Evaluating New Approaches to Assessing Learning

CSE Report 604

Richard Shavelson and Maria Araceli Ruiz-Primo
CRESST / University of California, Los Angeles

Min Li
University of Washington

Carlos Cuauhtemoc Ayala
California State University, Sonoma

August 2003

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
301 GSE&IS, Box 951522
Los Angeles, Ca 90095-1522
(310) 206-1532

Project 3.6 Group Activity on Cognitive Validity
Richard Shavelson, Project Director, CRESST / University of California, Los Angeles

The work reported herein was supported under the Educational Research and Development Centers Program, PR / Award Number R305B60002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

EVALUATING NEW APPROACHES TO ASSESSING LEARNING

Richard J. Shavelson and Maria Araceli Ruiz-Primo

CRESST/Stanford University

Min Li

University of Washington

Carlos Cuauhtemoc Ayala

California State University, Sonoma

Abstract

We present a framework for evaluating cognitive claims to the interpretation of assessment scores and provide evidence of its applicability to science achievement. We adapted the idea of the "assessment triangle" (Pellegrino, Chudowsky, & Glaser, 2001), in the form of an assessment square with four tightly linked corners: construct (definition), assessment (task / response / score analysis), observation (cognitive and statistical data), and interpretation (link between observation and construct). In an iterative process of assessment review, the model focuses on four analyses that feed back on one another: conceptual, logical, cognitive, and statistical and / or qualitative. The heart of the model is a knowledge framework consisting of declarative (knowing that), procedural (knowing how), schematic (knowing why), and strategic (knowing when knowledge applies) knowledge that underlie achievement. Concrete examples of the model's application are provided.

Introduction

Spurred by calls for education reform in the wake of international comparisons of student performance, developments in cognitive science, and augmentation of international achievement test formats, a myriad of new (and renewed) approaches to assess students' learning have emerged over the past 15 years. Some examples of these new assessments are *performance assessments* (mini-science-laboratory investigations); *concept maps* (a graph with nodes representing concepts, directed lines representing relations between concepts, and line labels explaining the relation); and *predict-observe-explain* demonstrations (e.g., students predict whether two objects will fall at the same rate, observe a demonstration, and then explain what happened). In most all cases, the intent of assessing learning is to go beyond ranking students on their performance to drawing inferences about what they know and are able to do with that knowledge. That is, assessments of learning are interpreted as providing information on *cognitive activities* as well as on performance.

With the burgeoning variety of learning assessments and the cognitive as well as performance interpretations placed on information collected with these assessments, the question arises, "How might we evaluate the quality of information produced by sometimes quite novel approaches to learning assessment?" That was the question posed to us by the conference organizers and the topic of this paper. To address the question, we draw on our work and that of colleagues in the Stanford Education Assessment Laboratory where we encounter this question over and over again. We focus on evidence of the validity of proposed interpretations of learning assessments but also touch on reliability where appropriate.

To be sure, there are other issues that we might have addressed, such as consequential validity—the impact of an assessment practice on students' education. However, these issues would take us beyond the limits of your patience and journal space. The good news is that what is presented here has years of concrete practice behind it; the bad news is that what is not presented here may constitute equally worthwhile alternative approaches, or approaches that address other issues that you are more interested in than what we have presented here. So, in all fairness, readers beware!

Conceptual Framework

Like any practical enterprise, in developing and evaluating assessments we drew on current ideas from theory and practice, specifically cognitive, reliability, and validity theory to do our work. On occasion, we have been asked to talk or write about what we were doing, and that gave us pause to step back and try to figure out what we were doing. This paper is another such opportunity. Fortunately, the work of Pellegrino, Chudowsky and Glaser (2001) and their National Research Council committee provided a stepping-stone for us to understand our own work on evaluating new approaches to learning assessment. Their assessment triangle identified three key elements underlying any assessment (Figure 1). The first element is: "a model of *cognition* and learning in the domain…. [The model explains] how students represent knowledge and develop competence" (Pellegrino, Chudowsky, & Glaser, 2001, p. 44, italics in original). The second element, *observation*, is "a set of beliefs about the kinds of *observations* that … provide evidence of students' competencies…" (Pellegrino, Chudowsky, & Glaser, 2001, p. 44, italics in original). These observations are based on "…tasks or situations that prompt students to say, do, or create something to demonstrate knowledge and skills" (Pellegrino, Chudowsky, & Glaser, 2001, p. 47). And the third element is *interpretation*, a "… process for making sense of the evidence" (Pellegrino, Chudowsky, & Glaser, 2001, p. 44). Interpretation involves "… all the methods and tools used to reason from fallible observations" (Pellegrino, Chudowsky, & Glaser, 2001, p. 48).

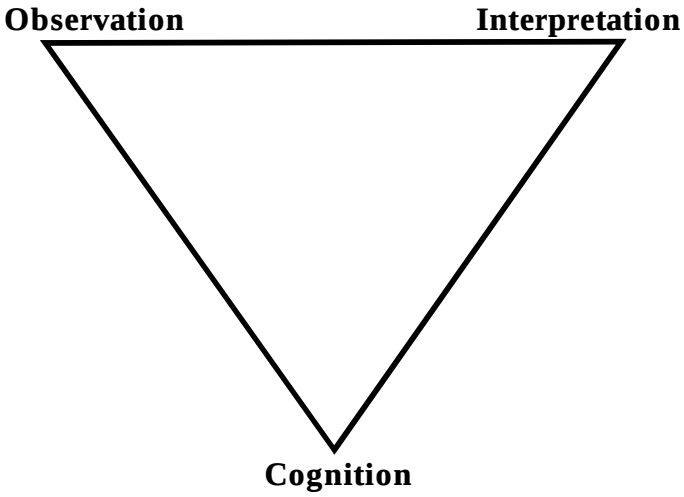


Figure 1. The assessment triangle (Pellegrino, Chudowsky, & Glaser, 2001, p. 44)

We (Ayala, Yin, Shavelson, & Vanides, 2002; Ruiz-Primo, Shavelson, Li, & Schultz, 2001) have taken the liberty of modifying the triangle, changing a label and expanding it into what we call "the assessment square" (see Figure 2). Although for simplicity the assessment square looks like a nice neat series of steps, in reality it is an iterative process in which in early stages of assessment development the corners of the square loop back to earlier corners.

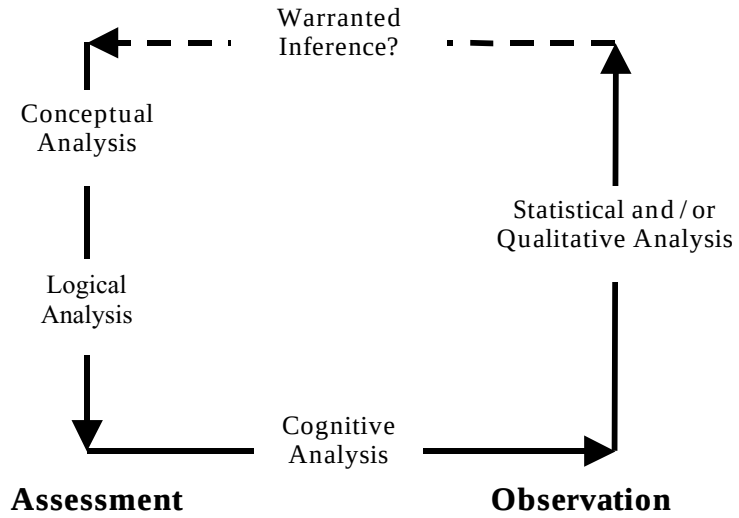


Figure 2. The assessment square (Adapted from Ayala, Yin, Shavelson, & Vanides, 2002, p. 3)

Assessments of learning are intended to measure a *construct* (a conceptual model), for example, *learning, knowledge,* or *motivation.* The term, *construct*, in our assessment square corresponds to *cognition* in the assessment triangle (see Figure 1). The construct definition is the heart of the learning assessment.

We prefer the term *construct* because it does not limit assessment to an underlying cognitive model. For example, an alternative approach to positing a cognitive model would be to define a construct (say *knowledge*) in terms of a subject-matter domain (e.g., Newtonian mechanics; see Shavelson & Ruiz-Primo, 1999).

A *conceptual analysis* would expect the *working* construct definition to circumscribe the domain of learning that the construct covers and the kinds of student responses (behavior) that the construct says should be produced by the construct. For example, if the construct were declarative knowledge *(knowing that),* a conceptual analysis would identify a range of tasks in a knowledge domain that could be presented to students (e.g., present a definition of a term) and the kinds of responses that would be expected (e.g., recall or recognize the term). The definition

might also rule out other tasks and responses, and it might posit other constructs that declarative knowledge should and should not be related to. The working construct definition, then, drives the tasks or situations, response demands, and scoring system that comprise a learning assessment (Shavelson & Ruiz-Primo, 1999).

Once the initial conceptual analysis has been completed, the remaining corners of the square involve the assessment, and we search for logical evidence that this construct will be evoked by the assessment tasks and empirical evidence that the construct was, indeed, evoked in a student's behavior. Of course, the logical and empirical analyses might very well lead to a modification of the construct, based on what we learned. Hence, we reiterate, this is an iterative, not linear, process.

An *assessment* is a systematic procedure for eliciting, observing and describing behavior, often with a numeric scale (cf. Cronbach, 1990). The assessment is a physical manifestation of the working construct definition. It is one of many possible manifestations of the construct in the form of a test that could have been produced; we think of an assessment as a "sample" from a universe of possible assessments that are consistent with the construct definition (Shavelson, Baxter, & Gao, 1993; Shavelson & Ruiz-Primo, 1999).

Once an assessment has been developed or selected for use, we analyze, logically, its tasks and response demands to see whether it falls within the construct domain, and whether it is likely to elicit the expected behaviors from a student. The task analysis involves reviewing the task and determining what kinds of thinking the task might evoke in students. This analysis posits cognitive activities that the task might evoke by examining the "opportunities and constraints" that the assessment task provides students to elicit their knowledge and skills (Li, 2001; Ruiz-Primo, Shavelson, et al., 2001). More specifically, we perform the assessment task, hypothesize what a competent student needs to know and be able to do to complete the task, and relate this analysis to the construct underlying the assessment. In this process, we assume a student who is proficient on the task; otherwise this logical analysis generates any number of inappropriate task demands (e.g., guessing, or trial and error; Ayala et al., 2002). The logical analysis links the assessment back to the construct (Figure 2).

Of course the logical analysis is, of necessity, incomplete. We cannot anticipate completely what responses the assessment will elicit from students, even competent

students. However, the logical analysis will point out limitations in the assessment, or the construct definition, that will need to be addressed.

The third corner of the square, *observation*, involves collecting and summarizing students' behavior in response to the assessment task. This empirical analysis focuses not only on observed and perhaps scored task performance, but also on cognitive activities elicited by the task. The analysis provides evidence on a student's cognitive activities that were evoked by the task as well as the student's level of performance. The analysis brings both to bear on the link between the assessment and the construct definition. We ask, did the assessment evoke the intended behaviors? Is there a correspondence between the intended behaviors and the performance scores? Specifically, in the empirical analysis, we examine, as appropriate: (a) students' cognitive activities as they carry out the assessment task using a concurrent "think-aloud" technique, (b) the empirical structure of students' responses as represented by scores (analyzing covariances), (c) the relationship between students' cognitive activities and their performance scores, and (d) the relationship between assessment scores and scores from other tests based on similar and different constructs (Ruiz-Primo, Shavelson, et al., 2001).

In our empirical analysis of observations, we compare (a) the observed structure of students' assessment scores to the expected structure based on the construct definition, (b) the observed students' cognitive activities to the cognitive activities expected in the construct definition, (c) the relation between cognitive activities and performance scores to that expected by the construct definition, and (d) the performance scores to scores on similar and different constructs as expected from the construct definition.

Finally, we put together evidence from the logical and empirical analyses and bring it to bear on the validity of the *interpretations* from an assessment to the construct it is intended to measure. Simply put we ask, "Are the inferences about the construct—a student's learning or domain knowledge—from performance scores warranted?" During the development of an assessment, we iterate through, somewhat informally, the assessment square until we have fine-tuned the assessment. In research and practice where learning is assessed, we formally evaluate the inferences.

All of this is pretty abstract. To make it practical, we apply the assessment square to three examples. The first example is the Population 2 science achievement

test from the Third International Mathematics and Science Study (TIMSS). We defined the construct, *achievement*, as consisting of four types of knowledge—declarative, procedural, schematic, and strategic—and examined the TIMSS test items to see if the test measured achievement as conceived in our definition (Li, 2001; Li & Shavelson, 2001). The second example is an analysis of tests that are used to measure the construct, *knowledge structure*. These tests are concept maps (Ruiz-Primo & Shavelson, 1996); they are based on "… the notion that concept relatedness is an essential property of knowledge and the empirical finding that an important competence in a domain is well-structured knowledge" (p. 101). And the third example is an analysis of performance assessments—tests that are intended to measure especially procedural knowledge (knowing how) but also schematic knowledge (knowing why) and strategic knowledge (knowing when, where, how knowledge applies).

Achievement Assessment

There is an adage that goes, "intelligence is what intelligence tests measure." The same could be said of achievement tests. However, four somewhat diverse lines of research and practice—brain research (e.g., Bransford, Brown, & Cocking, 1999), cognitive research (Bransford, et al.), U.S. science standards (Bybee, 1996), and testing practice (Pellegrino, Chudowsky & Glaser, 2001) have converged and have been synthesized into a heuristic framework for conceptualizing the construct, achievement, as declarative, procedural, schematic and strategic knowledge (e.g., Li, 2001; Li & Shavelson, 2001; Li, Shavelson, & White, 2002; Shavelson & Ruiz-Primo, 1999). In what follows, we focus on achievement in the domain of science; specifically, we focused on the TIMSS Population 2 science achievement test items (Li; Li & Shavelson).

Construct: Types of Knowledge Defining Science Achievement

We defined science achievement as consisting of four types of knowledge (see Figure 3). *Declarative knowledge* is "knowing that"—for example, knowing that force is a push or pull and light is a form of energy. Declarative knowledge includes scientific definitions and facts, mostly in the form of terms, statements, descriptions, or data. For instance, a statement like, "combining two or more materials together forms a mixture" is the scientific definition of "mixture." A scientific fact would be,

for example, “the density of water is 1 gram per milliliter at 4 degrees centigrade and at one atmosphere of pressure.”

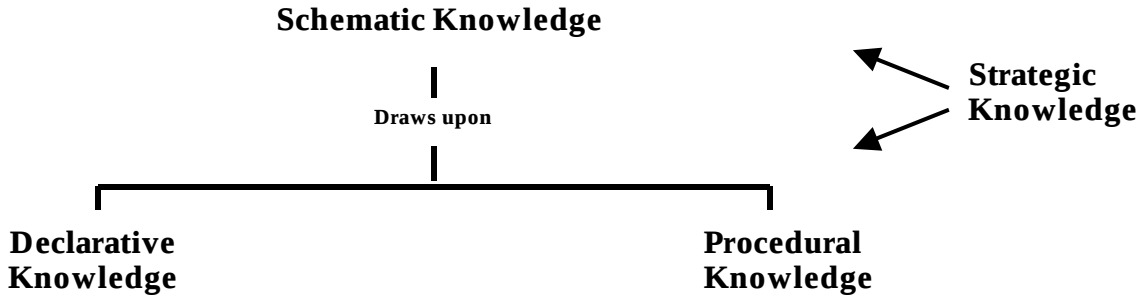


Figure 3. Knowledge framework for science achievement

Procedural knowledge is “knowing how.” For example, knowing how to design a study that manipulates one relevant variable and controls others, or how to measure the density of an object or how to graph the relation between the angle of an incline plane and the force needed to move an object up it. Procedural knowledge includes if-then production rules or a sequence of steps that can be carried out to achieve a sub-goal leading to task completion.

Schematic knowledge is “knowing why.” For example, knowing why Delaware has a change of seasons or knowing why we see different phases of the moon. To know why is to have a scientifically justifiable “model” or “conception” that explains the physical world. Schematic knowledge includes principles, schemes, and mental models. Schematic knowledge can be used to interpret problems, to troubleshoot systems, to explain what happened, and to predict the effect that changes in some concepts will have on other concepts (De Kleer & Brown, 1983; Gentner & Stevens, 1983).

And *strategic knowledge* is “knowing when, where and how” to use certain types of knowledge in a new situation and knowledge of assembling cognitive operations. Strategic knowledge includes domain-specific conditional knowledge and strategies such as planning and problem-solving as well as monitoring progress toward a goal. People use strategic knowledge to recognize the situations where some procedures can be carried out, to examine the features of tasks in order to decide what schematic knowledge can be applied, to set task goals, or to control and monitor cognitive processing.

This simple framework, linked to science-content domains, can be used to specify important aspects of an achievement test. It can also be used to guide the analysis of current tests to figure out what they are measuring. Having defined the construct, achievement, we turn to logical and empirical analyses of item characteristics and their link to the construct.

Logical Analysis of TIMSS Items

In order to identify TIMSS test item characteristics and link them to our multi-knowledge construct of science achievement, we analyzed and coded multiple-choice, short-answer, and performance-assessment test items logically; we focus here on the first two item types (cf. Li, 2001; Li & Shavelson, 2001). The logical analysis assumes, importantly, that a competent student's performance on an item involved a transaction between the cognitive, motivational, and emotional characteristics of the student and the task and response demands of the test item (cf. Cronbach, 2002; Shavelson, et al., in press). We assume that the student, encountering the assessment task, creates a mental problem space that she operates on to complete the task. She brings her aptitudes—cognitive, motivational, and emotional resources—to bear on task completion. As she proceeds along a solution path, the task affords new opportunities, and as those new opportunities arise, new aptitude complexes are evoked. Consequently, our logical analysis examines both task and cognitive/response demands. Of necessity, the logical analysis must be incomplete; many solution paths, some quite unpredictable, arise from one student to another. We restrict the range of cognitive responses by assuming competent students building on cognitive studies of expertise (e.g., Bransford, Brown & Cocking, 1999) where experts create problem spaces that represent the underlying principles embedded in the task and novices attend to surface features in unanticipated ways.

Following from our construct definition of science achievement, we developed an analytic coding system to describe the items and to predict how proficient examinees would interact with the items (e.g., Baxter & Glaser, 1998). The coding system (see Table 1) was divided into four main categories and arrayed in order of importance: a) task demands, b) inferred cognitive demands, c) item openness (e.g., selected vs. constructed response), and d) complexity. For example, a task that asks students to respond by enumerating procedures and actions taken in an investigation involves the use of procedural knowledge. In contrast, a task asking

for a definition of a term affords declarative knowledge. An open task format offers examinees an opportunity to apply schematic or strategic knowledge because the openness allows or forces them to put together theories / models and generate their own procedures and / or explanations. Finally, a task with attractive but misleading distracting cues would be considered more complex than one with few or no distractors.

Table 1

Logical Analysis: Item Coding System (After Table 5.1 From Li, 2001, p. 69)

- **Task Demands:** What does the item ask examinees to do?
  - Terms, symbols, vocabulary, and definitions
  - Statements, descriptions, and facts
  - Procedures, steps, and actions
  - Algorithms and equations, figures and diagrams, and tables
  - Models, relationships, theories, explanations, and principles
- **Cognitive Demands:** Inferred cognitive processes examinee may bring to task and how they use knowledge and reasoning to respond to the item.
  - Visualization
  - Mathematical calculation
  - Mechanical operation (e.g., draw diagram, balance chemical equation)
  - Perform experiment
  - Recall information / fact
  - Reason and interpret with models and principles
  - Describe and record information and outcomes
  - Select and use strategies
  - Plan and monitor
  - Guess or eliminate wrong options
  - Reason with common sense
- **Item Openness:** Degree of freedom student has in shaping response to item.
  - Hands-on vs. paper-and-pencil
  - Read options and choose vs. generate responses on own (selected vs. constructed-response format)
  - Require only information found in task vs. steps / information can be learned from the task
  - Require one vs. multiple correct solutions / approaches
  - Follow instructions or steps
- **Complexity:** Familiarity, relevance, reading difficulty, common experience of the item
  - Textbook-type task vs. ill-structured task (provides or contains new situation / information)
  - Inclusion of irrelevant background information
  - Long, reading demanding descriptions and complicated vocabulary
  - Answers contradict every experience / belief

To exemplify the application of the logical analysis, we analyze three items with the coding scheme. The results of the analysis for the TIMSS Booklet 8 items are presented in Table 2. TIMSS item P6 draws upon declarative knowledge, asking for a definition. Specifically, it asks for the digestive substance in the mouth and its function.

Table 2

Classifications of TIMSS Booklet 8 Items (Li, 2001, p. 75)

Item		Classification [a]				
label	Description	DE	PR	SC	ST	NA
A7	Organ not in abdomen	P				
A8	Stored energy in two springs			P		
A9	Fanning a wood fire	P				
A10	Seeing person in a dark room					P
A11	Overgrazing by livestock	P				
A12	Changes in river shape / speed		P	P		
B1	Layers of earth	P				
B2	Energy released from car engine	P		P		
B3	Greatest density from mass / volume table		P			
B4	Pulse / breathing rate after exercise	P				
B5	Elevation diagram of wind / temperature		P			
B6	Color reflecting most light					P
P1	Distance versus time graph		P			
P2	Flashlight shining on wall			P		
P3	Life on planet Athena			P		
P4	Animal hibernation	P				
P5	Heating water with balloon			P		
P6	Digestive substance in the mouth	P				
P7	Replication of measurement		P			
Q11	Daylight and darkness			P		
Q12	Jim and Sandy's flashlights					P
Q13	Lid on jar	P		P		
Q14	Heated iron and sulfur	P				
Q15	Chemical change	P				
Q16	Light from star					P
Q17	Advantage of two eyes	P				
Q18	Melting ice cubes	P		P		
R1	Light striking mirror		P			
R2	Why does shirt look blue?			P		
R3	New species in area	P				
R4	Ozone layer	P				
R5	CO2 fire extinguisher	P				

[a]DE = declarative knowledge, PR = procedural knowledge, SC = schematic knowledge, and ST = strategic knowledge. NA indicates the items could not be classified into any certain knowledge-type(s) because the items were too general or too ambiguous. These abbreviations are used hereafter.

What Digestive Substance Is Found in the Mouth?

What Does It Do?

Two of the three item characteristics lead us to conclude that it taps declarative knowledge rather than schematic or procedural knowledge. First, the response is expected to be in the form of terms, vocabulary (e.g., saliva), and factual statements. The item asked a very specific content question (i.e., a specific fact), leaving little room for students to provide relations between concepts or to apply principles / models. Second, the cognition involved is likely to be recall. Note that the item is not only similar to school-type problems but similar to the way texts are written and students are taught. It makes sense that when students answer the question they may recall exactly what they had been taught. Of course, some students might learn this information from everyday experience instead of formal schooling. Note that the cognitive state in which the knowledge was learned barely differs from the cognitive state in which it is used in the testing situation. Moving from the former to the latter state involves little transformation or re-construction of knowledge. That is, the answer to Item P6 is stored in a student's memory.

The cognitive process involved in answering the item is to directly retrieve information or do a minimum of scientific reasoning to organize the relevant information. However, the item openness and complexity have little relevance to our classifying the item as tapping the declarative knowledge. For instance, the response format (i.e., free response) is slightly more cognitively demanding than multiple choice. Instead of only recognizing correct answers from options given, students were invited to recall and organize their responses. Such a characteristic made the item slightly challenging to students but did not fundamentally change the type of knowledge students might use. Therefore weighing the four item characteristics, we classify Item P6 as a declarative-knowledge item.

Very different from Item P6, our logical analysis showed that Item Q11 tested students' schematic knowledge. This multiple-choice item asked students to recognize the explanation for why we have daylight and darkness on Earth (see Figure 4).

Q11. Which statement explains why daylight and darkness occur on Earth?

A. The Earth rotates on its axis.
B. The Sun rotates on its axis.
C. The Earth's axis is tilted.
D. The Earth revolves around the Sun.

Figure 4. TIMSS Item Q11 – Example of a schematic-knowledge item

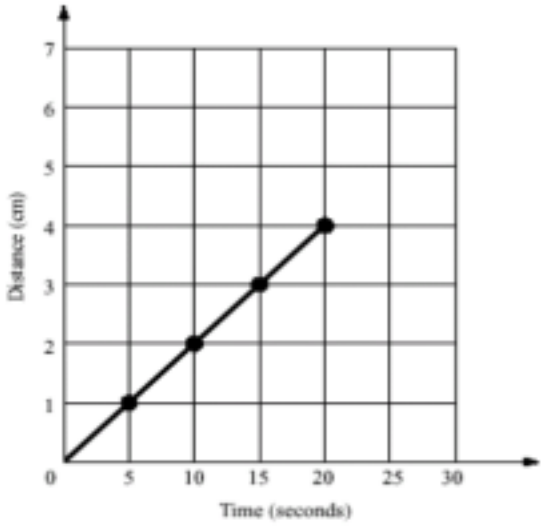
Three of the four item characteristics support the conclusion that the item affords the use of schematic knowledge. First, the item is intended to assess 14 year olds' knowledge about the cause of daylight and darkness because it invites students to select conceptual explanations by using the phrase "explain why" in the stem. The knowledge that students use to solve this item can be models or theories. Some students might use visualization or diagrams to represent and infer the process and consequence of the movement. Second, the dominant cognitive process is reasoning with theories, perhaps with the aid of visual representations and/or mentally spatial manipulation of objects. It is very likely that students have to figure out the correct answer from four explanative models by reasoning the key issues about the Earth and Sun's movement (e.g., rotating on axis and revolving). Third, the item is not restricted in terms of openness. That is, the item does not force students to read options first in order to solve the task. The necessary information for students to start and generate their own answers is provided in the descriptions in the stem. Therefore, it leaves room for students' thinking process. Of course, some students might read the options before reasoning. However, one critical element of task completion is to bring additional underlying principles and knowledge related to the Earth and Sun's movement instead of merely reading the options. For example, while reading the options, students might also think about and figure out that sentence C partially explains the season and sentence D partially explains the year. Finally, the item does not involve heavy reading or irrelevant information. This low complexity strengthens the posited link to schematic knowledge by reducing construct irrelevant variance.

Of course, if a student has encountered this question in class or at home repeatedly, she very well might recall the answer from memory. In that case, we would consider the item to tap declarative knowledge. The answer is recalled as a factual statement or model template. Until we examine students' cognitive activities

while answering the question, the item-characteristic analysis is incomplete, of necessity.

For our last example, P1 was coded as a candidate item to tap procedural knowledge (Figure 5). It provides students with a graph showing an ant's moving speed within 20 seconds and asks the distance that the ant will travel at 30 seconds.

The graph below shows the progress made by an ant moving along a straight line.



If the ant keeps moving at the same speed, how far will it have traveled at the end of 30 seconds?

A 5cm
B 6cm
C 20cm
D 30cm

Figure 5. TIMSS Item P1 – Example of a procedural-knowledge item.

Two of the four characteristics tend to engage examinees using procedural knowledge. First of all, the item requires students to interpret the diagram to find the distance and/or to apply an algorithm to calculate the speed. Either piece of knowledge falls into the category of procedural knowledge I defined. Second, the cognitive process students probably engage in is either applying an algorithm for speed by dividing distance with time or by extending the line in the graph to 30 seconds and simply reading the distance on the vertical axis. However, the item's lack of openness to some extent limits examinees applying procedural knowledge.

Although the multiple-choice format may allow students to generate their own responses before reading the options, very likely students can arrive at the right answer by working backwards from the options. Finally, the complexity codes were not informative for modifying the link since they were not at the extremes of the continuum.

Of course, questions about the consistency (reliability, agreement) arise when coding assessment tasks. Would a second coder see the item the same way as the first? In our studies, we have two coders code all items. In the TIMSS analysis, we found inter-coder agreement to be, on average, 80 percent (Li, 2001). Where disagreements arise, coders (and a third person if needed) reach consensus on the codes for the contested items.

Empirical Analyses of TIMSS Items

The logical analysis requires the analyst to "psychologize" about the nature of the problem space that a student constructs when confronted with an assessment task. That "psychologizing" can never be complete. What students actually think is always surprising. For this reason, the analysis of an assessment task is incomplete without empirical studies. Here we describe two types of empirical studies, one focused on the cognitive activities evoked by assessment tasks and the other on the covariance structure of the tasks / items on an assessment.

Cognitive Analysis. The most important of the empirical studies for both assessment development and validation is an analysis of the cognitive activities that a task evokes. Do these cognitive activities comport with the expectations of the construct definition?

Several methods can be used to examine cognitive processes evoked by assessment tasks. Perhaps the method with the strongest theoretical and empirical justification is the "think-aloud" method (Ericsson & Simon, 1993). Individual students are asked to think (verbalize) aloud as they carry out a task, revealing cognitive activities in working memory. Their verbalizations are recorded and a transcript produced. Analysis of a think-aloud[1] protocol provides evidence bearing on a student's thinking processes (Ericsson & Simon; for education research

[1] The terms, *think aloud* and *technique* were first introduced by Ericsson and Simon (1993).

applications, see Ayala, Yin, Shavelson & Vanides, 2002; Baxter & Glaser, 1998; Hamilton, Nussbaum, & Snow, 1997; Ruiz-Primo, Shavelson, et al., 2001).[2]

Li (2001) examined the cognitive activities evoked by TIMSS items using the think-aloud method (see Figure 6). She compared think-aloud protocols from "competent" graduate students drawn from either physics or biochemistry across the items categorized as tapping different types of knowledge. Would the distinction among TIMSS items as to the knowledge types they tapped be reflected in examinees' cognitive activities? She augmented the think-aloud protocols with retrospective interviews and notes she took while observing participants' performance.

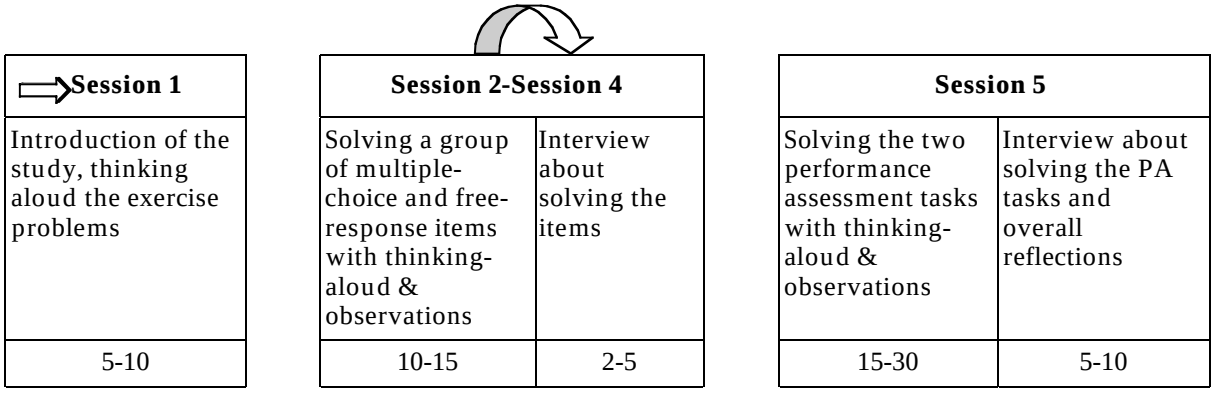


Figure 6. The sequence of events in the protocol study (after Li, 2001, Figure 7.3, p. 123)

Participants' think-aloud protocols were analyzed by looking for indicators of participants applying different types of knowledge, which could be related with some certainty to the possible characteristics—affordances and constraints—of the test items. The analysis took the following steps:

- Segment participants' verbalizations into response entries, each of which contained their responses to only one item or task no matter how many statements or types of knowledge it consisted of. Response entries were considered as the unit-of-analysis. Within each entry, the protocol was separated by linguistic pauses at the end of phrases and sentences, reflecting the natural pauses that the participants took.

- Code participants' response entries with a coding scheme that captured evidence of the their use of the four types of knowledge. Each

[2] Noticeably missing from these methods are interviews about thinking processes and focus groups. These techniques are helpful in analyzing possible cognitive demands. Nevertheless, they are likely to elicit respondents' theories about their cognitive activities and the activities themselves.

knowledge-type category in the scheme has subcategories, for example, strategic knowledge included four subcategories: framing a problem, planning, monitoring, and testwiseness (see Table 3). The knowledge-category codes were designed to capture what types of knowledge the participants applied instead of the levels of understanding. Those phrases, not relevant to task completion or directly indicating content understanding, such as "all right," or "I am reading the item A1", were excluded from the analysis.

Table 3

Sub-Category Codes for Strategic Knowledge (from Li, 2001, Table 7.4, p. 127)

Subtype	Description	Examples
Framing a problem	Statement or questions recognizing / labeling the *features* of tasks, procedures, and projects or *making sense* of tasks	It says "digestive," so digestive means it does something changing the food, but I don't know what the chemistry is. Okay, so three and four are theoretical questions, which I will get to after the experimental one. I think [it is] a poorly worded question.
Planning	Statements or questions about what will or should happen next (steps or actions)	All right, I will make a chart, I know the fastest way to do it is to make a chart, and I usually do so and figure it out. I read all the questions because there may be a better way to do the experiment. Either I can do all of them together one time or maybe it's not the best way to do the first one first.
	Statements or questions related a plan to a condition or specifies the basis for choosing between alternative plans (hypothesizing or conjecturing)	So let's try B and C. And I expect that this one's not going to be very bright at all. Now, just for the sake of completeness, although I'm pretty much convinced by now, I'll check D and B about the same as A and B.
Monitoring	Statement or questions noticing / regulating / checking the progress or lack of progress (an ongoing task)	So, I try to remember the differences between all these words. Something I learned a long time ago. Also from experience,
	Statement or questions concerning the conclusions at the end of the task (a completed or aborted task)	But it's one of many cases where one can guess what the test writer had in mind. Did not feel compelled to check all other combinations because I know enough about how batteries work to know that I'm reasonably sure that the result I saw could only be explained by A and D being good and B and C being worn out.
Test wiseness	Statement or questions as an educated guessing, completely guessing, or eliminating options	So it seems like a reasonable guess.

Table 4 contains an example of the schematic knowledge category sub-category coding. Notice that the participant used two models to work problem P2. One model explained the relation between amount of light, distance, and size of circles on the wall, whereas another one explained that light could heat the air and could be absorbed by the air.

Table 4

Coding of a Participant's Think Aloud (From Li, 2001, Table 7.4, p. 128)

Item Description	Code	Category	Protocol Response Entry
FR P2: "Flashlight close to a wall produces a small circle of light compared to the circle it makes when the flashlight is far from the wall. Does more light reach the wall when the flashlight is further away?"	3-1	Schematic theoretical model	Okay, so the flashlight close to the wall makes a small circle of light because the cone of the light coming off of the light bulb doesn't have time to expand. And so there's a constant amount of light coming out of the flashlight and as it expands out at a constant angle of light, the flashlight is, the wall is close here then you're going to get a circle of light the size and if the wall is farther away you're going to get a bigger circle of light, but it's the same amount of light. And, in fact, less light reaches the wall when the flashlight is farther away because it's scattered by the light absorbed by the air and is heating things up. And so, no, not, well it doesn't reach it.

- Examine the consistency of coding. Would two independent coders using the rules provided, agree on their classification of a segmented protocol? After training and calibrating on pilot data, the final agreement between two coders was 85%.

- Bring coded protocols data to bear on how participants employed different types of knowledge to represent and solve problems. The frequency with which different types of knowledge that participants used were aggregated for each TIMSS item to examine whether the predicted knowledge-type was used more frequently that those not predicted.

Assuming that different types of test items provided different affordances and constraints on participants' task completion, the cognitive activities participants displayed across the items should also vary. Based on the knowledge-type construct of science achievement, participants' use of knowledge inferred from the protocols (cognitive analysis) and the knowledge-types demanded by test items (logical analysis) should be congruent. Table 5 presents the occurrences of knowledge that

the participants utilized to solve the test items organized by two dimensions—knowledge-types inferred from protocols and knowledge-types into which the test items were pre-classified. For example, nine declarative-knowledge items were administered to the participants, and 54 responses (9r6 = 54) were collected and analyzed. Among these 54 responses, participants used declarative knowledge in 48 entries, employed schematic knowledge in 10 entries, and applied strategic knowledge in one entry. Since 54 does not equal the total of 48, 10, and 1, the numbers imply that at least in some protocol entries, the participant applied more than one type of knowledge. A quick look at Table 5 reveals that of the 85% of the test items,[3] participants' use of knowledge was consistent with the item-knowledge links predicted from the logical analysis, and expectation from the construct definition.

Table 5

Number of Entries Coded by Knowledge-Type (From Li, 2001, Table 7.6, p. 130)

Type of knowledge used	Preclassified knowledge type			
	Declarative (n = 9)	Procedural (n = 10)	Schematic (n = 9)	Strategic (n = 2)
Declarative	48	8	11	0
Procedural	0	54	7	9
Schematic	9	16	41	0
Strategic	2	12	2	10

Covariance analysis. With the consistency between the logical and cognitive analyses, we selected TIMSS items as representing three knowledge types: declarative, procedural, and schematic as no TIMSS item was built to measure strategic knowledge directly. The goal was now to see if the logical and cognitive links of items to knowledge types could be found in the covariances among item scores. The structure of the item covariances was examined with confirmatory factor analysis (see Figure 7). The model fit was extremely good (for details, see Li & Shavelson, 2001).

[3] The proportion was calculated using the number of response entries, for example, 54 for the declarative-knowledge items to show whether and to what extent a pre-classified type of knowledge occurred when experts responded to the items.

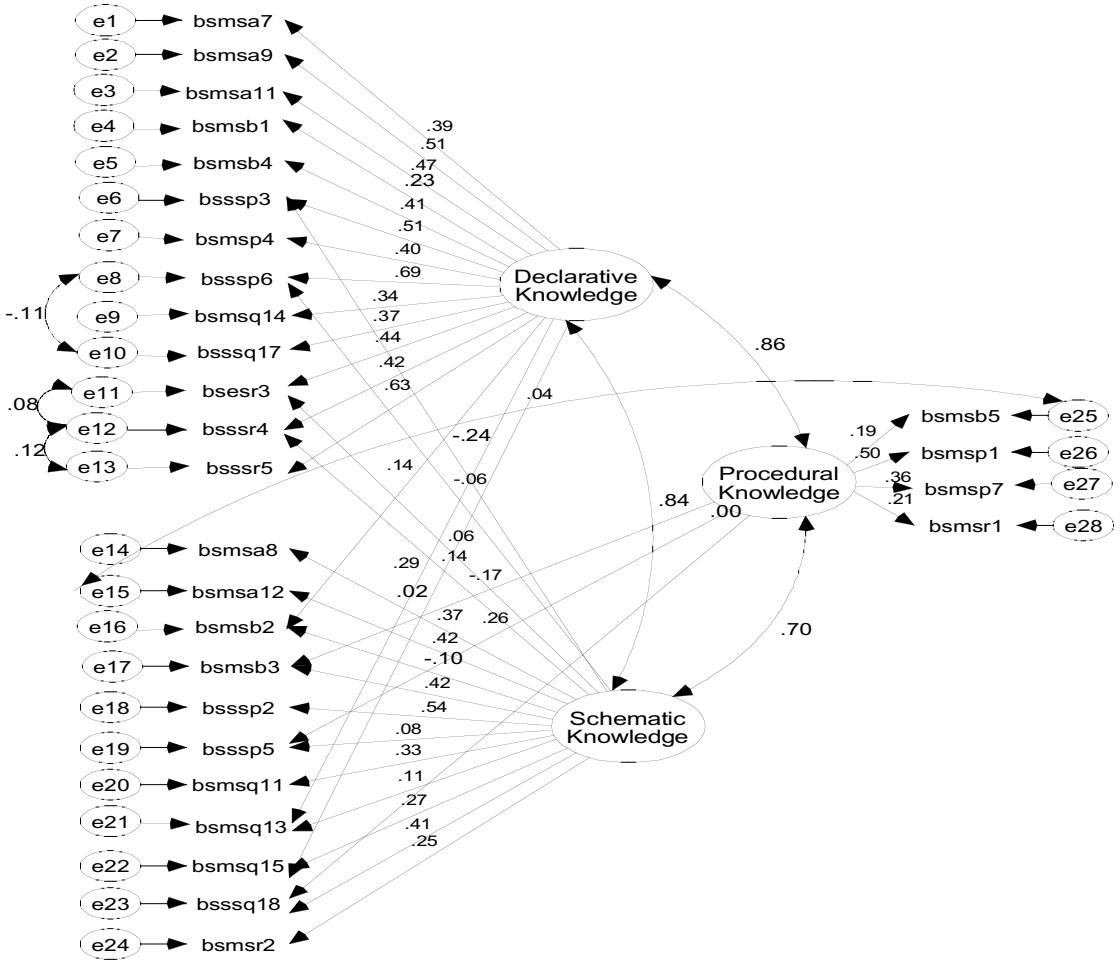


Figure 7. confirmatory factor analysis of TIMSS Booklet 8 science achievement items (from Li & Shavelson, 2001, Figure 4, p. 22).

Three main findings are apparent. First, the factor analysis predictions from the logical and cognitive analyses were confirmed. Thirteen items were pre-classified as declarative-knowledge, 11 as schematic-knowledge, and 4 as procedural-knowledge. The results supported our predictions with the declarative and procedural knowledge. However, only 8 of the 11 pre-classified items showed their primary loadings on the schematic-knowledge factor, leaving 2 items to the declarative-knowledge and 1 item to the procedural-knowledge factor. Most item regressions were significant and generally high, which is also true for at least one loading with those double-relation items. The fit and the regression estimates supported the construct definition and the TIMSS test with items selected for consistency in our logical and cognitive analyses.

The *declarative knowledge* factor included fifteen items, regardless of their content or format. Those items required students to deal with scientific vocabulary, provide definitions, or recall facts or information. Students could obtain the knowledge either from formal schooling or everyday experience. Eight items were heavily regressed on the *schematic knowledge* factor. Those items required students to apply scientific rules, theories, or principles in a context of providing explanations or predictions. Finally, five items showed primary regressions on the factor, labeled *procedural knowledge*. Those items involved use of scientific actions, algorithms, steps, and procedures needed to solve problems. Note that the loadings of procedural-knowledge items were relatively lower compared to those of the declarative-knowledge or schematic-knowledge items. The reasons, we would hypothesize, are: a) number of items—only a few items included in this study were designed to tap procedural knowledge, and b) quality of items—multiple-choice items may not be the best method to adequately tap procedural knowledge (Baxter & Shavelson, 1994).

Second, the relation between each pair of the three latent knowledge factors was strong, about .80 on average. Further, the relation between procedural and schematic knowledge was slightly weaker than the procedural-declarative and schematic-declarative ones. Our explanation is that those procedural knowledge items in multiple-choice format mainly asked for routine procedures without requiring examinees to use principles or theories. However, further analyses are necessary to support this interpretation.

Third, the pattern of double loadings and the correlations between error variances were partly consistent with what we expected. Most of the built relations were in the right direction (i.e., positive loadings) and were statistically significant. Adding an extra relation may weaken a relation. The four negative loadings can be considered evidence to suggest that the items tended more likely to tap one knowledge type than another type. For example, item P3's double loading confirmed our exploratory analysis that it taps primarily declarative-knowledge, whereas Item Q15 taps schematic knowledge.

We also compared the triplet knowledge model with different models. We tested alternative models based on different theoretical claims, such as one factor as general science achievement, two factors as test format (i.e., multiple-choice and free-response), and four factors as different domains (i.e., life science, earth science, chemistry, and physics). The three-factor model better reflected the underlying

pattern among item scores than a total-score (general science achievement), science-sub-domain or item-format model. The three-factor model corresponding to knowledge types provided the best fit to the item scores, consistent with our definition of the science achievement construct.

3.3 Summation

The evidence from logical, cognitive, and covariance analyses converges to support the interpretation of the TIMSS science achievement items as measuring three underlying types of knowledge consistent with our construct definition. Any one analysis would increase confidence in the interpretation link in the assessment square model. However, the convergence of different methods substantially increases the justifiability of the link from construct definition to assessment to observation to inference back to the construct.

We now turn to another example, stepping through the assessment square, but this time focusing on new pieces of analysis while only briefly mentioning methods already described in the TIMSS analysis. Our intent is to provide a vision of the various types of analyses that can be brought to bear on the quality of new learning assessments.

Concept Map Assessment

Concept maps are labeled directed graphs (e.g., Figure 8) that are interpreted as measuring declarative-knowledge structure (Ruiz-Primo & Shavelson, 1996). Concept maps come in many different varieties, yet they all claim to measure knowledge structure (Ruiz-Primo & Shavelson). We have tested such claims in a number of studies (e.g., Ruiz-Primo, Schultz, Li, & Shavelson, 2001; Vanides, Yin, Ayala & Shavelson, 2002), most notably for our purpose here, a study examining the "cognitive validity" of concept-map-score interpretations (Ruiz-Primo, Shavelson, et al., 2001). In the study, we compared knowledge-structure interpretations of high-school chemistry students' scores on two types of concept maps, a "construct-a-map" and two versions of a "fill-in" map. From a theoretical perspective (see below) we expected the construct-a-map to provide a more accurate representation of knowledge structure than the fill-in maps. But the fill-in maps are preferred on practical grounds; they are easily, quickly, and inexpensively scored.

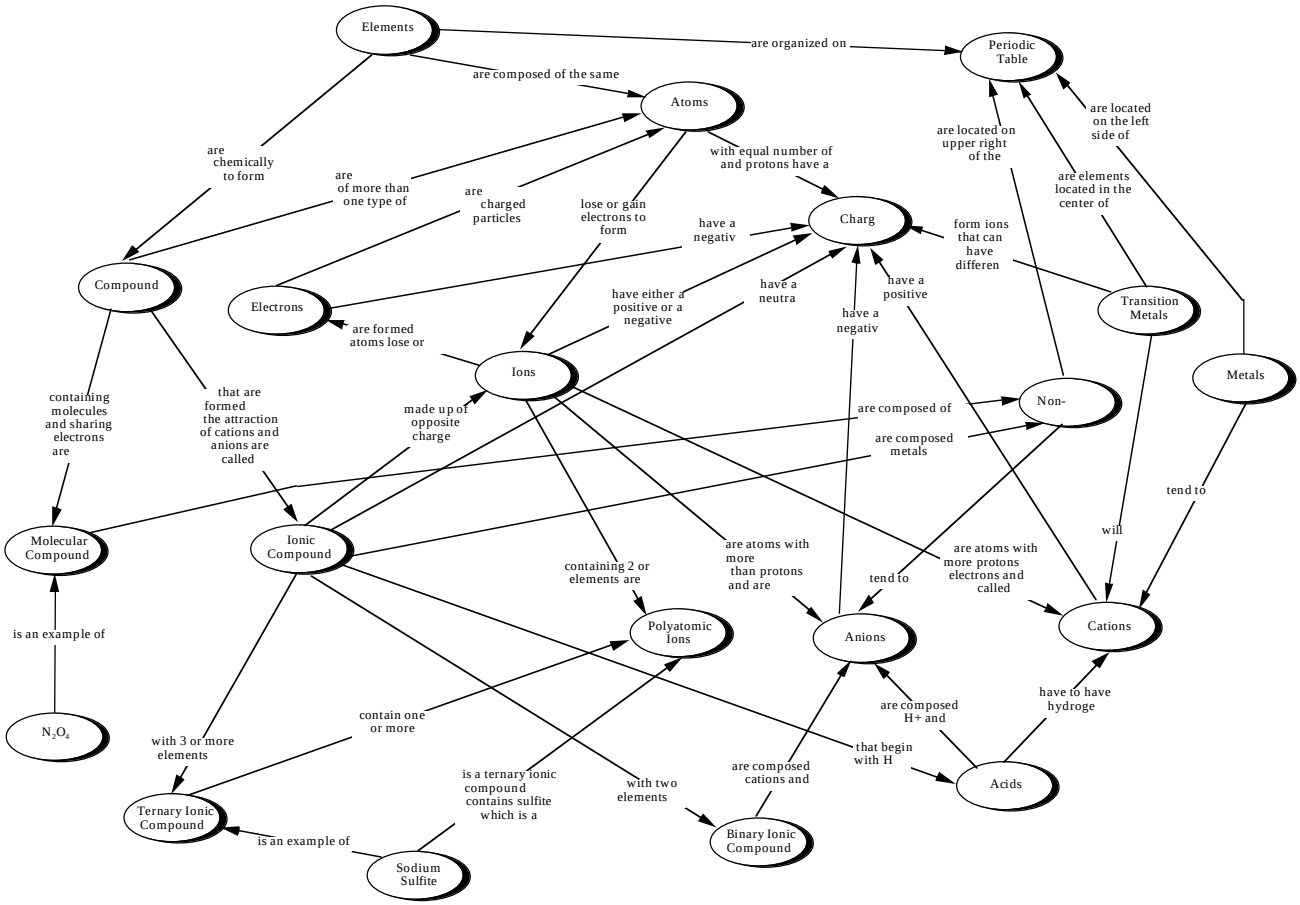


Figure 8. Criterion concept map (from Ruiz-Primo et al., 2001, Figure 3, p. 106)

We (Ruiz-Primo, Shavelson, et al., 2001) set forth a framework for examining the "cognitive validity" of the alternative mapping techniques (see Figure 9), that is, to examine the validity of interpretations of the map scores as tapping the construct, *structural knowledge*. We sought to bring empirical evidence to bear on the: a) cognitive activities evoked by each assessment, b) relationship between the cognitive activities and performance scores, c) impact of variation in assessment task on cognitive activities, and d) correlations between assessment measuring similar and different constructs. As we have dealt extensively with construct definition, logical analysis, and think-aloud for addressing much of the assessment square already, we touch on these methods briefly and focus on novel applications as suggested in Figure 9.

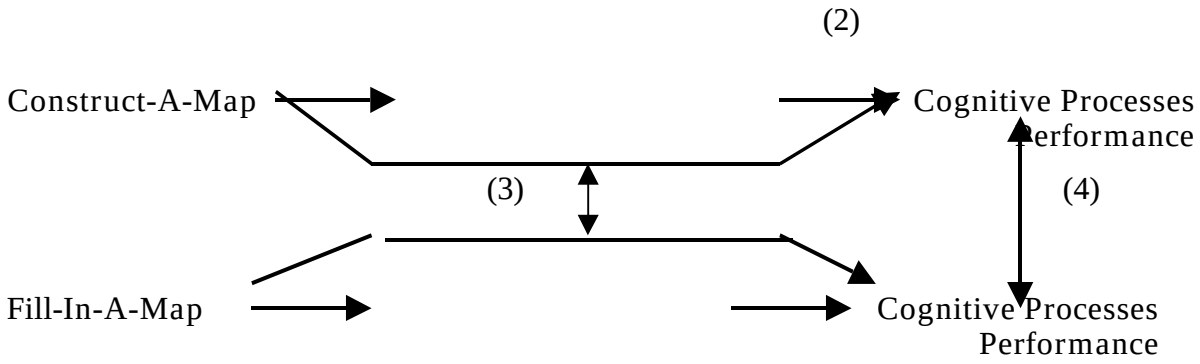


Figure 9. An Approach to Empirically Evaluating Cognitive Interpretations of Concept Maps (adapted from Ruiz-Primo, Shavelson, et al., 2001, Figure 2, p. 103)

Construct Definition: Knowledge Structure

The theory guiding concept mapping posits a long-term memory organized as an associative network. We call the construct to be measured "connected knowledge" or "knowledge structure." Concept maps follow from this model of memory and provide one possible approach for tapping into structural aspects of this network. By having "mappers" connect key concepts (terms) with labeled lines that explain the relation between concept pairs, they purport to tap connected knowledge.

Logical Analysis

Here we examine, logically, the credibility of the link between two concept-mapping techniques and the interpretative claim that each measures knowledge structure. We do so by examining their task and response demands. Our analysis posits the nature of the cognitive activities that the maps will evoke by examining their affordances and constraints.

One such important affordance or constraint, especially for this comparison, is the degree of mapping-task directedness inherent in the techniques (Ruiz-Primo & Shavelson, 1996). At one end of the directedness continuum, the assessor provides the concepts terms, linking lines, linking explanations and structure on a highly directed map ("process constrained"; Baxter & Glaser, 1998). At the other end of the continuum, students choose the concepts in a domain, create their own linking lines, linking explanations, and determine structural relations among concepts ("process open"; Baxter & Glaser, 1998). From this logical analysis of the concept-mapping tasks presented by the construct-a-map and fill-in-a-map techniques, we concluded the former is considerably more "process open" than the latter. Moreover, the

construct-a-map seems to fit our definition of the nature of knowledge structure more closely than the fill-in map.

This logical analysis conjectured, but did not necessarily delimit, the cognitive activities that a student might employ in response to one or the other mapping technique (systematic search, trial-and-error). To see if the conjecture was justifiable, we needed to collect data on the cognitive activities evoked by the two mapping techniques and students' observed performance.

Empirical Studies

To address the validity of the knowledge-structure interpretation of the two mapping techniques in the domain, atoms and molecules (see Figure 8), Ruiz-Primo, Schultz, et al. (2001) followed the model shown in Figure 9. They collected data from 152 high-school chemistry students in the performance portion of the research and talk-aloud data from two experts (chemistry teachers) and six novices (three high- and three low-performing students). The *expert-novice design* for talk alouds was used because research on expertise had consistently demonstrated that experts' knowledge is far more highly structured than novices'. "We reasoned … that competent examinees … would a) provide coherent, content-based explanations rather than descriptions of superficial features or single statements of fact; b) generate a plan for solution; c) implement solution strategies that reflect relevant goals and subgoals; and d) monitor their actions and flexibly adjust their approach" (Ruiz-Primo, Schultz, et al., p. 104). Moreover, consistent with cognitive activity differences, experts would be expected to score higher on their concept maps than novices, and they would systematically vary by mapping technique with greater differences between experts and novices observed on the less-directed construct-a-map.

Cognitive Analysis. Ruiz-Primo, Schultz, et al. (2001) used talk alouds to capture experts' and novices' cognitive activities while concept mapping with the different techniques. They used the same concurrent talk aloud technique that Li did with two important differences. First, they distinguished between micro- and macrolevels of the protocol analysis. At the microlevel, they coded *explanations* (e.g., "N_2O_2 is a molecular compound because they are both nonmetals"; Ruiz-Primo et al., 2001, p. 115), *monitoring* ("I can't remember exactly what this is"; p. 115), *conceptual errors* reflecting misconceptions ("molecules are atoms"), and *inapplicable events* (e.g., reading instructions). At the macrolevel, they coded *planning*

(verbalizations at the start of each protocol; e.g., “I will read the concepts and select the most important concept to put … in the center”; Ruiz-Primo, Schultz, et al., p. 116) and *strategy* (from entire protocol) for working through the mapping exercise. Second, their unit of analysis differed somewhat from Li’s (2001). Like Li, at the macrolevel, they used the entire verbal protocol evoked by a particular mapping technique. However, for their microlevel analysis, they segmented verbal protocols into fine-grained response entries (as small as phrases) for their microlevel analysis.

On average, inter-coder reliability for the microlevel analysis was high across the different mapping techniques (i.e., construct-a-map, fill-in-the-nodes, and fill-in-the-lines). Both percent agreement and agreement adjusted for chance agreement (kappa) were reported at the macrolevel with a range of 86% to 100% agreement and 71% to 100% adjusted.

Microlevel results. If the different mapping techniques imposed different affordances and constraints in tapping knowledge structure, these differences should appear in the think-aloud data (points 1 and 3 in Figure 9). Ruiz-Primo, Schultz, et al.’s (2001) predictions about differences from prior research and the logical analysis were supported by the data with one exception, monitoring (see Table 6). The construct-a-map technique led to a greater percent of explanations and revealed a greater percentage of conceptual errors than did the two fill-in maps. With respect to conceptual errors, “we interpreted this … as indicating that the low-directed technique provided sufficient response latitude to allow respondents to more accurately show their conceptual understanding” (Ruiz-Primo, Schultz, et al., p. 122). Contrary to expectation, the fill-in technique led to much greater monitoring than did the construct-a-map technique. “A possible explanation … is that because respondents made more decisions on their own in the construct-a-map technique, those decisions might be considered as final answers…. In the fill-in-the-map techniques, students were more aware of the accuracy of their responses … leading them to monitor themselves more frequently” (Ruiz-Primo, Schultz, et al., p. 120).

Table 6
Comparison of microlevel cognitive activities (percents) across mapping techniques (After Ruiz-Primo, Schultz, et al., Table 6, 2001, p. 120)

		Conceptual		No-Code	
Mapping Technique	*n*	Explanation	Monitoring	Errors	Applicable
Construct-a-map	8	39.08	28.22	9.78	22.91
Fill-in-the-nodes	8	4.64	40.93	0.16	54.27
Fill-in-the-lines	8	2.84	35.22	0.14	61.80

Macrolevel results. We focus here on strategies because Ruiz-Primo, Shavelson, et al. (2001) presented not frequencies but a flow diagram to reflect a *sequence* in the talk-aloud protocol. We suspect this representation might prove useful to others analyzing talk-aloud protocols. Their intent was to capture how a student started the concept-mapping task, continued or advanced in the task, monitored her performance, and completed the task. In Figure 10, strategies are numbered by sequence and by descriptive code. For example, the high-proficiency student began with the most general concept (1.1), proceeded by selecting another concept to relate to it (2.1), and monitored performance by reading the complete proposition (node-link-node, 3.1), and so on. When a single strategy was used repeatedly to advance the task, a looping arrow was used. Teachers and high-proficient students used efficient, intentional strategies such as that described in the figure. Low-proficient students used primarily trial and error.

Statistical analyses. The question arises as to whether there is a link between cognitive activities and performance on the concept map (see Figure 9, point 2). Ruiz-Primo, Shavelson et al. (2001) found the same pattern at the microlevel with the teachers and students as reported for all 152 participants (Ruiz-Primo, Schultz, et al.) but the pattern (more explanations, less monitoring) was more accentuated for teachers' and high-ability students than for low-ability students. Ruiz-Primo, Shavelson, et al., concluded that the "construct-a-map technique better tapped into differences in respondent's cognitive activities according to their level of competence" (p. 124). At the macrolevel, planning and strategies (in graphical representations of participants' sequence of actions) were more related to performance level than to mapping technique.

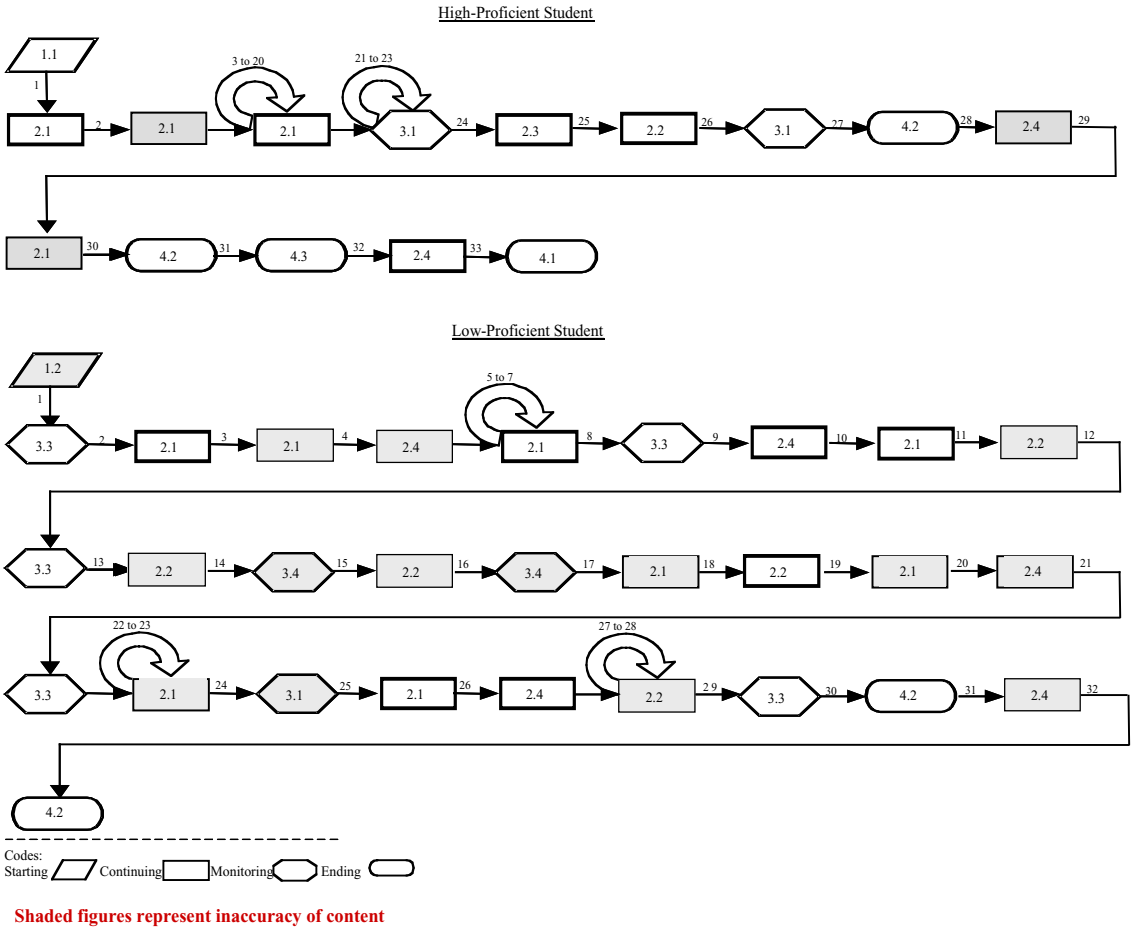


Figure 10. Sequence of cognitive strategies employed by high- and low-proficient students carrying out the construct-a-MAP TASK.

Another way to look at the link between cognitive activities and performance is to correlate cognitive activity percentages with concept-map performance scores. The main drawback to doing this kind of analysis is that, typically, talk-alouds are done with small samples. With this caveat, Ruiz-Primo, Shavelson, et al. (2001) found percent of explanations to be positively correlated with concept-map scores for the construct-a-map technique (.33) but much less so for the nodes (.12) and lines (.01) techniques. In contrast, the correlation between monitoring and performance was negative with the lines technique correlating at the greatest magnitude (-.83) followed by nodes (-.31) and construct-a-map (-.29). A -.27 correlation was found between conceptual errors and scores on the construct-a-map; similar correlations

could not be calculated for the nodes and lines fill-in maps because so few errors cropped up in the talk-aloud protocols.

Finally, scores on the two types of mapping techniques were compared (point 4, Figure 9) using as criteria the characteristics of strictly parallel tests (equal means, variances and covariances). The techniques differed on all three criteria based on data from the original study (N = 152) and from the talk-aloud study (N = 8). Most notably, participants' mean score with the fill-in technique was near maximum possible score (10.67 / 12 = .89); in contrast, the construct-a-map score was considerably lower (.63 of maximum).

Summation

Evidence from the logical, cognitive, and statistical analyses converged and supported knowledge-structure interpretation of construct-a-map scores; not so the fill-in scores by comparison. Moreover, the cognitive validity framework in Figure 9 proved particularly helpful in making clear the link from observation to inference back to the construct of interest: *knowledge structure*.

Concluding Comments

In this paper we described a framework for evaluating new learning assessments that evolved out of our research on alternative types of science assessments. That framework, the assessment square (see Figure 2), linked together the: (a) *construct* to be measured (e.g., knowledge structure), (b) the *assessment* used to measure it (task, response format, scoring system), (c) the *observed* behavior (qualitative and quantitative means for collecting and summarizing responses); and (d) the *interpretations* made of the observations (i.e., the justifiability of the inferences drawn from the observed behavior to the construct).

Embedded in the assessment square are a set of analyses, including a *conceptual analysis* of the construct to be measured, a *logical analysis* of the assessment instrument, a *cognitive analysis* of the thinking evoked by the assessment, and *quantitative* and *qualitative analyses* of the observations that provide justification for the proposed *interpretation* of the assessment.

We hope that the framework and proposed analyses prove as useful to others as they have to us in analyzing the quality of new learning assessments.

References

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

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 7(3), 37-45.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. International Journal of Educational Research, 21, 279-298.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.) (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Bybee, R. W. (1996). The contemporary reform of science education. In J. Rhoton & P. Bowers (Eds.), *Issues in science education* (pp. 1-14). Arlington, VA: National Science Teachers Association.

Cronbach, L. J. (1990). *Essentials of psychological testing 5th Edition*. New York: HarperCollns.

Cronbach, L. J. (Ed.) (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, New Jersey: Erlbaum.

De Kleer, J., & Brown, J. S. (1983). Assumptions and ambiguities in mechanistic mental models. In D. Gentner & A.L. Stevens (Eds.) *Mental models* (pp. 155-190). Hillsdale, NJ: Erlbaum.

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

Gentner, D., & Stevens, A. L. (Eds.) (1983). *Mental models.* Hillsdale, NJ: Erlbaum.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10,* 181-200.

Li, M. (2001). *A framework for science achievement and its link to test items.* Unpublished doctoral dissertation, Stanford University.

Li, M., & Shavelson, R. J. (2001). (April 12, 2001). *Examining the links between science achievement and assessment*. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Li, M., Shavelson, R. J., & White, R. T. (2002). *Toward a framework for achievement assessment design: The case of science education*. Stanford CA: School of Education, Stanford University.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

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., & Shavelson, R. J. (1996). 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., 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.

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

Shavelson, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., et al. (in press). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Overview and Study design. Educational Assessment.

Shavelson, R. J., & Ruiz-Primo, M. A. (1999). On the psychometrics of assessing science understanding. In J. J. Mintzes, J. H. Wamhersee & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view*. New York: Academic Press.

Vanides, J. Yin, Y., Ayala, C. & Shavelson, R. (2002). *Concept Mapping for Science Assessment: A Comparison of Constructed- and Selected-Response Approaches.* Stanford, CA.: Stanford University School of Education.