Higher Education (2005) 49: 413–430
DOI 10.1007/s10734-004-9448-9

Windows into the mind *

RICHARD J. SHAVELSON[1]*, MARIA ARACELI RUIZ-PRIMO[1] & EDWARD W. WILEY[2]
[1]*School of Education, Stanford University, Stanford, CA 94305-3096, USA;* [2]*McKinsey & Company, Inc. (*author for correspondence, e-mail: richs@stanford.edu)*

Abstract. As faculty, our goals for students are often tacit, hidden not only from students but from ourselves as well. We present a conceptual framework for considering teaching goals – what we want our students to achieve – that encourages us to think more broadly about what we mean by achieving in our knowledge domains. This framework includes declarative knowledge ("knowing that"), procedural knowledge ("knowing how"), schematic knowledge ("knowing why") and strategic knowledge ("knowing when, where and how our knowledge applies"). We link the framework to a variety of assessment methods and focus on assessing the structure of declarative knowledge – knowledge structure. From prior research, we know that experts and knowledgeable students have extensive, well-structured, declarative knowledge; not so novices. We then present two different techniques for assessing knowledge structure – cognitive and concept maps, and a combination of the two – and provide evidence on their technical quality. We show that these maps provide a window into the structure of students' declarative knowledge not otherwise tapped by typical pencil-and-paper tests. These maps provide us with new teaching goals and new evidence on student learning.

As university faculty teaching in, for example, science, mathematics and engineering, our goals for students' learning are often tacit. They are hidden not only from our students, but also from ourselves. However, when we assess students' learning we begin to make our goals public, or at least those goals that are easily tested. Especially with large classes, these tests tend to be pencil-and-paper, often some combination of multiple-choice and open-ended constructed responses. Explicitly, then, our goals for student learning are often highly content oriented, focusing on facts and concepts and algorithms and procedures, looking for "right answers". If we could make all our goals explicit to our students and ourselves, we might expect much more of their learning and our teaching.

* Based on an invited address, Facoltà di Ingegneria dell'Università degli Studi di Ancona, June 27, 2000. This research was supported, in part, by the Center for Research on Evaluation, Standards, and Student Testing (Grant R117G10027), and by the National Science Foundation (Nos. ESI 95-96080). The opinions expressed here represent those of the authors and not necessarily those of the funding agency.

In this paper, we attempt to make explicit goals for teaching and learning in the sciences and show how they might be linked to the assessment of learning. We then focus on an important but often neglected aspect of learning science – the conceptual structure of the domain. We describe some measurement techniques that provide windows into the structure of our students' minds. We also present empirical evidence showing that we can measure some important aspects of conceptual structure reliably, and that the measures provide information different from the usual pencil-and-paper measures that focus on the amount of conceptual and procedural knowledge learned. These structural measures provide added insight into learning and teaching.

Knowledge goals for teaching and learning

We believe that the goals of teaching and learning science include knowledge (cognition), emotion and motivation. Here we focus on knowledge, broadly speaking, and specifically on the structure of conceptual ("declarative") knowledge.

Our working definition of the cognitive goals, which we call "science achievement" when expected of students (Shavelson and Ruiz-Primo 1999a; see Figure 1 below), involve "knowing that" – *declarative* (factual, conceptual) knowledge. For example, force equals mass times acceleration. Achievement also involves knowing how to do something – *procedural* (step-by-step or condition-action) knowledge. For example, knowing how to measure the density of an object. And we also seek to teach *schematic* knowledge – "knowing why". For example, schematic knowledge involves knowing why New England has a change of seasons. Finally, we want students to develop *strategic* knowledge – knowledge of when, where and how their knowledge applies, and to check to see if their application of this knowledge is reasonable. For example, experts know when to apply Newton's first law given a problem to solve dealing with force and motion whereas novices are attracted to the surface features of the problem (Chi, Feltovich and Glaser 1981). Presumably experts have scientifically well justified "mental models" (Gentner and Stevens 1983) that deal with knowing *why* something is so, getting at deep conceptual understanding or misunderstanding. Finally, we recognize that the cognitive tools of planning and monitoring in general and especially in the domain of science tacitly influence how this knowledge is used and thereby influence achievement, although they are not easily measured directly by tests.

Each type of knowledge has a set of characteristics. For example, we can ask about the *extent* of declarative – factual and conceptual – knowledge. That is, how extensive is a student's conceptual knowledge of force and motion?

Or we can ask about the *structure* of this conceptual knowledge, the focus of this paper. For example, how well structured is this student's knowledge. We know that physics experts have extensive *and well-structured* knowledge (Chi et al. 1981).

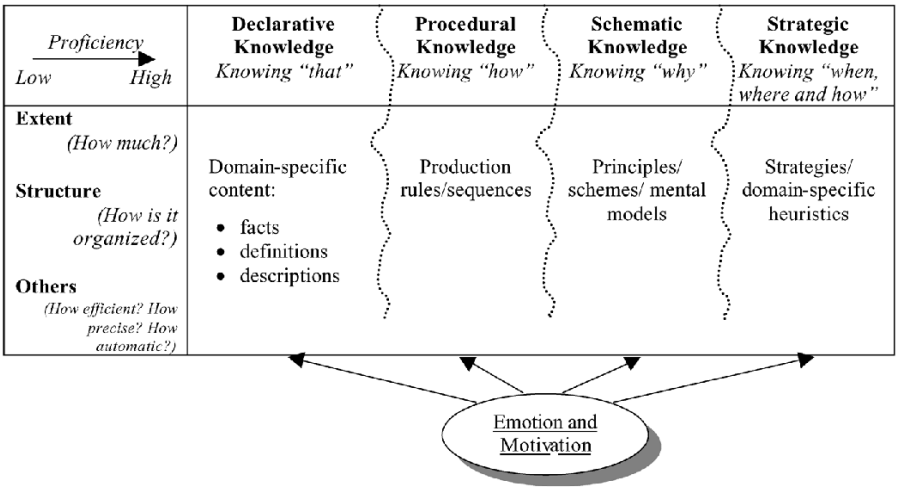

Figure 1. Conceptual framework for characterizing science goals and student achievement.

For each combination of knowledge type and characteristic, we have begun to identify assessment methods (e.g., Li and Shavelson 2001, see Figure 2). For example, to measure the *extent* of *declarative* knowledge, multiple-choice and short-answer questions are cost-time efficient and very reliable. To measure the *structure* of *declarative* knowledge, however, multiple-choice tests fall short and concept- and cognitive-maps provide valid evidence of conceptual structure (Ruiz-Primo and Shavelson 1996a; Shavelson and Ruiz-Primo 1999). And to measure procedural knowledge, performance assessments, not paper-and-pencil assessments, are needed (e.g., Ruiz-Primo and Shavelson 1996b; Shavelson, Baxter and Pine 1992). However, life is not quite this straightforward. While we can conceptually distinguish knowledge types, in practice they are difficult to distinguish. For example, conducting an investigation to find out which substance (powder) is most soluble in water requires knowing what "solubility" means. And assessment methods do not line up perfectly with knowledge types and characteristics. For example, Sadler (1998) provided evidence of the validity of multiple-choice tests for measuring mental models (schematic knowledge) in astronomy (see also Stecher, Klein et al. 2000). Moreover, strategic knowledge is rarely ever directly measured. Rather, it is implicated whenever other types of knowledge are accessed.

	Declarative Knowledge	Procedural Knowledge	Schematic Knowledge
Extent	Multiple Choice Fill In Science Notebooks	Performance Assessments Science Notebooks	Performance Assessments Predict, Observe, Explain (POE)* Multiple Choice
Structure	Concept Maps Cognitive Maps	Procedure Maps	Models/Mental Maps

* (White & Gunstone, 1992)

Figure 2. Possible links between types and characteristics of science knowledge and some assessment method.

Nevertheless, there is an empirical basis for the framework based on our research in pre-college education (e.g., Li and Shavelson 2000; Schultz, 1999; most research on concept maps has been conducted in pre-college science education). In her dissertation, for example, Schultz (1999) conducted a randomized experiment comparing two treatment conditions – individual- and complex-instruction-based science (ecology) inquiry – and measured middle-school students' learning with multiple-choice, concept-map, and performance tests. In addition, she used the California Test of Basic Skills (CTBS) reading test (declarative knowledge – verbal ability) as a covariate (N = 109). The pattern of correlations among these measures was what we expected from our working definition of achievement. Measures of declarative knowledge correlated higher with each other than with the performance assessment. The two measures of the extent of declarative knowledge, the multiple-choice and reading tests, correlated 0.69. These two measures correlated 0.60 and 0.53 (respectively) with a measure of the structure of declarative knowledge (concept map). And these three measures correlated 0.25, 0.33, and 0.43 (reading, multiple-choice and concept-map) with the performance assessment.

This paper focuses on some instruments for measuring science achievement that are consistent with our broader notion of achievement. More specifically, the paper focuses on *concept and cognitive maps* as instruments to assess students' knowledge structure (or connected understanding) and provides information about the technical quality of these instruments. Concept and cognitive maps have been said to provide a "window into students' minds". Although this might be an overstatement, there is evidence that they *do* seem to provide information on the structure of complex ideas.

Conceptual structure and its measurement

Cognitive psychologists posit "the essence of knowledge is structure" (Anderson 1984, p. 5). Research on the cognitive aspects of science learning

has provided evidence that professional scientists and successful students develop elaborate, well differentiated, and highly integrated frameworks of related concepts (e.g., Chi, Feltovich and Glaser 1981; Glaser and Bassok 1989; Mintzes, Wandersee and Novak 1997; Pearsall, Skipper, and Mintzes 1997; Shavelson 1972). This means that as expertise in a domain grows, through learning, training, and/or experience, the elements of knowledge become increasingly interconnected (cf. Chi, Glaser and Farr 1988). Indeed, expert performance seems to lie in the organization of the expert's domain knowledge. Experts possess a large knowledge base (what we have called extent of knowledge) that is organized into elaborate, integrated structures, whereas novices tend to posses less domain knowledge and a less coherent organization of it (cf. Chi, Glaser and Farr 1988; Zajchowski and Martin 1993).

Theory and measurement techniques

Assuming that knowledge within a content domain is organized around central concepts, to be knowledgeable in the domain implies a highly integrated conceptual structure among those concepts. Researchers have taken different representational approaches to capture this organizational property of knowledge (e.g., Goldsmith, Johnson and Acton 1991; Novak and Gowin 1984; White and Gunstone 1992). What seems common among them is the assumption that knowledge structures can be modeled as associative network nodes linked together, the strength of the links depending on, for example, the frequency with which the link is traversed.

Direct approaches to the measurement of structure, called concept maps (Novak and Gowan 1984), assume a semantic model of memory (Fisher 2000) which builds on the associative model by labeling the lines linking concepts with specific relationships describing the meaning of the relationship (see Figure 3). When a node is stimulated, a memory search radiates out along relevant relationships (lines).

A concept map, then, is a graph in which the nodes represent concepts, the lines represent relations, and the labels on the lines represent the nature of the relation between concepts. A pair of nodes and the labeled line connecting them is defined as a *proposition*. Students are asked to construct a "map" that makes explicit how they relate concept pairs. Students may be asked to create a map in an explicit science domain with concepts provided by the teacher. Or they may be provided a skeleton map with some nodes filled in and all lines labeled and their task is to fill in the missing nodes using a set of concepts provided by the teacher (see Ruiz-Primo, Schultz, Li and Shavelson 2001 for details).

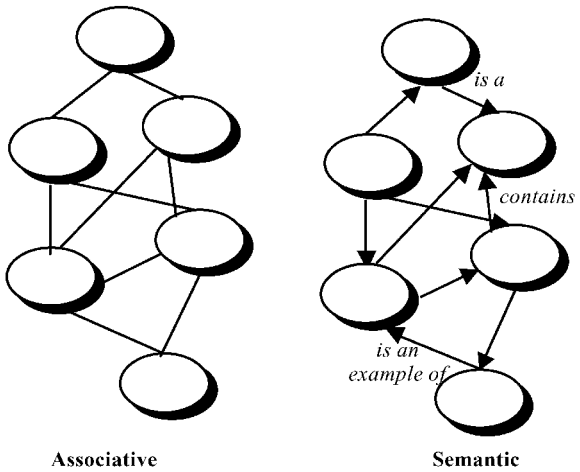

Figure 3. Associative and semantic models of memory: nodes represent concepts and lines represent relationships.

Indirect approaches to measuring knowledge structure, called *cognitive maps* (e.g., Goldsmith, Johnson and Acton 1991), assume an associative model of memory (Anderson 1983; Shavelson 1974). The model is a network where the nodes are concepts and the lines are relations between concepts (see Figure 3). When one (or more) nodes are stimulated, a search radiates out from that node first to the most proximate nodes, and then beyond. Consequently, the order of recall, for example, provides information about memory structure and the sequence of or similarity between concepts provides information about distances among concepts. Cognitive maps probe a student's knowledge structure by asking, for example, the student to rate the similarity between concepts (e.g., Goldsmith, Johnson and Acton 1991), to associate words (e.g., Shavelson 1972, 1974), or to sort concepts into groups based on their similarity (e.g., Shavelson and Stanton 1975). These ratings, associations, and groupings are treated as distances (or proximities) among concepts and translated, mathematically, into network representations where the nodes represent concepts and the lines represent relationships between concept pairs.

Evidence on the reliability and validity of interpretations of concept maps *as measures of declarative knowledge structure*

Variations in concept maps abound, as we have pointed out. Perhaps the most commonly used type of concept map is what we call, "construct a map". Students are provided with a set of (10–20) concepts and asked to draw a map. The result might look like the map of the water cycle in Figure 4. We

have explored the reliability and validity of interpretations of these and other concept maps with students primarily in elementary and middle school.

Reliability. In a typical reliability study, we have two or more "raters" score concept maps using a scoring system that identifies and evaluates possible propositions. A student's concept-map score is defined as the proportion of valid propositions in her map to the total possible valid propositions identified in a criterion (e.g., instructor's) map. We interpret both the agreement between raters on proposition scores and the correlation between raters' proposition scores as evidence of reliability. We have found that raters can reliably score concept maps. Interrater reliabilities, for example, are around 0.90 on a scale from 0 to 1.0.

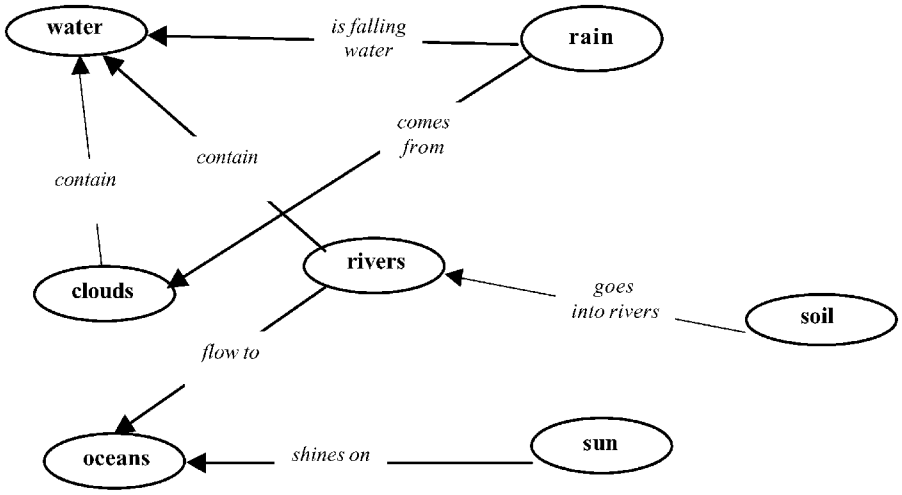

Figure 4. An eleven-year old's construct-a-map concept map from White and Gunstone (1992, p. 16).

Another aspect of reliability is the sensitivity of concept-map scores to the particular set of concepts provided by the instructor in the map. To what extent would a student's score differ from one sample of concepts to another? We have explored this question and found high reliability – "generalizability" (Ruiz-Primo, Schultz and Shavelson 1996c; Shavelson and Webb 1991) – from concept sample to sample, *and* from rater to rater (see Table 1). The first column of the table shows the sources of variation in students' test scores. One purpose of the concept map is to reliably distinguish among persons; this source of variation is expected and desirable. All other sources contribute to "absolute" error in the measurement and are reflected in the dependability coefficient (Φ) in Table 1; those sources that contain the person component

(e.g., p × s) form the basis of error for the relative reliability estimate (ρ^2) in Table 1.

Notice that raters contribute very little variation to scores, regardless of mapping technique or concept sample, a finding that further extends our finding of high interrater reliability (0.90 above). Hence our conclusions that raters can score concept maps reliably. Secondly, the variability contributed by the unique combination of person and sample of concepts is moderate – students have partial knowledge and vary in which parts of their conceptual structure are consistent with a scientifically justifiable representation. Importantly, the variability hardly changes moving from students' performance across three maps including the "no concept map" where students generated their own concepts, and the two maps with randomly selected concepts from the domain of molecular structures. Finally, the reliability generalizing across raters and concept samples is quite respectable, 0.78 (one rater and one concept sample) or greater (more raters, samples or both).

Table 1. Percent of total variability and generalizability (reliability) coefficients* for concept-map scores

Source of variation		Percentage of variance components		
		No Concepts		Sample A & B
Person (p)		71.64		78.67
Rater (r)		0.15		0.0
Sample (s)		0.0		0.0
p × r		0.0		0.79
p × s		22.81		17.64
r × s		0.01		0.18
prs,e		5.37		2.69
ρ^2	$(n_1 = 2, n_2 = 3)$	0.89	$(n_1 = 2, n_2 = 2)$	0.88
ϕ		0.89		0.88
ρ^2			$n_1 = 1, n_2 = 1$	0.78
ϕ				0.78

*ρ^2 is the generalizability (reliability) coefficient for consistency in rank-ordering scores; ϕ is the dependability coefficient for absolute scores.

Validity. In a number of studies, we have examined the evidentiary basis of the claim that concept maps measure important aspects of knowledge structure – i.e., validity claims (Ruiz-Primo, Schultz and Shavelson 1996c). In some cases, our studies focus on whether we can distinguish concept-map scores

from multiple-choice scores. While we claim that both measurement methods can tap declarative knowledge, the maps reflect the structure of knowledge and the multiple-choice reflect the extent of knowledge. We have commonly found correlations among these scores in the range of 0.45–0.64 (across no-concepts-given maps and maps constructed with a set of concepts provided by the person assessing).

We have also compared different concept-mapping techniques, asking whether they provide the same information about students' knowledge structures (Ruiz-Primo, Schultz, Li and Shavelson 2001). Different mapping techniques may tap different aspects of knowledge structure. Take, for example, the nature of a mapping task. One dimension in which tasks can vary is the constraints imposed on students in representing their connected understanding. We have considered construct a map, as the "benchmark" technique. In this technique, students are asked to construct a map from scratch. This technique varies as to how much information is provided by the assessor (e.g., the assessor may provide the concepts). Students draw their maps on a piece of paper. Scoring systems vary from counting the number of nodes and linking lines to evaluating the accuracy of propositions. Several variations of maps provide cost and time efficient means for assessment. For example, one mapping technique asks student to fill in the randomly missing nodes with terms provided by the instructor. Another version provides nodes filled in but randomly deletes the labels from (say) 40 percent of the lines (see Figure 5). In either case, scoring is rapid – right or wrong – and students can complete the task rapidly.

In our study of these three mapping techniques, we created two randomly parallel forms of the fill-in-the-nodes and fill-in-the-lines maps and compared students' performance on them with a construct a map (Figure 6). To evaluate the equivalence of the two randomly parallel versions of the fill-in maps, we began by comparing their means and standard deviations (Table 2). The means for the two versions of the fill-in-the-nodes maps were very close, well within sampling error, as were the standard deviations. The means and standard deviations for the fill-in-the-lines maps were quite similar. Importantly, the latter maps were more difficult for the students than the former (lower means). Indeed, the students' mean scores on the fill-in-the-nodes maps were at ceiling (mean of 11 out of 12 possible!).

We then compared students' performance on the fill-in maps with their performance on the construct a map (see Figure 6). We found that the former provided high estimates of structural knowledge whereas the latter provided a more sobering view with a mean of 53.91 out of a possible 135 (instructor's map criterion).

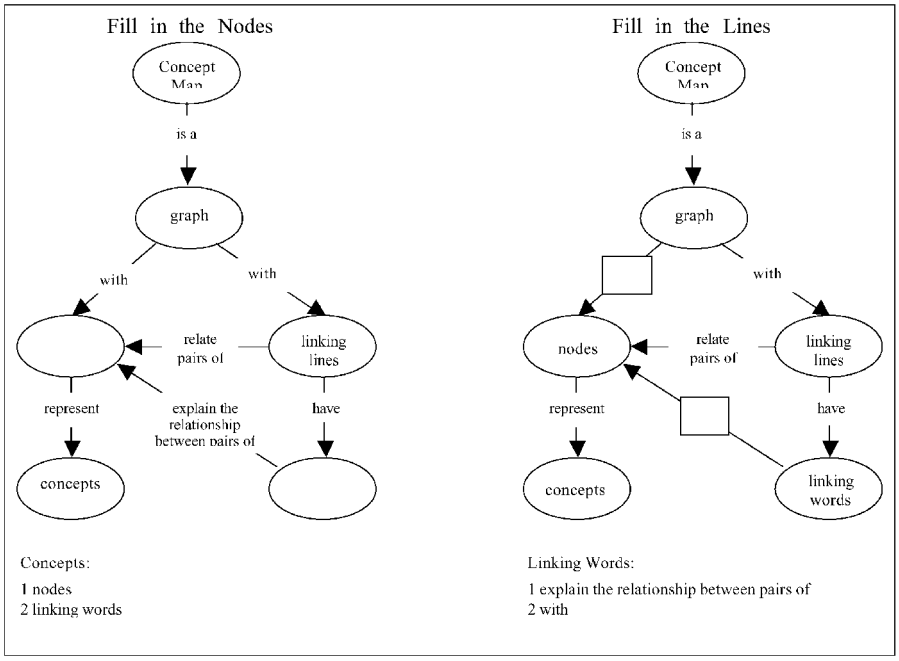

Figure 5. Examples of fill-in-the-nodes and fill-in-the-lines skeleton maps.

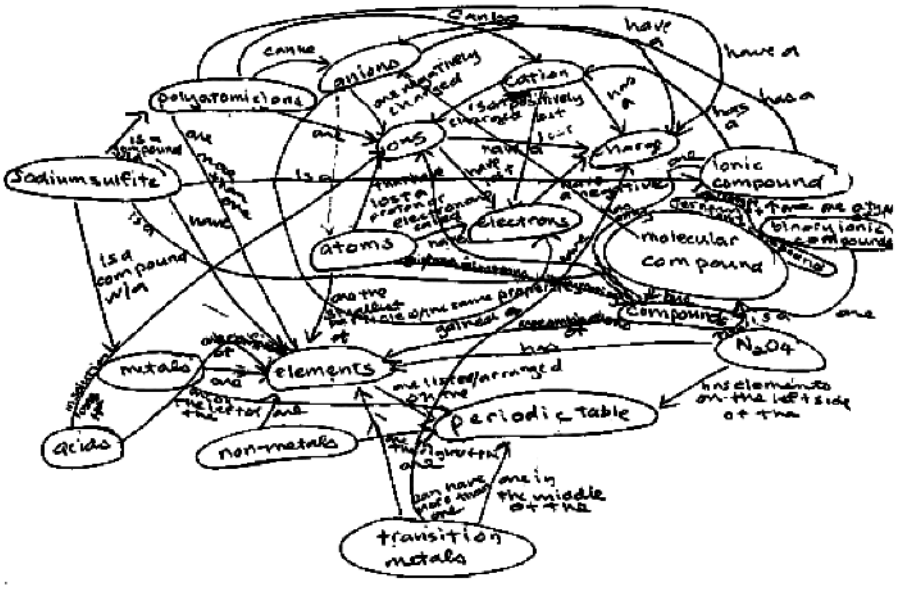

Figure 6. High school chemistry student's construct-a-map concept map.

We also examined the correlations between construct-a-map and fill-in-a-map scores (ranging from 0.40–0.47) with one another and with multiple-choice test scores (0.37–0.53) on the same concepts (elements, atoms, compounds, etc.; see Table 2). Apparently each of these methods measure somewhat different aspects of declarative knowledge.

Table 2. Correlations (and reliabilities) among concept maps and a multiple-choice test

Type of assessment		Structure		Extent		Standard
	C-M	FI-N	FI-L	M-C	Mean	deviation
Construct a map (C-M)	(0.99)					
Fill in nodes (FI-N)	0.47	(0.71)				
Fill in lines (FI-L)	0.44	0.40	(0.85)			
Multiple choice (M-C)	0.44	0.37	0.53	(0.74)		

An examination of validity claims needs to go beyond correlations. A thorough investigation of the claim that concept maps measure important aspects of structural knowledge must also include evidence of students' "thinking" as they construct their maps. If these various techniques tap the same structural knowledge, we would expect a cognitive analysis to reveal similar patterns of thinking. To find out about their thinking, we asked students to "think aloud" as they carried out the mapping activity (Ericsson and Simon 1993, 1998). Their think aloud protocols were audio-recorded, transcribed, and analyzed. We found a striking difference between construct-a-map and fill-in techniques. Students, talking aloud as they performed the construct-a-map assessment tended to explain to themselves the concepts and their links, and to a lesser extent, monitor what they were doing to make sure they had covered all relevant possibilities. In contrast, students rarely offered explanations while filling in maps. Rather, they most often monitored their performance to see that everything had been filled in.

Based on this research, albeit primarily from pre-college education, we have reached the following set of tentative conclusions about concept maps:

- Students' maps can be consistently scored by different raters even when complex judgments are required
- Different mapping techniques provide different pictures of students' declarative knowledge structure. The construct-a-map method provides opportunities to reveal students' conceptual understanding to a great extent than do fill-in maps
- Different mapping techniques impose different cognitive demands on students. Highly structured mapping techniques like fill-in allow

students to respond by elimination or guessing (hence high monitoring of performance); whereas constructed response requires considerable explaining of concepts and their relationships

- Correlations between concept-map and multiple-choice scores are positive and moderate suggesting that these two types of assessments measure overlapping but somewhat different aspects of declarative knowledge.

Evidence on the reliability and validity of interpretations of cognitive maps *as measures of declarative knowledge structure: A view to the future*

Variations in cognitive maps abound. Perhaps the most commonly used type of cognitive map is generated using similarity judgments. With this method, students are provided with a set of (10–20) concepts and asked to evaluate the similarity between all possible pairs. An example from a college course on introductory statistics is provided in Figure 7.

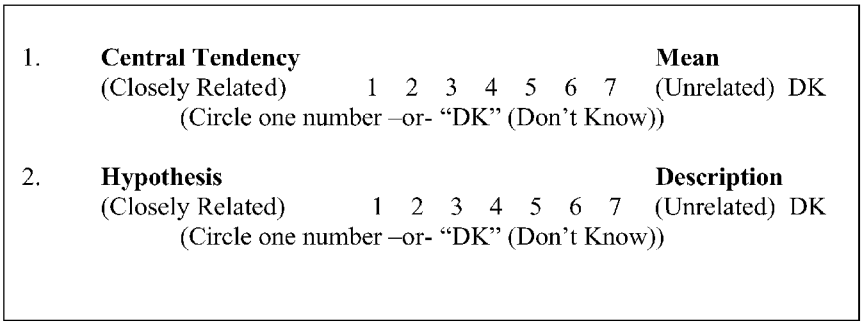

Figure 7. Similarity judgment task – rate the similarity of each pair of terms by circling one of the numbers provided or "DK – Don't Know".

We have, in a small way, studied the impact of an introductory statistics course on the extent and structure of 28 graduate-students' knowledge both before and after 3 months of instruction (Wiley 1998; see also Schau and Mattern 1997). In addition to collecting similarity judgments, we collected construct-a-map concept-map and multiple-choice data. Consequently, we had two different measures – one direct (see Figure 8) and one indirect – of students' declarative knowledge structures, and a measure of the extent of their declarative knowledge.

The similarity judgments were placed in a concept x concept proximity matrix. The entries in a cell of the matrix represent a student's judgment of the similarity between that pair of concepts (Figure 9).

These data were then submitted to an algorithm that translated the proximity matrix into a network representation (Figure 10). This network can be

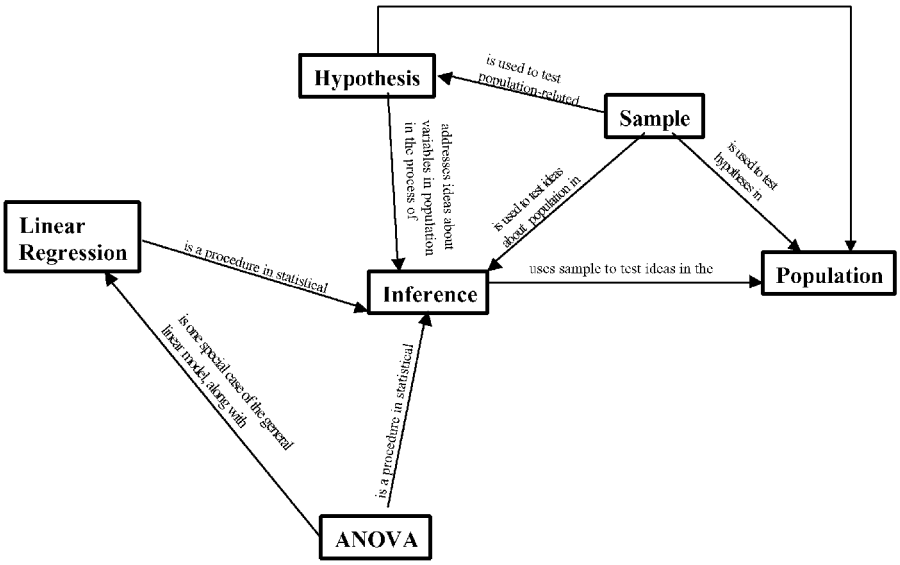

Figure 8. Portion of a student's concept map of the domain of introductory statistics (unpublished study by Wiley, Stanford University, 1998).

interpreted as a representation of a student's declarative knowledge structure (e.g., Goldsmith, Johnson and Acton 1991; Shavelson 1972).

We correlated scores on the three measures both at pretest and at posttest. If students were not particularly knowledgeable about statistics at pretest, we would expect the correlations to be low, drawing on general ability more than on domain-specific statistical knowledge. In contrast, at posttest, we expected the correlation between the two alternative measure of structure – cognitive and cognitive-map scores – to be higher than either measure's correlation with the multiple-choice test. This is roughly what we found with a small sample of students (Table 3).

The results of this small study are directionally consistent with those reported in an extensive literature on cognitive maps: cognitive (and concept) maps provide representations of students' declarative knowledge structures, and this information is somewhat different from that provided by multiple-choice tests. This suggests that these assessment methods provide a valuable complement to multiple-choice tests toward providing a comprehensive measurement of science achievement. Moreover, computer-based assessment of cognitive structure is straightforward and can be widely implemented (with a multitude of software now available).

We might speculate about the possibility of combining *indirect* and *direct* measures of knowledge structure, as well. What if we collected structure

	Cent. Tend.	Corr.	Desc.	Regr.	ANOVA	Hyp.	Inf.	Mean	Median	Mode	Pop.	Sample	Stat.	t-test	Variab.	Variance
Central Tendency	0															
Correlation	8	0														
Description	1	4	0													
Regression	9	3	8	0												
ANOVA	9	7	9	1	0											
Hypothesis	8	7	8	3	3	0										
Inference	8	8	5	2	3	1	0									
Mean	1	6	2	9	9	9	8	0								
Median	1	8	6	8	9	9	8	3	0							
Mode	1	8	4	9	8	9	8	3	2	0						
Population	8	8	8	3	3	2	1	8	7	7	0					
Sample	3	5	2	8	8	6	2	6	5	8	1	0				
Statistics	1	6	1	4	7	4	1	6	5	6	3	2	0			
t-test	8	8	7	2	1	4	2	7	8	9	4	9	7	0		
Variability	7	8	1	9	7	9	9	8	8	9	9	3	4	8	0	
Variance	8	7	3	8	5	8	7	7	9	8	8	3	6	8	1	0

Figure 9. A student's concept × concept proximity matrix of similarity judgment ratings (Wiley 1998) (since the matrix is symmetric, only the data above or below the main diagonal were preserved).

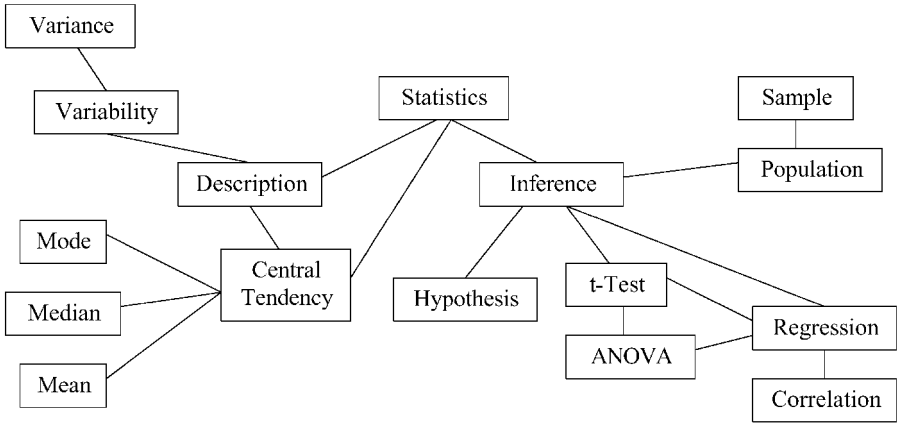

Figure 10. Network representation of a student's similarity judgments (proximities; Wiley 1998).

Table 3. Correlations among concept-map, cognitive-map, and multiple-choice scores

Type of assessment	Multiple choice	Concept map
Pretest		
• Concept map	0.155	
• Similarity	0.554*	0.251
Posttest		
• Concept map	0.330	
• Similarity	0.434*	0.706*

*$p < 0.01$.

information indirectly with a computer, having a student provide similarity judgments, translating these judgments into a proximity matrix and then translating the proximities into a network representation? What if we then provided a network representation immediately to the student and asked the student to add and/or remove lines? And what if we then asked the student to label the lines with short descriptions of how concept pairs went together (propositions)? By combining both methods, we might just get a very good representation of declarative knowledge structure, one that did not depend on the student's artistry or on the student's practicing certain maps in preparation for our test! Computer assessment of knowledge structure that combines a statistical and a student-generated representation seems to be an important next step in this work.

Concluding comments

Concept and cognitive maps provide visual representations of some aspects of a student's declarative knowledge structure (connected understanding) in a particular knowledge domain. The evidence reported here suggests that they may make useful windows into a student's mind if care is taken in which types of maps are used and interpretations validated.

Moreover, concept and cognitive maps can be useful tools for representing what knowledge students have acquired over a long period of time and how they interrelate that knowledge. Changes in the structure of the representation may mean changes in students' conceptual frameworks that can help determine the type of changes that take place as a result of instruction.

Concept and cognitive maps can be thought as a unique source of information that may help to know what and how students are learning and accommodating the new information learned. Research has shown that successful students have well developed and interconnected knowledge structures.

Concepts maps can be used in a variety of forms (mapping techniques) and the nature of the particular mapping technique(s) employed will affect the potential costs and benefits of the assessment process. Cognitive maps also may be constructed in many ways – e.g., through similarity judgments, concept sorting, tree construction, and word association – and have the benefit that they do not encourage "teaching to the test"; however, they contain less information than concept maps with their labeled lines. Hence, the combination of concept and cognitive maps explored in our small graduate statistics study provides an interesting possibility for further exploration.

Note

1. Complex instruction integrates individual and group work where the groups are designed to include individuals varying in demographic characteristics and prior knowledge and the tasks that groups work on are open ended.

References

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University.

Anderson, R.C. (1984). Some reflections on the acquisition of knowledge. *Educational Researcher* 13(10), 5–10.

Chi, M.T.H., Feltovich, P.J. and Glaser, R. (1981). 'Categorization and representation of physics problems by experts and novices', *Cognitive Science* 5, 121–152.

Chi, M.T.H., Glaser, R. and Farr, M.J. (1988). *The Nature of Expertise*. Hillsdale, NJ: Lawrence Earlbaum Associates, Publishers.

Ericsson, A.K. and Simon, H.A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT.

Ericsson, A.K. and Simon, H.A. (1998). 'How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking', *Mind, Culture & Activity* 5(3), 178–186.

Fisher, K.M. (2000). 'SemNet software as an assessment tool', in Mintzes, J.J., Wandersee, J.H. and Novak, J.D. (eds.), *Assessing Science Understanding: A Human Constructivist View*. New York: Academic Press, pp. 197–221.

Gentner, D. and Stevens, A.L. (eds.) (1983). *Mental Models*. Hillsdale, NJ: Erlbaum.

Glaser, R. (1991). 'Expertise and assessment', in Wittrock, M.C. and Baker, E.L. (eds.), *Testing and Cognition*, pp. 17–39.

Glaser, R. and Bassok, M. (1989). 'Learning theory and the study of instruction', *Annual Review of Psychology* 40, 631–666.

Goldsmith, T.E., Johnson, P.J. and Acton, W.H. (1991). 'Assessing structural knowledge', *Journal of Educational Psychology* 83(1), 88–96.

Li., M. and Shavelson, R.J. (2001). 'Examining the links between science achievement and assessment'. *Presented at the annual meeting of the American Educational Research Association*, Seattle, WA.

Mintzes, J.J., Wandersee, J.H. and Novak, J.D. (1997). *Teaching Science for Understanding*. San Diego: Academic Press.

Novak, J.D. (1990). 'Concept mapping: A useful tool for science education', *Journal of Research in Science Teaching* 27(10), 937–949.

Novak, J.D. and Gowin, D.R. (1984). *Learning How to Learn*. New York: Cambridge Press.

Pearsall, N.R., Skipper, J.E.J. and Mintzes, J.J. (1997). 'Knowledge restructuring in the life sciences. A longitudinal study of conceptual change in biology', *Science Education* 81(2), 193–215.

Ruiz-Primo, M.A. and Shavelson, R.J. (1996a). 'Problems 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. and Shavelson, R.J. (1996b). 'Rhetoric and reality in science performance assessments: An update', *Journal of Research in Science Teaching* 33(10), 1045–1063.

Ruiz-Primo, M.A., Schultz, E.S. and Shavelson, R.J. (1996c). 'Concept map-based assessments in science: An exploratory study'. *Presented at the annual meeting of the American Educational Research Association*, New York, NY.

Ruiz-Primo, M.A., Schutlz, S.E., Li, M. and 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.

Sadler, P.M. (1998). 'Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments', *Journal of Research in science Teaching* 35(3), 265–296.

Schau, C. and Mattern, N. (1997). 'Use of map techniques in teaching applied statistics courses', *The American Statistician* 51(2), 171–175.

Schultz, S.E. (1999). *To Group or not to Group: Effects of Grouping on Students' Declarative and Procedural Knowledge in Science*. Stanford, CA: Unpublished doctoral dissertation.

Shavelson, R.J., Baxter, G.P. and Pine, J. (1992). 'Performance assessments: Political rhetoric and measurement reality', *Educational Researcher* 21(4), 22–27.

Shavelson, R.J. and Ruiz-Primo, M.A. (1999a). 'Leistungsbewertung im naturwissenschaftlichen unterricht' (Evaluation in natural science instruction), *Unterrichtswissenschaft* 27, 102–127.

Shavelson, R.J. and Ruiz-Primo, M.A. (1999b). *On the Psychometrics of Assessing Science Understanding*. New York: Academic Press.

Stecher, B.M., Klein, S.P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R.J. and Haertel, E. (2000). 'The effects of content, format, and inquiry level on performance on science performance assessment scores', *Applied Measurement in Education* 13(2), 139–160.

Shavelson, R.J. (1972). 'Some aspects of the correspondence between content structure and cognitive structure in physics instruction', *Journal of Educational Psychology* 63, 225–234.

Shavelson, R.J. (1974). 'Some methods for examining content structure and cognitive structure in instruction', *Educational Psychologist* 11, 110–122.

Shavelson, R.J. and Stanton, G.C. (1975). 'Construct validation: Methodology and application to three measures of cognitive structure', *Journal of Educational Measurement* 12, 67–85.

Shavelson, R.J. and Webb, N.M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: SAGE.

White, R.T. and Gunstone, R. (1992). *Probing Understanding*. New York: Falmer Press.

Wiley, E.W. (1998). *Indirect and Direct Assessment of Structural Knowledge in Statistics*. Stanford, CA: Stanford University School of Education.

Zajchowski, R. and Martin, J. (1993). 'Differences in the problem solving of stronger and weaker novices in physics: Knowledge, strategies, or knowledge structure', *Journal of Research in Science Teaching* 30(5), 459–470.