurement in Education. All views expressed and any remaining errors are the authors’ alone.

REFERENCES

Workshop on Instructional Design Models for Computer-Based Learning Environments, The Netherlands.


