Application of Generalizability Theory to Concept-Map Assessment Research

CSE Report 640

Yue Yin and Richard J. Shavelson
Stanford University

November 2004

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

Project 2.2: Classroom and Teachers' Assessment. Strand 2: Curriculum Embedded Assessments
Richard Shavelson, Project Director, StanfordUniversity.

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

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

APPLICATION OF GENERALIZABILITY THEORY TO CONCEPT-MAP ASSESSMENT RESEARCH

Abstract

In the first part of this paper we discuss the feasibility of using Generalizability (G) Theory (see Footnote 1) to examine the dependability of concept map assessments and to design a concept map assessment for a particular practical application. In the second part, we apply G theory to compare 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). We explore some measurement facets that influence concept-map scores and estimate how to optimize different concept mapping techniques by varying the conditions for different facets. We found that C and S were not technically equivalent. The G coefficients for S were larger than those for C. Furthermore, a D study showed 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. Therefore, S might be a better candidate than C in the large-scale summative assessment, while C would be preferred as a formative assessment in classroom.

Assessment of learning is typically narrowly defined as multiple-choice and short-answer tests; achievement is typically what multiple-choice tests measure. There is more to achievement than this, and a definition of achievement might well consider the structure of a student's knowledge, not just the quantity. To this end, concept maps provide one possible approach. Once this approach is taken, often the technical quality of the concept-map assessment is assumed; it is not clear just how to evaluate reliability, for example. In this paper, we present a Generalizability Theory framework (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) for examining the dependability of concept-map assessments and demonstrate its application.

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* 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 non-hierarchical).

Concept maps were originally proposed to be used as an instructional tool (e.g., Novak & Gowin, 1984) and later as an assessment as well (Ruiz-Primo & Shavelson, 1996). Concept maps hold promise in tapping students' declarative knowledge structures which traditional assessments are not good at. 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 dimensions: 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 (Appendix 1). To get a comprehensive review of the variations, readers can refer to the paper written by Ruiz-Primo and Shavelson (1996).

Even though thousands of concept-map assessment permutations are possible, not all alternatives are suited for assessment (Ruiz-Primo & Shavelson, 1996). Ruiz-Primo and Shavelson pointed out that reliability and validity information about different mapping techniques should be supplied before concept maps are used for assessment. Our study is one such effort. In particular, in the first part of this paper, we discuss the feasibility of using G theory to evaluate the dependability of concept map scores. In the second part of this paper, we illustrate how G theory can be applied in this kind of research 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).

Part 1. Application of G theory to Concept-Map Assessment

Issues and Problems Related to the Technical Properties of Concept-map Assessments

Concept maps vary greatly from one another both for 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 et al. (Ruiz-Primo, Shavelson, & Schultz, March 1997, p. 7) 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". Even though criterion (c) only talked about the scoring system, we (Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, In Press) found that the accuracy of the scores is not only related to the scoring systems, but also related to the task format. For example, using the same scoring form, some task formats might be scored more reliably and accurately than others (Yin et al., In Press).

This paper, then, mainly focuses on criteria (b) and (c), which have typically been gauged by traditional statistical analyses and classical test theory. For example, mainly using those methods, researchers examined scores for inter-rater reliability/agreement (Herl, O'Neil, Chung, & Schacter, 1999; Lay-Dopyera & Beyerbach, 1983; Lomask, Baron, Greig, & Harrison, March 1992; McClure, Sonak, & Suen, 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, July 1991; Markham, Mintzes, & Jones, 1994; Novak, Gowin, & Johansen, 1983; Rice, Ryan, & Samson, 1998; Schreiber & Abegg, 1991); predictive validity (Acton, Johnson, & Golldsmith, 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., In Press).

Those studies have supplied important information about the technical properties of different concept map tasks, response formats, and scoring systems, which can undoubtedly help to eliminate improper alternatives. However, because the variations among concept map assessments are so great that classical test theory cannot handle those variations simultaneously and efficiently.

Examining Concept-Map Assessments' Technical Properties with G Theory

If we view a concept map assessment score as a sample from a universe of conditions with all kinds of variations—for example, tasks, response formats, and scoring systems—we can examine concept map assessments in the framework of G theory.

Strength of G theory. Compared with classical test theory, G theory can (a) integrate conceptually and simultaneously evaluate test-retest reliability,

internal-consistency, convergent validity, and inter-rater reliability; (b) estimate not only the influence of individual measurement facets, but also interaction effects; (c) permit us to optimize an assessment's dependability ("reliability") within given dollar and time cost constraints. For example, the 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; (4) as a general advantage in assessing students' performance, G study can supply dependability information on students' absolute level of knowledge structure quality as well as its relative level.

Object of measurement. Typically, education research using concept maps focuses on the variation in the quality and complexity of students' declarative knowledge structures in a certain subject. This is the variability that the concept-map assessments intend to measure—the object of measurement.

Variation among students, then, is a desirable variation and should not be confused with variation caused by other sources. There are many other sources of variation in concept map scores that contribute error to the measurement. They include individual factors and the interaction between/among more than one factor. The individual factors leading to measurement error are called facets in G theory.

Facets. The following are some possible facet examples of individual factors that are characteristic of concept maps.

(1) Concept/term Sampling: Concept maps sample concepts from some domain, for example, key concepts. Concept sampling may give rise to variability in a student's performance; performance might differ with another sample of concepts. Concept sampling, then, introduces error to the measurement when we want to generalize a student's performance from one map with certain concept sample to a universe of maps with concept-terms sampled.

(2) Proposition Sampling: Each proposition in a concept map can be regarded as an independent item in a test, sampled from some domain. Different propositions vary in difficulty level. Proposition sampling, then, can cause variation in the measures of the proficiency of students' declarative knowledge structures. Notice that facet (2) is similar to facet (1) in that they are both related to the variation due to concept sampling; however, facet (1) focuses on sampling

at a macro level, analogous to alternate form reliability in the classic test theory, while facet (2), analogous to internal consistency, focuses on sampling at a micro level. The two facets' similarity and difference again show the strength of G theory in that it allows researchers to flexibly focus on the interested error type in the analysis to meet specific needs in one single analysis.

(3) Concept-Map Task Sampling: Concept-map tasks vary greatly. A concept map task may supply nothing but a topic and ask students to construct the map from scratch; it may supply concepts only and require students to construct the map with the concepts supplied; it may supply both concepts and linking phrases and ask students to construct the map by assembling the concepts and linking phrases supplied; it may supply a partially complete map and ask the students to fill in the nodes (concepts) or fill in the lines (relationships). Also, a concept map task may set a certain structure for the map—for example, hierarchical or linear—or it may leave the students to decide how to structure the map. The list of variations in concept-map tasks can go on endlessly. 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, a task can be regarded as a facet captured by G theory.

(4) Response-format Sampling: Concept map assessments can be administered as a paper-and-pencil test or a computer interactive test; students may perform better in one response format than the other. Previous research on other kinds of assessments—for example, performance assessment—has shown the existence of variation brought by response formats (Shavelson, Baxter, & Gao, 1993).

(5) Occasion Sampling: When we repeat a concept-map assessment on a group of students, as in a test-retest design, we sample occasions. Students may perform inconsistently when taking concept map assessments on different occasions (Cronbach, Linn, Brennan, & Haertel, 1997; Shavelson, Ruiz-Primo, & Wiley, 1999). Classical test theory 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.

(6) Rater Sampling: Unless a computer automatically scores a concept map, raters are involved and introduce unintended variation into the measurement.

Some raters might be more lenient than others or raters may interpret an examinee's map differently due to their personal background, experience, or just perspective. Especially, when a concept-map task exerts few constraints—for example, only the concept terms are supplied—it leaves great freedom/flexibility to the students in map construction. Consequently, it is almost impossible to establish a rubric exhausting all the possibilities that students may construct in an open-ended map. In this case, open-ended maps leave raters even greater possibilities of interpreting students' concept maps differently than other map types. All in all, raters may introduce unwanted variation into students' scores.

(7) Scoring-system Sampling: Concept maps have been scored in a wide variety of ways to measure a student's knowledge structure. For example, different scoring systems can be used to evaluate propositions, structures, or the whole map: semantic content score (Herl et al., 1999), individual proposition accuracy score (Yin et al., In Press), total accuracy score, convergence score, salience score (Ruiz-Primo, Schultz, & Shavelson, 1996, April), link score (Astin & Shore, 1995), structure score (Kinchin, 2000) holistic score, and relational scoring (McClure et al., 1999) can all be used to judge students' performance in concept map assessment and infer students' declarative knowledge structure quality. What matters is that we may draw different inferences about a students' declarative knowledge structure based on variation in scoring systems. In addition, some scoring systems are easier to use than others. Therefore, the scoring system is a very important facet often masked by the use of a single system in a particular study and thereby confounded with task sampling.

Besides the variation created by each facet above, the interactions among the individual facets, and with the object of measurement, can lead to variation in students' concept-map scores. The following are some examples of interactions.

(1) Rater _ Concept-map task: Raters may score students' performance on different concept map tasks more or less stringently. For example, Rater A may score concept map task I more strictly than he/she scores concept map task II, because the rater may unconsciously believe task I is easier to do than task II.

(2) Rater _ Student: Rater A might be more lenient to students whose handwriting is clear, while Rater B may treat all the students the same.

(3) Scoring system _ Concept-map task: When used to score concept map task I, Scoring System A may lead to a higher estimation of students' proficiency level than Scoring System B; in contrast, Scoring System A and Scoring System B may reach the same judgment on students' proficiency when used to score Task II.

(4) Concept map task _ Student: Concept map task I might be easier for some students than concept map task II, but the two concept map tasks might be equally difficult for other students.

Following a similar logic, we can continue to extend this interaction list, because interactions can happen between/among more than one facet and the object of measurement. All of the interactions have the possibility of supplying important information about concept map assessments' technical properties. Furthermore, based on the information conveyed from G study, D study can help optimize the concept map task, response form, and scoring system for a specific purpose.

The flexibility of G theory allows researchers to "fix" certain facet(s) and examine concept-map assessments' dependability on other facets. For instance, when we try to compare two concept map tasks' dependability over occasions and propositions, we can the fix concept-map type—i.e., examine the impact of these facets for each concept-map task (based on the same scoring system). In contrast, when we are interested in examining a concept-map task's dependability on different scoring systems and occasions, we treat the scoring system and occasion as random facets in the analysis along with other facets not varied.

Table 1 presents some of the reliability questions that can be addressed by G theory and Table 2 presents some of the practicality questions that can be answered by D study in examining concept map tasks and scoring systems.

Table 1.
Some Questions about the Reliability of Concept Map Assessment that can be Answered by G Study.

Technical Properties	Task Type	Scoring System
Stability over time	Which concept map tasks can be scored consistently over time?	Which scoring system can be used to score a concept map consistently over time?
Inter-rater reliability	Which concept map tasks can be scored consistently over raters?	Which scoring system can be used to score a concept map consistently over raters?
Internal consistency	Which concept map samples and tasks can be scored consistently over propositions?	Which scoring system can be used to score a concept map consistently over propositions?
Equivalence of alternate forms	Can concept map tasks with different concept samples reveal similar information about students' knowledge in a domain?	Can concept map scoring system reveal the similar information about students' knowledge in that domain when different concept samples are used?

Table 2.
Some Questions about the Practicality of Concept Map Assessments that Can be Answered by D Study.

Technical Properties	Task Type	Scoring System
Stability over time	How many times should a certain concept map task be administered in order to obtain a reliable result?	How many times should a certain concept map be administered in order to obtain a reliable result under a specific scoring system?
Inter-rater reliability	How many raters should score the concept maps in order to obtain a reliable result?	How many raters should score the concept maps in order to obtain a reliable result under a specific scoring system?
Internal consistency	How many propositions should be included in the concept maps in order to obtain a reliable result?	How many propositions should be included in the concept maps in order to obtain a reliable result under a specific scoring system?
Equivalence of alternate forms	How many sample lists should be administered as the concept maps in order to obtain a reliable result?	How many sample lists should be administered as the concept maps in order to obtain a reliable result under a specific scoring system?

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 Concept Sample A, and a concept map task with Concept Sample B. In their analysis, two facets were implemented—rater and condition (concept map task/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 link over all criterion links; and salience (proportion of valid student link over all student links). They found that: using the three scoring systems, (a) Raters introduced negligible error variability into the measurement; (b) Students' relative standing varied on different conditions; (c) Both the relative and absolute G coefficients were quite high. 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 G coefficients.

In the following section, we use one of our recent studies to illustrate the application of G theory to concept-map assessment research.

Part 2. Comparison of Two Concept Map Techniques by G Theory

We use two concrete concept-mapping tasks as exemplars: construct-a-map with created linking phrases (C) and construct-a-map with selected linking phrases (S). In C, students are given concept terms and asked to create a map; in S, students are given both linking phrases and concept terms to construct a map.

We chose C and S concept-mapping tasks because they 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 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, 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 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 we consider the use of concept maps as a potential large-scale assessment tool.

Given the openness of the C mapping technique and the constraints of the S, we raised the following questions: Do the two techniques vary in technical characteristics? For example, do they vary in stability and internal consistency? What can be done if certain reliability levels are needed? For different techniques, does the way in which they are optimized vary? How can quality and efficiency be balanced in the concept map assessment design? We address these questions in the framework of G theory. We explore factors influencing concept-map scores' generalizability in G study and estimate how to optimize different concept mapping techniques by varying the conditions for different measurement facets in D study.

Method

Participants

Ninety-two eighth-graders from the California Bay Area participated in the study; 46 were girls and 46 boys. The students were drawn largely from upper middle class homes and 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 we used a 4 ¥ 2 (mapping sequence ¥ occasion) design. Students were randomly assigned to one of four

mapping sequences across the two occasions: (a) CS—construct-a-map with created linking phrases then 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 then construct-a-map again with created linking phrases (n = 26); or (d) SS—construct-a-map with selected linking phrases and then with selected linking phrases again (n = 21). The elapsed time between occasions was 7 weeks, with no instructional intervention related to the content assessed—in this case mass, volume, density, and buoyancy—during that time.

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 an assessment design team working with 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). By using an iterative selection process involving ranking and voting by the team members, we selected 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

In our study, 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 vice versa, "matter has the property of density". The relationships described by the two propositions are quite similar; two-way arrows are not allowed. Therefore, we considered the direction of the relationship when evaluating the adequacy of the proposition but treated both "Matter ⇵ Density" and "Matter ⇓ Density" scores as the same proposition for "Density - Matter".

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' maps, we constructed a criterion concept map (Figure 1) and identified sixteen concept pairs with scientific relationships. We labeled those concept pairs with their corresponding relationships as "mandatory" propositions (Table 3). They are the propositions in the criterion map (Figure 1). 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 frequently by students. We only scored the sixteen propositions and viewed them as a sample from the subject-matter universe.

Table 3.
Scientifically Relevant Concept Pairs

Buoyancy									
CC									
Density	Depends on								
Volume		Measure of	Mass per						
Mass			/ Volume	* Density					
Matter			Property	Property	Property				
Gram					Measure of				
Water			*Has*	*Has*	*Has*	Form of			
Wood			*Has*	*Has*	*Has*	Form of			
	Buoyancy	**CC**	**Density**	**Volume**	**Mass**	**Matter**	**Gram**	**Water**	**Wood**

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.

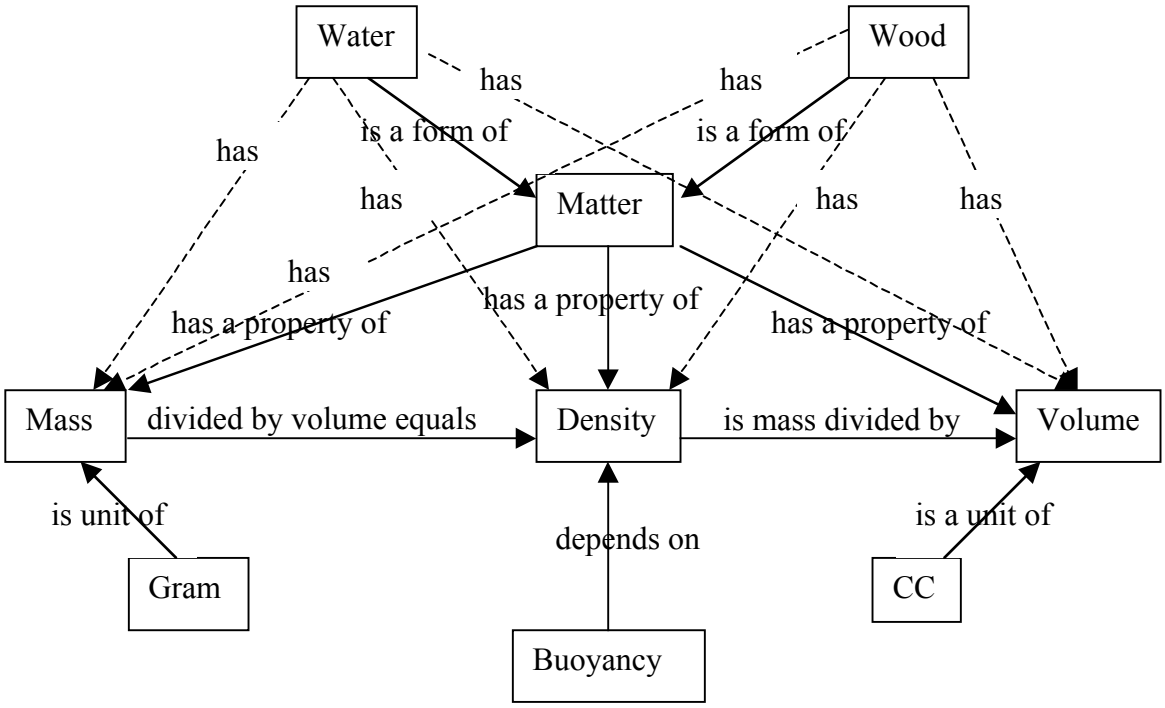

Figure 1. Criterion concept map.

Since a proposition is relatively easy to score and is interpreted as revealing depth of understanding (McClure et al., 1999), we scored propositions in our study. We scored the mandatory propositions using a four-point scale—0 for wrong/scientifically irrelevant propositions or if the mandatory proposition was not constructed, 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 containing all of the propositions submitted 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 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.

Facets

We view a concept-map's proposition scores as a sample representative of a student's declarative knowledge structure drawn from a universe defined by a combination of all possible propositions, test formats (e.g., C and S), and occasions (1 and 2). Since students' map scores were the consensus score of two raters, scoring was automated with the aid of the Excel program and rater is not regarded as a facet in our design. Persons are the object of measurement and propositions, occasions and formats are the facets of the measurement.

Proposition, then, is a facet of the concept-map assessment. We think of a proposition in a concept map as analogous to an item in a multiple-choice test. Propositions sampled in our study could be considered exchangeable with any other possible proposition in the topic, therefore, we treated proposition as a random facet. In order to differentiate P (proposition) from P (person), we use I (item) to represent proposition in the following discussion.

We examined two concept-mapping techniques that varied in format, C and S, in four sequences: 1) from C to S; 2) from S to C; 3) from C to C; 4) from S to S. Accordingly, besides proposition, the second facet in sequences 1 and 2 of our study is format and the second facet in sequences 3 and 4 is occasion. Format in our study is a fixed facet because the two task types were purposively selected and we cannot generalize the conclusion drawn about the two task types to other task types. However, according to Shavelson and Webb, we could first "run an analysis of variance treating all sources of variance as random" before further analysis (Shavelson & Webb, 1991, p. 67). Therefore, we ran a fully random analysis in CS and SC before further steps were taken. In CC and SS we treated occasion as a random facet, which is exchangeable with any other occasion drawn from the universe.

Results and Discussion

G study

Four different sequences are involved in our study. We first studied CS and SC, treating the format facet as random, following Shavelson and Webb (1991), to determine whether format created variation in scores. In both CS and SC, even though the format itself did not create much variability in both sequences, it created substantial variance by interacting with P, I, or both (see Table 4 and Figure 2).

Table 4.
Variance Component Estimate for the Person _ Proposition _ Format

	CS			SC		
Source	df	Estimate Variance Component	Percentage of Total Variance	df	Estimate Variance Component	Percentage of Total Variance
Persons (P)	21	.1105	5.9%	22	.1415	8.1%
Formats (F)	1	0	0.0%	1	.0236	1.4%
Items (I)	15	.0327	1.7%	15	.1638	9.4%
PF	21	.0410	2.2%	22	.0595	3.4%
PI	315	.1142	6.1%	330	.2563	14.7%
FI	15	.3931	21.0%	15	.1537	8.8%
PFI, *e*	315	1.1822	63.1%	330	.9479	54.3%
σ_{δ}^{2}		1.3374			1.2637	
σ_{Δ}^{2}		1.7632			1.6049	
$\hat{\rho}^{2}$		.0763			.11194	
f		.0590			.08814	

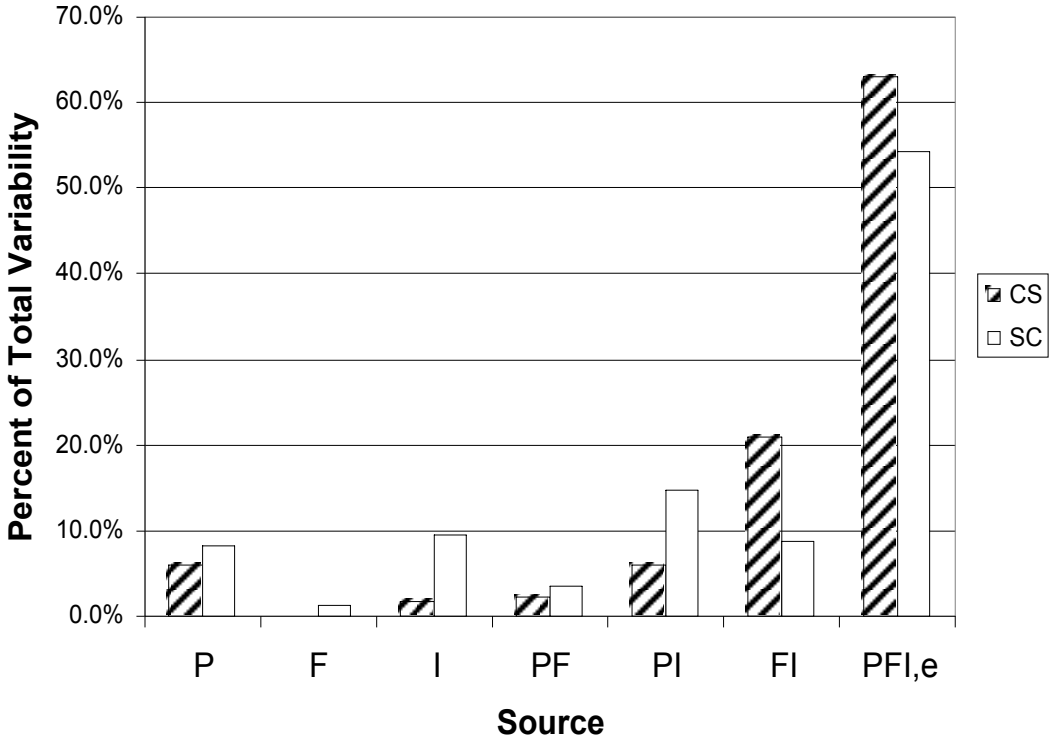

Figure 2. Comparison of source of variability in CS and SC.

However, the interaction between the format and other facets accounted for different proportions of the total variability in CS and SC scores. For example, format _ person interaction accounted for more of the total variability in CS (21.0%) than in SC (8.8%). With the similar trend, the person _ proposition/item interaction and error also accounted for more of the total variability in CS (63.1%) than in SC (54.3%). Differently, item and the item _ person interaction contributed larger variation in SC than in CS.

Considering that CS and SC groups only differed in the order of students constructing S and C maps, we suspected that in the SC and SC group, the format effect was confounded with the occasion effect and led to variance component differences in the two groups. Also, as mentioned earlier, format in our study should be treated as a fixed facet. Since we are more interested in comparing the two task formats than averaging their properties, in our later discussion, we studied SS and CC separately, examining each format in person _ proposition _ occasion G studies.

We focus on G study results for CC and SS. Table 5 and Figure 3 present variance component comparisons. The greatest difference between the variance component patterns involves a small OI in CC/SS and a correspondingly large FI in CS/SC. The difficulty level of propositions was more consistent across the two occasions with the same task type—either CC or SS— than across the two occasions with different task types—either SC or CS. In SC and CS, the effect of format was confounded with the effect of occasion. We interpret the difference in OI and FI as indicating that different task formats, and not occasions, contributed to the difference in magnitudes in the previous CS and SC analysis.

Table 5.
Variance Component Estimate for the Person _ Proposition _ Occasion

Source	CC			SS		
	df	Estimate Variance Component	Percentage of Total Variance	df	Estimate Variance Component	Percentage of Total Variance
Persons (P)	25	.1756	10.4%	20	.3015	18.6%
Occasions (O)	1	.0038	0.2%	1	.0036	0.2%
Items (I)	15	.1547	9.1%	15	.0822	5.1%
PO	25	(0.0)	0.0%	20	.0080	0.5%
PI	375	.27399	16.2%	300	.4101	25.3%
OI	15	.02821	1.7%	15	(.0)	0.0%
POI, *e*	375	1.0581	62.4%	300	.8146	50.3%
σ_{δ}^{2}		1.3321			1.2327	
σ_{Δ}^{2}		1.5188			1.3185	
$\hat{\rho}^{2}$		.1164			.1966	
f		.1036			.1861	

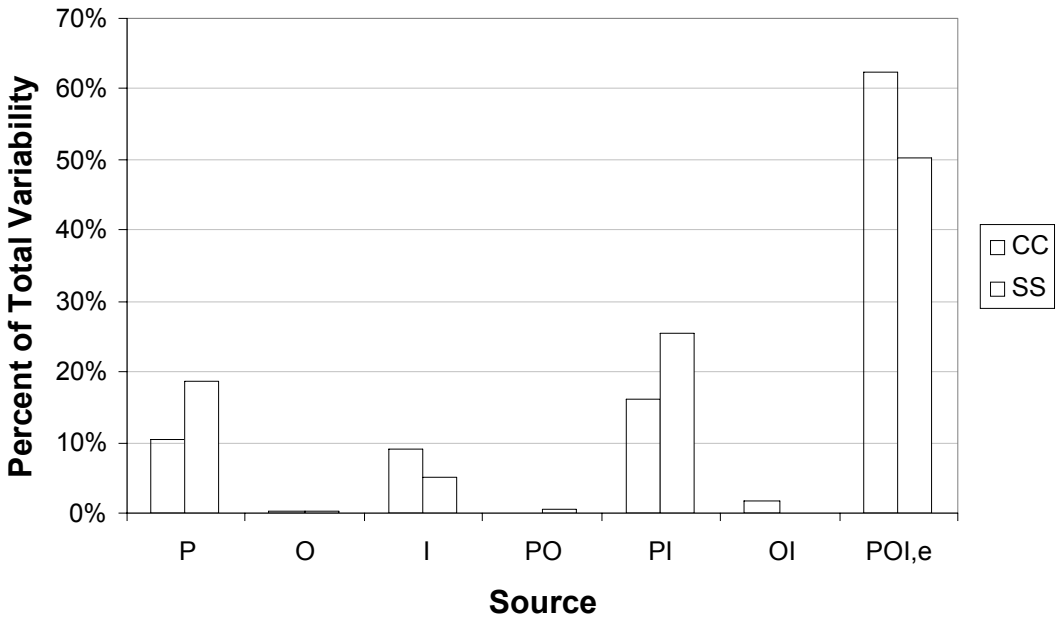

Figure 3. Comparison of source of variability in CC and SS.

The Person _ Proposition _ Occasion interaction confounded with random error was still the major source of measurement error, suggesting that a substantial proportion of the variability was due to the three-way interaction and/or unmeasured systematic/random error. This variance component accounted for more of the total variability in C (62.4%) than in S (50.3%), indicating that we could have a better measure of the variability in S than in C.

The Person _ Item 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 items—e.g., some students did better on certain items 1-8 than items 9-16, while other students did better on items 9-16 than item 1-8. This pattern was stronger in S than in C. This pattern fit our experience in scoring students' maps. We noticed that in S, due to the constraints of linking phrases, students either got a proposition perfectly correct (score=3) by connecting the right linking words with concept pairs or completely missed it by connecting the wrong linking words with concept pairs (score=0) (Yin et al., In Press).

Item created a larger portion of variability in C (9.1%) than in S (5.1%). It 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 of concepts might be more difficult than others if no hints (linking phrases) were available. For example, we noticed that 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…" Consequently, students in S were more likely than students in C to construct proper relationships for those concept pairs.

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

Overall, the G study suggested that C and S conditions were not equivalent—the patterns of variance components were similar, but the magnitude of error was greater in C than in S.

Clearly, the variance components for persons, the object of measurement, were rather low in both S and C compared with the measurement errors. Then how would the overall G coefficients change when the numbers of items and occasions vary? Do S and C have the similar requirements? When financial and time costs become important concerns in practice, it is necessary to find the less "costly" technique and the proper combination of items and occasions.

D study

Based on the information from the G study, we examined the effects of increasing the numbers of proposition and occasion in a series of D studies. Since our D study showed similar patterns in relative and absolute G coefficients, for simplicity, here we only discuss the D study with relative G coefficients.

Figure 4 and Figure 5 show the relative G coefficient change with varying numbers of proposition and occasion for C and S. To reach the generalizability of approximately 0.80 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 occasion could only lead to a relative generalizability coefficient of 0.70 in C. To reach 0.80 generalizability 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.

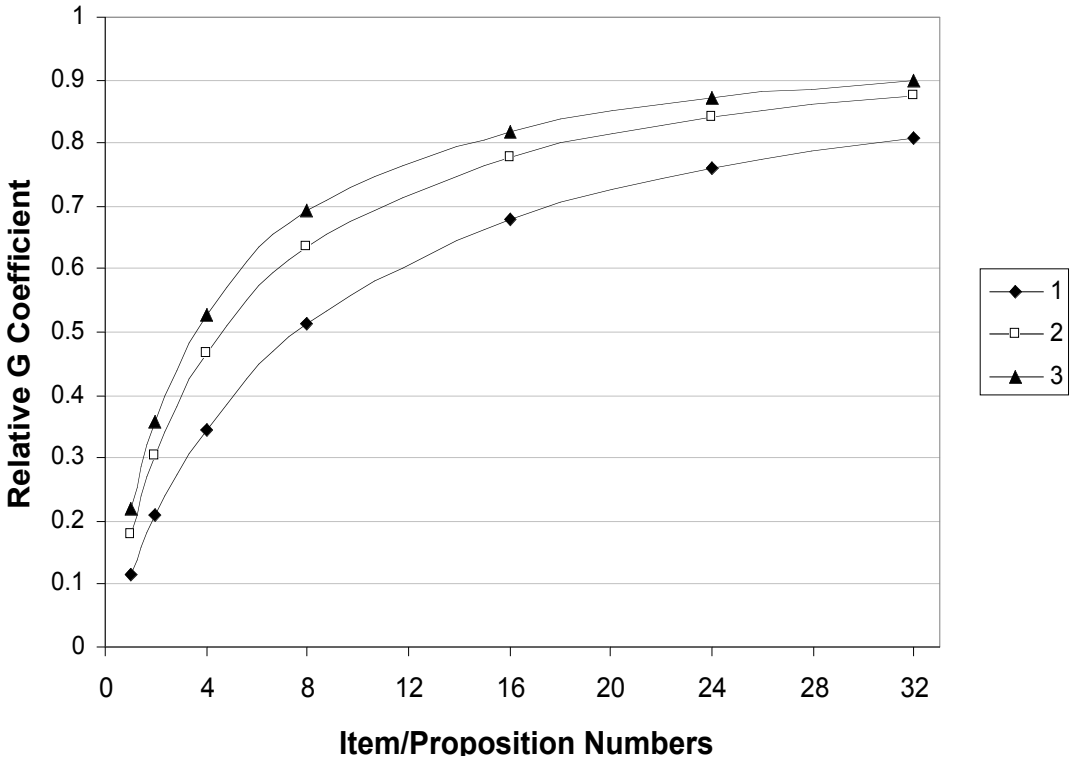

Figure 4. Trade-offs between numbers of mandatory items and occasions needed to achieve various levels of relative Generalizability in C mapping techniques.

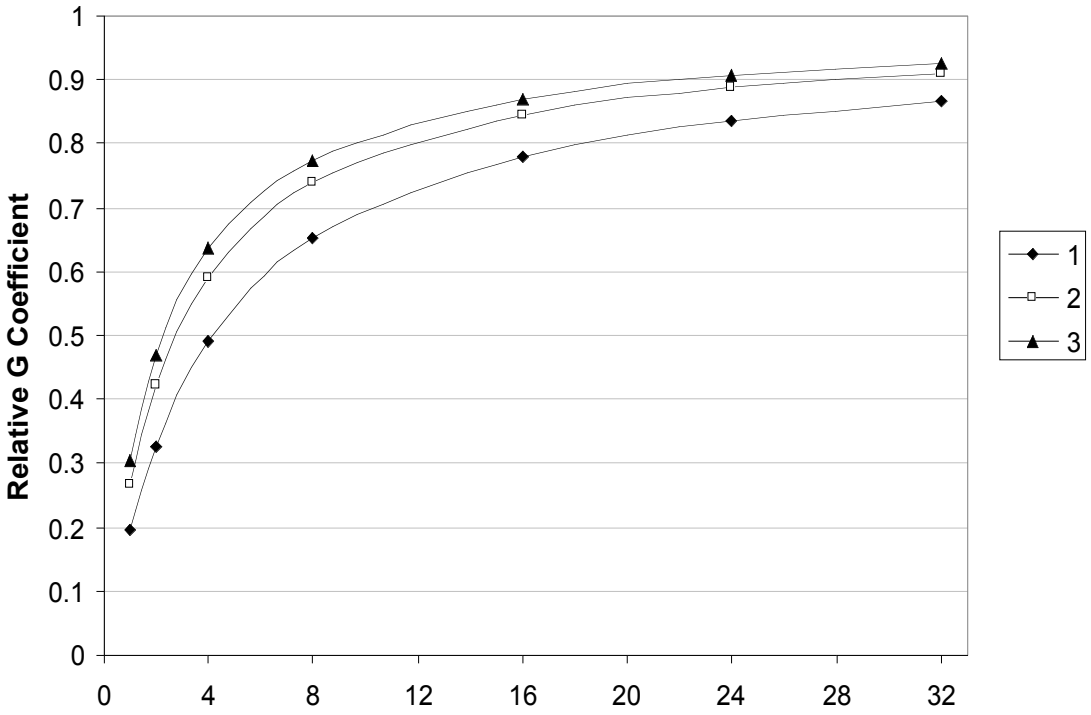

Figure 5. Trade-offs between numbers of mandatory items and occasions needed to achieve various levels of relative Generalizability in S mapping techniques.

Based on the data obtained from our study, S has higher reliability than C under the same condition. Moreover, our scoring experience showed that agreement was much more difficult to reach in C scoring than S scoring because of the openness of C maps. That is, if we included raters as another facet, it might be a greater source of variation in C than in S; consequently, G coefficients for C might have been much lower than those for S. In addition, as mentioned earlier, S maps can be computer scored, which theoretically could eliminate the error made and the cost brought in by raters. Therefore, based on our reliability analysis, our tentative conclusion is that S might be a more dependable and efficient assessment than C.

Conclusion

Due to G theory's power, convenience, and flexibility, we believe it is preferable to classical test theory in examining measurement error and reliability for concept maps. In this paper, we demonstrate the application of G theory to concept maps with the hope of widening the theory's use in concept map research.

We discussed the possible roles G theory may play in a concept map study. We summarized facets of a concept-map measurement that might enter into a concept map G study: concepts, propositions, tasks, response formats, occasions, raters, and scoring systems. With these facets, we could examine many properties of concept map assessments, for example, test-retest stability, inter-rater reliability, internal consistency, and equivalence of different concept map tasks/scoring systems. With information obtained from the G study, we could answer many questions related to the application of concept maps in large-scale assessment. 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? Answers to these (and other) questions can contribute to narrowing the large number of options involved in concept map assessments used in practice.

In addition to the theoretical application of G theory to concept maps, we examined two concept-mapping techniques: construct-a-map with created linking phrases (C) and construct-a-map with selected linking phrases (S). We found that C and S were not equivalent in their measurement errors and reliability. The variance component analysis showed: (1) Person _ Proposition _ Occasion interaction confounded with error accounted for a larger proportion of variability in C than in S; (2) Person _ Item interaction accounted for a larger proportion of variability in S than in C; (3) the G coefficients for S for one item and one occasion were larger than those for C.

A D study showed that fewer items would be needed for S than C to reach the desired level of G coefficients. Otherwise, more occasions have to be applied for C technique in order to get dependable scores. Based on the current study, S is more efficient and reliable mapping technique than C. S might especially be a better candidate than C in the summative assessment. In contrast, our previous research also showed that C better reflected students' partial knowledge and

misunderstandings (Ruiz-Primo, Schultz et al., 2001; Ruiz-Primo, Shavelson et al., 2001; Yin et al., In Press). Accordingly, C might be an effective tool for *formative assessment* in a classroom setting, where fully understanding a student's current thinking is more important than scores (Black & Wiliam, 1998).

This study was just a little demonstration of how G theory could contribute to the study of concept-map assessment. As mentioned earlier, many more unknowns in this field could be investigated with G theory. With more technical information about concept map assessments, assessment researchers could better utilize concept mapping, this promising assessment tool, to help understand what students know and finally improve students' learning.

Reference

Acton, W., H., Johnson, P. J., & Golldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86*, 303-311.

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

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

Baker, E. L., Niemi, D., Novak, J. D., & Herl, H. (1991, July). *Hypertext as a strategy for teaching and assessing knowledge representation.* Paper presented at the NATO Advanced Research Workshop on Instructional Design Models for Computer-Based Learning Environments, The Netherlands.

Black, P., & Wiliam, D. (1998). Inside the Black Box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: John Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373-399.

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

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

Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., Herl, H. E., & O'Neil, H. F. (2001). *Examing the validity of knowledge mapping as a measure of elementary students' scientific understanding*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lay-Dopyera, M., & Beyerbach, B. (1983). *Concept mapping for individual assessment.* Syracuse, NY: School of Education, Syracuse University.

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

Markham, K. M., Mintzes, J. J., & Jones, M. G. (1994). The concept map as a research and evaluation tool: Further evidence of validity. *Journal of Research in Science Teaching, 31*(91-101).

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36*(4), 475-492.

Nakhleh, M. B., & Krajcik, J. S. (1991). *The effect of level of information as presented by different technology on students' understanding of acid, base, and pH concepts.* Paper presented at the annual meeting of the National Association for the Research in Science Teaching, Lake Geneva, WI.

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

Novak, J. D., Gowin, D. B., & Johansen, G. T. (1983). The use of concept mapping and knowledge vee mapping with junior high school science students. *Science Education, 67*(625-645).

Pottenger, F. M., & Young, D. B. (1996). *The Local Environment, Foundational Approaches in Science Teaching* (Second Edition). Honolulu, Hawaii.

Rice, D. C., Ryan, J. M., & Samson, S. M. (1998). Using concept maps to assess student learning in the science classroom: Must different methods compete? *Journal of Research in Science Teaching, 35*(10), 1103-1127.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*(2), 99-141.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38*(2), 260-278.

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

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problem and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569-600.

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

Schreiber, D. A., & Abegg, G. L. (1991). *Scoring student-generated concept maps in introductory college chemistry.* Paper presented at the annual meeting of the National Association for the Research in Science Teaching, Lake Geneva, WI.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessment. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, G. C. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*, 61-71.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. (In Press) Windows into the Mind. *International Journal of Higher Education*.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, J. R. (In Press). A comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching*.

Footnote

1. Since this AERA session is "Issues in Population Generalization", we assumed the audiences have understood Generalizability Theory; therefore, we did not define or explain most terms related to G theory but simply use them as an analysis tool.

Appendix 1 Concept Map Components and Variations Identified

Map assessment components	Variations	Instances
Task	• Task demands	Students can be asked to: • fill in a map • construct a map from scratch • organize cards • rate relatedness of concept pairs • write an essay • respond to an interview
	• Task constraints	Students may or may not be: • asked to construct a hierarchical map • provided with the concepts used in the task • provided with the concept links used in the task • allowed to use more than one link between nodes • allowed to physically move the concepts around until a satisfactory structure is arrived at • asked to define the terms used in the map • required to justify their responses • required to construct the map collectively
	• Content structure	The intersection of the task demands and constraints with the structure of the subject domain to be mapped.
Response	• Response mode	Whether the student response is: • paper and pencil • oral • computerized
	• Format characteristics	Format should fit the specifics of the task
	• Mapper	Whether the map is drawn by a: • student • teacher or researcher
Scoring system	• Score components of the map	Focus is on three components or variations of them: • propositions • hierarchy levels • examples
	Use of a criterion map	Compare a student's map with an expert's map. Criterion maps can be obtained from: • one or more experts in the field • one or more teachers • one or more top students
	• Combination of map components and a criterion map	The two previous strategies are combined to score the student's map.

Ruiz-Primo and R.J. Shavelson, 1996, *Journal of Research in Science Teaching*, 33, p. 586. Copyright 1996 by the National Association for Research in Science Teaching. Reprinted with permission.