On the Assessment of Science Achievement
Conceptual Underpinnings for the Design of Performance Assessments: Report of Year 2 Activities

CSE Technical Report 491

Richard J. Shavelson and Maria Araceli Ruiz-Primo
CRESST/Stanford University

November 1998

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

Project 1.1 Models-Based Assessment Design: Individual and Group Problem Solving—Individual Problem Solving in Science Richard J. Shavelson, Project Director, CRESST/Stanford University

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

ON THE ASSESSMENT OF SCIENCE ACHIEVEMENT[1]

Richard J. Shavelson and Maria Araceli Ruiz-Primo

Stanford University/CRESST

Abstract

In this paper we provide one possible definition of science achievement and hypothesize links between the definition and instruments that can used be used to measure achievement. We define three types of science: *declarative*—knowing that something is true; *procedural*—knowing how to do something; and *strategic*—knowing the which, when, and why of doing something. The working definition identifies several characteristics of knowledge that should be considered in achievement testing and that can help identify competent and less competent students in a domain: structure—experts have highly organized knowledge, novices do not; and metacognition—experts monitor their actions and flexibly adjust them based on feedback. We also describe some instruments for measuring science achievement following from this broader notion and provide evidence bearing on their technical quality.

There is a saying that intelligence is what intelligence tests measure. This saying is an even more apt description of academic achievement: Achievement is what multiple-choice and short-answer tests measure. Lacking in all achievement testing is a reasonable, working definition of achievement to guide measurement. In this paper, we sketch, very briefly and incompletely,[2] a broader notion of achievement than is implied by current achievement testing practice. Our focus is on achievement in the domain of science, but we suspect that some ideas presented here apply to other subject matter domains. We then describe instruments for measuring science achievement that are consistent with this broader notion, and provide evidence bearing on their technical quality. We conclude by suggesting new areas for research on science achievement measurement.

[1] A version of this report will be published in a volume edited by the Max Planck Institute in Berlin.

[2] For example, the definition does not address situated cognition (e.g., Greeno, 1998). Hence, it is a working definition in need of development.

Working Definition of Science Achievement

What does it mean to achieve in a subject matter domain such as science? Surely such achievement must include knowing the important facts and concepts within the domain. That is, we expect students of physics to know what "force," "mass," and "acceleration" mean. We also expect them to know that an object in motion will continue in motion indefinitely unless other forces act on it to slow or stop it. This kind of knowledge—*knowing that* something is true—is often called *propositional* or *declarative* knowledge; knowledge about facts, concepts, and principles. Current multiple-choice and short-answer achievement tests do a reasonably good job of measuring certain aspects of propositional knowledge. Indeed, a remarkable technology of multiple-choice and short-answer testing has been developed and used extensively in the twentieth century.

Our notion of what it means to achieve in science, however, goes beyond the idea that propositional knowledge is a set of factual and conceptual "beads on a chain." For propositional knowledge to be "usable," the bits of information need to be interrelated conceptually. Experts' declarative knowledge, for example, is highly structured (e.g., Chi, Glaser, & Farr, 1988, Glaser, 1991). "Learning science" has been described, at least in part, as a process of building an increasingly sophisticated knowledge structure; that is, as a process of becoming expert in a science domain (see, e.g., Shavelson, 1972, 1974). Current paper-and-pencil achievement tests do a poor job of measuring the structural aspect of declarative knowledge. What is needed is a "picture" of how key concepts in a science domain are organized mentally by a student. That is, we need a picture of a student's "cognitive structure." To get this snapshot, alternatives to traditional achievement tests must be sought, alternatives that probe cognitive structure directly (e.g., concept maps) and indirectly (e.g., concept-similarity judgments). Some of these techniques are described later in the paper.

However, even this fuller notion of declarative knowledge stops short of what might be conceived of as science achievement. After all, scientists conduct investigations, testing curiosities, hunches and theories. They test their ideas by, for example, manipulating some variables and controlling others to gather empirical support for their hunches and theories. Shouldn't our conception of science achievement include the knowledge and skills needed to conduct such investigations? That is, shouldn't our conception of achievement include *procedural* knowledge—*knowing how* to do something? Current paper-and-pencil tests do a poor job of measuring procedural knowledge (Ruiz-Primo & Shavelson, 1996b). In conducting an investigation, for example, a scientist does something to the external world (e.g., manipulates objects in a laboratory) and the external world reacts in turn. Multiple-choice and short-answer science achievement tests do not react to the actions taken by the student. To tap a student's procedural knowledge, we might put her in a laboratory, pose a problem or set forth a hypothesis, and observe (and evaluate) how she goes about solving the problem or testing the hypothesis. Recognizing that laboratories are not ubiquitous in schools, we might construct a test in the form of a "mini-laboratory." This laboratory might be composed of materials that can be transported from one classroom to another and set up on a student's desk. Or the mini-laboratory might be provided via a computer simulation that can be transported from one computer to another. We call tests that involve student-conducted, "hands-on" investigations *performance assessments*. Some examples are provided in the paper.

There still seems to be more to achieving in science than admitted by the notions of declarative and procedural knowledge. For example, research has shown that experts combine concepts and procedures in the form of rules for action under certain task demands and work conditions. The result is a set of alternative plans to solve a problem. This kind of knowledge is known as *strategic* knowledge—knowing *which, when*, and *why* specific knowledge would be applicable. Experts seem to structure this knowledge in the form of *mental models* (e.g., Glaser, Lesgold, & Gott, 1991). They are able to use this model to bring their declarative and procedural knowledge to bear on solving a new problem or testing a new hypothesis. We do not know of any systematic assessment research and development in this area of science achievement, although there is a large research literature on mental models in science (e.g., Gentner & Stevens, 1983).

In sum, we can identify at least three kinds of knowledge that might constitute the domain of science achievement: declarative, procedural and strategic. One important characteristic of each of these three kinds of knowledge is that they are more or less structured—more structured for the expert; less structured for the novice (e.g., Glaser, 1991). Undoubtedly, there are other kinds and characteristics of knowledge that should be included in a definition of science achievement. This is a work in progress, a beginning.

Assessing Some Dimensions of Achievement

Our working definition of achievement demands a broader array of measuring instruments than typically used in testing achievement. The need for different tests to tap different kinds and characteristics of achievement constitutes a critical research agenda. Furthermore, the instruments developed must generate trustworthy results; that is, they must be reliable and valid. Here we describe some measuring instruments developed with the intent of tapping some of the forms of knowledge included in our conception of achievement. We also provide evidence on their technical quality—reliability and validity. Before doing so, we sketch our approach to evaluating technical quality.

Approach to Evaluating Technical Quality of Assessments

We have used a sampling framework to evaluate the technical quality of assessments (Shavelson, Baxter, & Gao, 1993). We view an assessment as a concrete, goal-oriented *task* with an associated response demand and scoring system. The task is performed by a student on a particular *occasion* (e.g., second week in May) and scored by an expert *rater*, who judges the scientific validity of student's procedures according to the task as well as the final product. The measurement *method* depends on the kind of assessment used. For example, the following methods have been used as a science performance assessment: observation of a hands-on investigation, a student's notebook on the investigation, a computer simulation of the investigation, or paper-and-pencil tests based on the investigation.

The *tasks* in an assessment are assumed to be a representative sample of the content in a subject domain; the content may be substantive, methodological, or both. Task sampling is easily described, and even possible to implement, at least approximately, when the domain is a concrete curriculum such as the National Science Resource Center's *Science and Technology for Children* (see Hein & Price, 1994, for a description).[3] Task sampling becomes more difficult at the state or nation level when a curriculum framework, such as California's *Science*

[3] Similarly, in military job performance measurement, domain sampling is possible, approximately, because job tasks are enumerated in doctrine (Wigdor & Green, 1991).

Framework for California Public Schools, Kindergarten Through Grade Twelve, serves as the domain specification and the curriculum itself varies from one school or classroom to another (but see Baxter, Shavelson, Herman, Brown, & Valadez, 1993).

Occasions are assumed to be sampled from a universe of all possible occasions on which a decision maker would be equally willing to accept a score on the student's performance. Occasion sampling, especially with performance assessments, has seldom been studied, due to expense (Cronbach, Linn, Brennan, & Haertel, 1997).

Raters are assumed to be a representative sample of all possible individuals who could be trained to score performance reliably. Rater sampling is not difficult to implement, but it is costly due to training, scoring, re-calibration time, and human resources.

Finally, *methods* are sampled from all permissible measurement methods that a decision maker would be equally willing to interpret as bearing on a student's achievement.

A student's performance can be viewed as a *sample* of behavior drawn from a complex universe defined by a combination of all possible tasks, occasions, raters and measurement methods. Student performance may vary across a sample of tasks, raters, occasions, or methods. Traditionally, task, occasion, and rater have been thought of as sources of unreliability in a measurement (cf. Shavelson et al., 1993; Shavelson & Webb, 1991), whereas the incorporation of measurement method in the specification of the universe moves beyond reliability into a sampling theory of validity (Kane, 1982). When performance varies from one task to another, or from one occasion to another, or from one rater to another, we speak of measurement error due to sampling variability. When performance varies from one measurement method to another, we speak of the lack of convergent validity due to method-sampling variability.

Once conceived as a sample of performance from a complex universe, generalizability (G) theory can be brought to bear on the technical quality of assessment scores (cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Brennan, 1992, Kane, 1982; Shavelson & Webb, 1991). From a G theory perspective, an assessment score is but one of many possible samples from a large domain of assessments defined by a particular task, an occasion, a rater, and a measurement method. The theory focuses on the magnitude of sampling variability due to tasks, raters, and so forth, and their combinations, providing estimates of the magnitude of measurement error in the form of variance components. These variance components can be combined to estimate a standard error of measurement for relative decisions (rank-ordering students) and absolute decisions (e.g., describing levels of student performance). In addition, G theory provides a summary coefficient reflecting the "reliability" of generalizing from a sample score to the much larger universe of interest (e.g., the score achieved over all possible tasks, occasions and raters) called a *generalizability coefficient*.

From a generalizability perspective, sampling variability due to raters, for example, speaks to a traditional concern about the viability of judging complex performance in an assessment—interrater reliability (cf. Fitzpatrick & Morrison, 1971). Sampling variability due to tasks speaks to the complexity of the subject matter domain. Traditionally, task sampling has been thought of as related to internal consistency reliability. One goal of test developers has been to make "items" homogeneous to increase reliability. Within the sampling framework, task sampling variability is dealt with not by homogenizing the tasks but by stratifying the domain, increasing sample size, or both. Sampling variability due to occasions corresponds to the classical notion of retest reliability. From a sampling perspective, the occasion facet reminds us that decision makers are willing to generalize a student's performance on one particular occasion to many possible occasions. Finally, sampling variability due to measurement method bears on convergent validity. Large-method sampling variability indicates that measurement methods do not converge as has commonly been assumed in arguing for the cost efficiency of multiple-choice testing.

With G theory we can evaluate the complex assessments described in this paper. We now turn to these assessments.

Concept Maps

Interest in assessing the structure of declarative knowledge is based on the assumption that understanding in science involves a rich set of relations among important concepts in that domain. To access students' cognitive structures—i.e., relations between concepts—two approaches have been used. *Indirect* approaches probe a student's knowledge structure by asking her to rate the similarity between concepts (e.g., Goldsmith, Johnson, & Acton, 1991), to associate words (e.g., Shavelson, 1972, 1974), or to sort concepts into groups based on their similarity (e.g., Shavelson & Stanton, 1975). A more *direct* approach is to ask a student to construct a "map" or labeled graph that makes explicit how he relates concept pairs. We focus here on the use of a direct, "construct-a-concept-map" approach to evaluate the structural aspect of declarative knowledge. Concept map assessments are interpreted to represent, at least partially, the structure of an individual's declarative knowledge in a content domain.

Definition. A concept map is a graph in which the nodes represent concepts, the lines between nodes represent relations and the labels on the lines represent the nature of the relation between two concepts (Figure 1). A pair of nodes and the labeled line connecting them is defined as a *proposition*, the basic unit of a concept map. We conceive of a concept-map-based assessment to be composed of (a) a *task* that invites students to provide evidence bearing on their knowledge structure in a content domain; (b) a format for the student's *response*; and (c) a *scoring system* by which the student's concept map can be evaluated accurately and consistently.

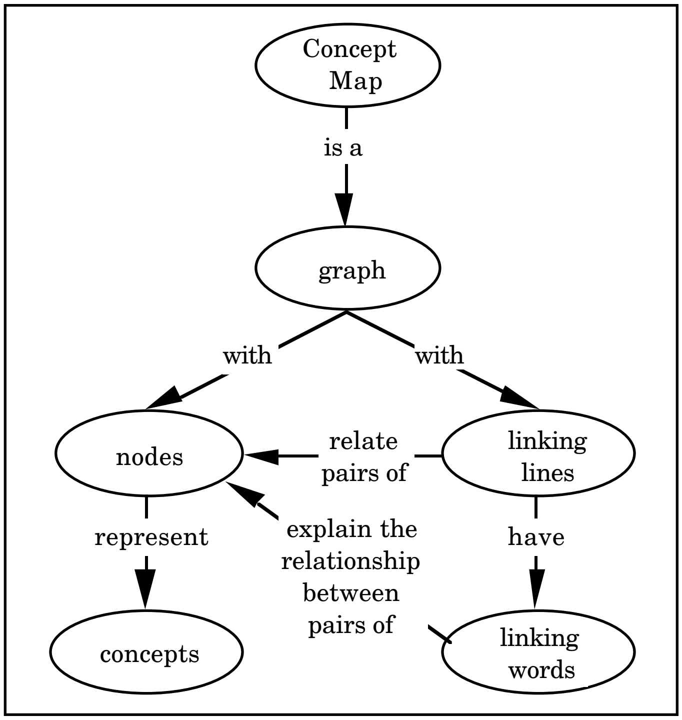

Figure 1. A concept map of what a concept map is.

Types of maps. We (Ruiz-Primo & Shavelson, 1996a) have identified different ways in which concept map tasks, response formats, and scoring systems vary in practice (Table 1). Concept map tasks vary as to (a) *task demands*—the instructions given to students in generating their concept maps (e.g., fill in a skeleton map, construct a map from scratch, or talk about the relation between concepts to an interviewer); (b) *task constraints*—the boundaries of the task (e.g., students may or may not be asked to construct a hierarchical map, or to use more than one link between concepts, or to provide the concepts for the map); and (c) *task content structure*—the intersection of the task demands and constraints with the structure of the subject domain to be mapped (i.e., there is no need to impose a hierarchical structure if the content structure is not hierarchical).

Three types of response variations have been identified in concept mapping: (a) *response mode*—whether the student's response is paper-and-pencil, oral, or on a computer (e.g., students may be asked to draw the concept map on a piece of paper or to enter the concepts and relations on a computer); (b) *response format*—the characteristics of the response requested usually fitting the specifics of the task (e.g., if the task asks students to fill in a skeleton map, a skeleton map and concepts are provided); (c) *the mapper*—who draws the map (e.g., student, teacher, interviewer).

Three scoring strategies have been used in practice: (a) *score map components* (e.g., the number of nodes, links, cross-links); (b) *compare a student's map with a criterion map* (e.g., an expert's concept map); and (c) *a combination of both strategies* (e.g., an expert's concept map is used to validate a student's links and concepts).

Technical quality. An assumption made when using concept maps is that they provide a "picture" of a student's knowledge structure. That is, the characteristics of the observed structural representation portray an important aspect of the student's underlying knowledge structure. A highly connected, integrated, organized structure characterizes experts and competent students. Isolated, less integrated structures are typical of students who are novices, less competent. However, the observed characteristics of a representation of a student's knowledge structure may depend to a large extent on how the representation is elicited, not a minor issue. From the characterization presented above, it is clear that concept mapping techniques can vary widely in the way

Table 1

Variations Among Concept Map Components

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

Note. From Ruiz-Primo and Shavelson (1996b).

they elicit a student's knowledge structure, which in turn can produce different representations and scores.

Our research has focused on providing evidence on the impact of different mapping techniques and their technical characteristics. The techniques used in our research were selected from the same task demand, "construct a map." Within this demand, task constraints were varied in different studies (see below; Figure 2). The response format (viz. draw the map on a piece of paper) and the scoring system (viz. scoring based on a criterion map and the quality of the propositions) have been held constant across the studies.

Three types of scores have been examined in our research: (a) *proposition-accuracy* score—the sum of the accuracy ratings assigned to each proposition in a student's map (assessed on a 5-point scale from 0 for inaccurate/incorrect, to 4 for excellent/outstanding in that the student provided a complete proposition that shows deep understanding of the relation between two concepts); (b) *convergence score*—the proportion of accurate propositions in the student's map out of the total possible valid propositions in a criterion map; and (c) *salience score*—the proportion of accurate propositions out of all the propositions in the student's map.

We first examined whether anything was lost in a representation of a student's knowledge structure when the assessor provided concepts for building a map instead of the student generating the concepts. We varied the source of the concept sample: student-generated sample or assessor-generated sample (Ruiz-Primo, Schultz, & Shavelson, 1996). One mapping technique asked students to provide the concepts with which to construct the map and the other technique provided the concepts (Figure 2). Results indicated that the two techniques were equivalent. No significant differences were found among means or variances for both the proposition accuracy and salience scores.[4] However, based on practical grounds, we recommend the assessor-generated concept sample method because a scoring system can be developed before data are collected and applied to all maps. The student-generated sample technique is clearly less practical in a large-scale assessment context.

[4] Convergence scores were not available for the "No Concepts" technique because no criterion could be established to determine the expected number of propositions.

Instructions When No Concepts Are Provided to the Students	**Instructions When Concepts Are Provided to the Students**
You recently studied the chapter on Chemical Names and Formulas. Construct a concept map that reflects what you know about *Ions, Molecules, and Compounds*. The concept map should have 10 concepts in it. We are providing you with 3 concepts: ions, molecules, and compounds. Select another 7 concepts to construct your map. The 7 concepts should be the ones that you think are the most important in explaining ions, molecules, and compounds. Organize the terms in relation to one another in any way you want. Draw an arrow between the terms you think are related. Label the arrow using phrases or only one or two linking words. You can construct your map on the blank pages attached. When you finish your map check that: (1) all the arrows have labels; (2) your concept map has 10 concepts, and (3) your map shows what you know about *ions, molecules, and compounds*. After checking your map *redraw* it so someone else can read it. Staple your *final map* to this page.	Examine the concepts listed below. They were selected from the chapter on Chemical Names and Formulas that you recently studied. The terms selected focus on the topic *Ions, Molecules, and Compounds.* Construct a concept map using the terms provided below. Organize the terms in relation to one another in any way you want. Draw an arrow between the terms you think are related. Label the arrow using phrases or only one or two linking words. You can construct your map on the blank pages attached. When you finish your map check that: (1) all the arrows have labels; (2) your concept map has 10 concepts, and (3) your map shows what you know about *ions, molecules, and compounds*. After checking your map *redraw* it so someone else can read it. Staple your *final map* to this page. **LIST OF CONCEPTS:** acids anions cations compounds electrons ions metals molecules molecular compounds polyatomic ions

Figure 2. Instructions to construct concept maps using techniques that differ in the demands imposed on the students.

To study the sensitivity of concept map scores to the sampling variability of assessor-generated concepts, we randomly sampled concepts from a subject domain (Ruiz-Primo et al., 1996). The same students constructed a map with two different samples of concepts (Sample A and Sample B). Half of the students constructed their maps first using Sample A (sequence 1) and the other half first using Sample B (sequence 2). No sequence effects or significant differences in means or variances were found on any type of score (i.e., proposition accuracy, convergence, salience). This result might be due to the procedure used in selecting the concepts. The list of concepts used to randomly sample the concepts was a cohesive list of critical concepts in the domain. Therefore, any combination of concepts could provide critical information about a student's knowledge on the topic.

We also evaluated the differences between mapping techniques that imposed a hierarchical and a nonhierarchical structure on students' representations of two types of content domains—one that is naturally hierarchical and one that is not (Ruiz-Primo, Shavelson, & Schultz, 1997). Regardless of the type of organization, we expected that as subject matter knowledge increases, the structure of the map should increasingly reflect the structure, hierarchical or not, in the domain as held by experts. Therefore, topics for this study were selected as having different structures according to experts' concept maps. On average, students' scores did not depend on whether the instruction to produce a hierarchical map matched a like content domain (i.e., no topic by mapping technique interaction was found in any type of score). We are still working on indicators to evaluate the hierarchical structure of the students' maps.

G theory was brought to bear on the reliability of concept map scores. Results across all the studies are clear about the effect of human judges ("raters"): Raters can reliably score students' maps. Raters, in general, did not introduce error variability into the scores (Ruiz-Primo et al., 1996, 1997). Results from the first two studies showed that the largest variance component was due to systematic differences among students' map scores—the purpose of measurement. The major source of measurement error was the interaction of persons by mapping technique; some students performed better with the student-generated concept sample, others performed better using the assessor-generated concept sample. Both relative and absolute reliability ("generalizability") coefficients were high (> .79) and of similar magnitude. This suggests that map scores can consistently rank students relative to one another as well as provide a good estimate of a student's level of performance, regardless of how well her classmates performed.

Results obtained across all the studies suggested that the type of score selected for scoring concept maps might be an issue. Results from the G studies showed that the percent of variability among persons ("true score" variability) is highest for the proposition-accuracy score, followed by the convergence score, and finally the salience score. Relative and absolute generalizability coefficients were higher for the proposition accuracy score (~ .90) than for the other two scores (~ .79). Proposition-accuracy scores, then, better reflect systematic differences in students' knowledge structures than convergence or salience scores. However, based on the amount of work and time involved in developing a proposition-accuracy scoring system, we recommend the use of convergence scores for large-scale assessment.

Finally, correlations between multiple-choice tests and concept map scores across the different studies are all positive and moderately high (*r* ~ .50). We interpret these findings to mean that concept maps and multiple-choice tests measure overlapping, yet different aspects of declarative knowledge.

Issues in the use of concept maps. There is some evidence that concept-maps tap different aspects of achievement than do multiple-choice tests. However, the virtues of concept maps may not be sufficient to overcome the challenges they face. Research is needed to determine, for example, which mapping techniques are more suitable for assessment purposes, especially for large-scale testing. We also need to know more about the cognitive demands imposed on students with the different techniques. Some researchers consider that asking students to draw a map from scratch imposes too high a cognitive demand on them to produce a meaningful representation of knowledge structure (Schau & Mattern, 1997). A technique considered as imposing a lower cognitive demand asks students to fill in the blanks in a concept map (Schau, Mattern, Weber, Minnick, & Witt, 1997). However, nothing is known about the differences in the representations provided by techniques that ask students to construct a map and techniques with lower cognitive demands, such as fill in a skeleton map. Do both techniques provide the same picture of a student's knowledge structure? Do the features of a particular technique limit the measurement of achievement maps are intend to tap?

A similar level of ignorance exists about the effects of different response modes on students' scores (e.g., drawing maps on a piece of paper or on a computer). Findings in the field of performance assessment have shown that a student's performance is sensitive to the method of assessment (see below). Whether or not these findings can be generalized to concept maps still remains to be studied.

Still another issue lies in the use of criterion maps in scoring concept maps. Simply put, there are many accurate (and inaccurate) ways to represent the interrelatedness of a set of concepts in a knowledge domain. Experts' concept maps disagree for this very reason. Which expert's map should serve as the criterion map? Different criterion maps can lead to different decisions about the adequacy of students' knowledge structures. Furthermore, are criterion maps the best way to score concept maps? At present, there is no widely accepted system for scoring concept maps. Research is needed to explore a wide variety of scoring systems that address the adequacy of propositions in concept maps. If concept maps are to be used in the measurement of achievement, a great deal of research needs to be done (see Ruiz-Primo & Shavelson, 1996a).

Performance Assessment

Science performance assessments invite students to conduct a "hands-on" investigation to test a hypothesis or solve a problem. Students plan and carry out an investigation, and report and interpret their findings. Performance assessments provide evidence bearing on procedural and strategic knowledge (e.g., Baxter, Elder, & Glaser, 1996) and sometimes propositional knowledge.

Definition. A science performance assessment is composed of (a) a *task* that poses a meaningful problem and whose solution requires the use of concrete materials that react to the actions taken by the student, (b) a *response format* that focuses the student's report of the investigation (e.g., record procedures, draw a graph, construct a table, write a conclusion), and (c) a *scoring system* that involves professionals judging both the reasonableness—scientific defensibility—of the procedures used to carry out the task and the accuracy of findings (Ruiz-Primo & Shavelson, 1996b).

To exemplify this definition we use the "Bugs" performance assessment (Shavelson, Baxter, & Pine, 1991). In this assessment, the student's *task* is to use laboratory equipment (e.g., bugs, dish, blotter paper, black paper, lamp, spray bottle) to design and conduct an investigation to find out which environment (e.g., damp or dry) sow bugs seek out. The student *responds* by drawing a conclusion about the environment and describing the procedures she used to carry out the investigation. The student's performance is *scored* by considering the scientific validity of the procedure used (e.g., comparing alternative environments in the same investigation) and the conclusion drawn from the results of the investigation.

Types of performance assessments. There are as many performance tasks as there are investigations in science. We have attempted to reduce the range by classifying them according to regularities in their characteristics (Figure 3; see Ruiz-Primo & Shavelson, 1996b; Shavelson, Solano-Flores, & Ruiz-Primo, in press). Although our categories (as would any other category system) oversimplify the complexity and uniqueness of each assessment, they focus on commonalities that have proven useful in developing other assessments within the same category. The *Others* category in our classification scheme acknowledges our ignorance and the possibility of discovering other types of tasks.

Table 2 provides examples of each type of assessment and the focus of its corresponding response format and scoring system. Here we briefly define each category and describe each example.

Types of Scoring Systems		**Types of Tasks**				
		Comparative Investigation	Component Identification	Classification	Observation	Others
Analytic	Procedure-Based	• Paper Towels • Bugs • Incline Planes • Saturation				
	Evidence-Based		• Electric Mysteries • Mystery Powders			
	Dimension-Based			• Rocks & Charts • Sink & Float		
	Data Accuracy-Based				• Day-Time Astronomy	
	Others					?
Holistic	Rubric			• Leaves (CAP Assessment)		
	Others					?

Figure 3. Types of tasks and scoring systems in performance assessments.

Table 2

Examples of Different Types of Assessments

Type of assessment	Task	Response format	Scoring system
Comparative Investigation: Saturated Solutions	Given three powders students determine which one saturates water most readily and which least readily.	Asks students to write in detail how they conducted the investigation as well as their finding.	Procedure-based. Focuses on the scientific defensibility of the procedure used and the accuracy of the findings.
Component Identification: Mystery Powders	Given bags of powder mixtures students determine which powders are in each bag.	Asks students to report the tests they used to confirm and/or disconfirm the presence of a substance as well as their observations.	Evidence-based. Focuses on the evidence provided to confirm or disconfirm the presence of a particular powder and the accuracy of the findings.
Classification: Rocks and Charts	Given some rocks, students create a classification scheme by selecting the relevant properties that help to classify these and other rocks.	Asks students to show the classification scheme they constructed and to explain why they selected the attributes used in their classification scheme.	Dimension-based. Focuses on the relevance of the attributes selected to construct the scheme and the accuracy of the use of the classification scheme.
Observation: Daytime Astronomy	Given an earth globe students model the path of the sun from sunrise to sunset and use direction, length, and angle of shadows to solve location problems.	Asks students to provide results of their observations and to explain how they collected the information.	Data accuracy-based. Focuses on the adequacy of the model used to collect the data and the accuracy of the data collected.

In *comparative investigations* students are asked to compare two or more objects on some attribute while controlling other variables. The "Saturated Solutions" investigation falls within this category (Figure 4). The *task* in this investigation asks students to compare the solubility of three powders in water (see Table 2). The Saturated Solutions *response format* invites students to provide not only their conclusion about saturation but also a description of how they conducted the investigation (i.e., the procedure used). The *scoring system* for a comparative investigation is *procedure-based*. To score students'

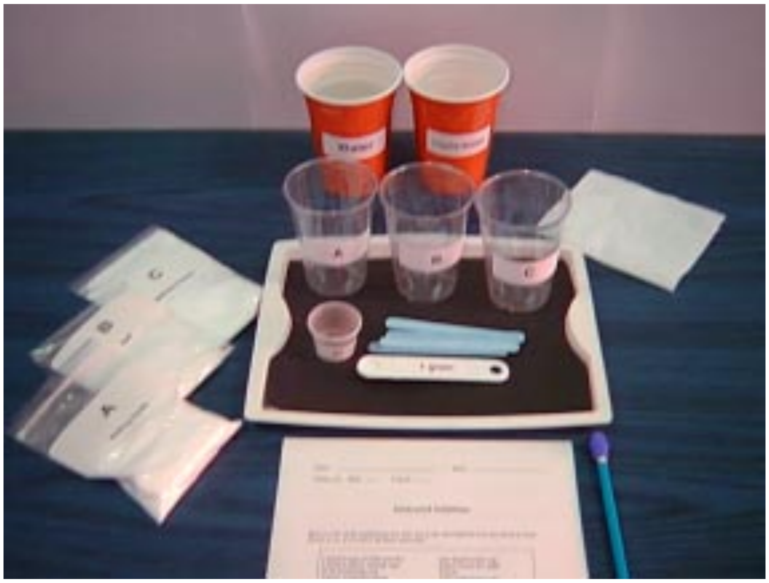

Figure 4. The Saturated Solutions investigation.

performance, information about both the quality of the procedures used and the accuracy of the problem solution is essential. For example, a student's score in the "Saturated Solutions" assessment is based on the scientific defensibility of the procedure used (e.g., did the student use the same amount of water with each of the three powders? did she carefully measure the amount of powder used in each solution?) and on the accuracy of the student's results (i.e., which of the powders is required most and which one least to saturate water?). If students are careless in measuring the water or the amount of powder, the comparison of powder solubility is flawed.

The *component identification* investigation asks students to determine the components that make up the whole. For example, in the "Mystery Powders" investigation, the student's *task* is to determine the household substances (e.g., baking soda, cornstarch, salt) that are contained in a "mystery" bag (Figure 5). The *response format* asks the student to provide information about the *tests* (e.g., iodine, vinegar) he used to *confirm* or *disconfirm* the presence of a substance (e.g., baking soda) and his observation of what happened when each test was used. The *scoring system* is *evidence-based*—it focuses on the evidence the student used to confirm the presence of one component and/or disconfirm the presence of another, as well as on the accuracy of his response as to the contents of the mystery bag.

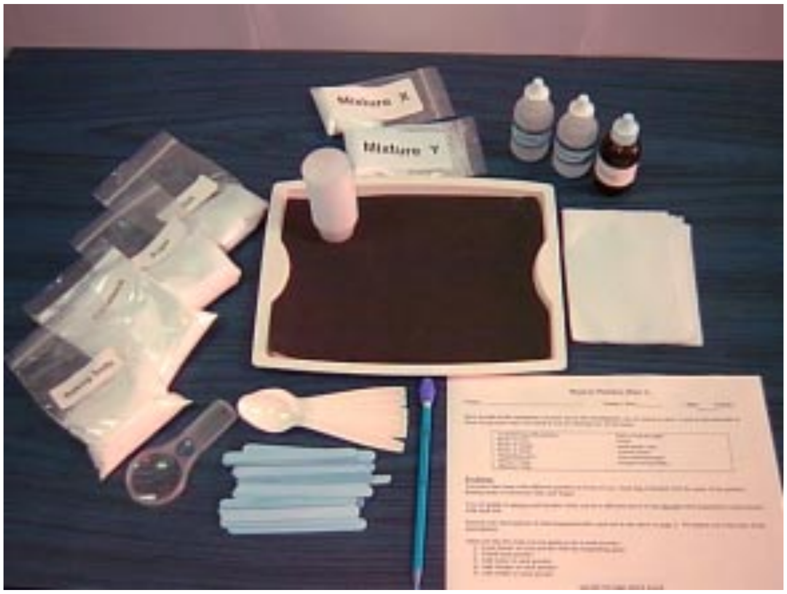

Figure 5. The Mystery Powders investigation.

The *classification investigation* asks students to create a classification scheme using attributes of a set of objects that can serve a practical or conceptual purpose. For example, the *task* in the "Rocks and Charts" investigation asks a student to consider different properties of rocks (e.g., hardness, layers, streak) to create a classification scheme and use it to classify a sample of rocks (Figure 6). In a typical *response format*, the student is asked to report the classification scheme, provide justification for the dimensions selected in his scheme, and use the scheme developed to classify a sample of objects provided. The *scoring system* is *dimension-based* and focuses on whether the student used attributes relevant to the purpose either singly or in combination to classify objects. For example, the scoring system for "Rocks and Charts" focuses on the relevance of the rocks' properties selected by the student to develop the classification scheme and how accurately he uses the scheme to classify the sample of rocks provided.

The *observation investigation* asks students to observe and systematically record an attribute of an object over a period of time. For example, the *task* in the "Daytime Astronomy" investigation asks a student to use her knowledge of Earth-Sun relations to model shadows at different times of day in the Northern and Southern Hemispheres to solve location problems (Figure 7). In the *response format*, the student provides information about the data she collected during her observations as well as the methods used to collect the data. The

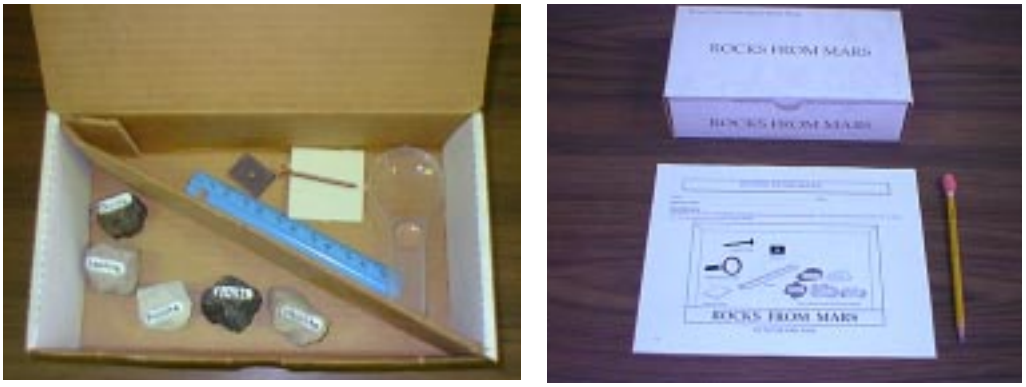

Figure 6. The Rocks and Charts investigation.

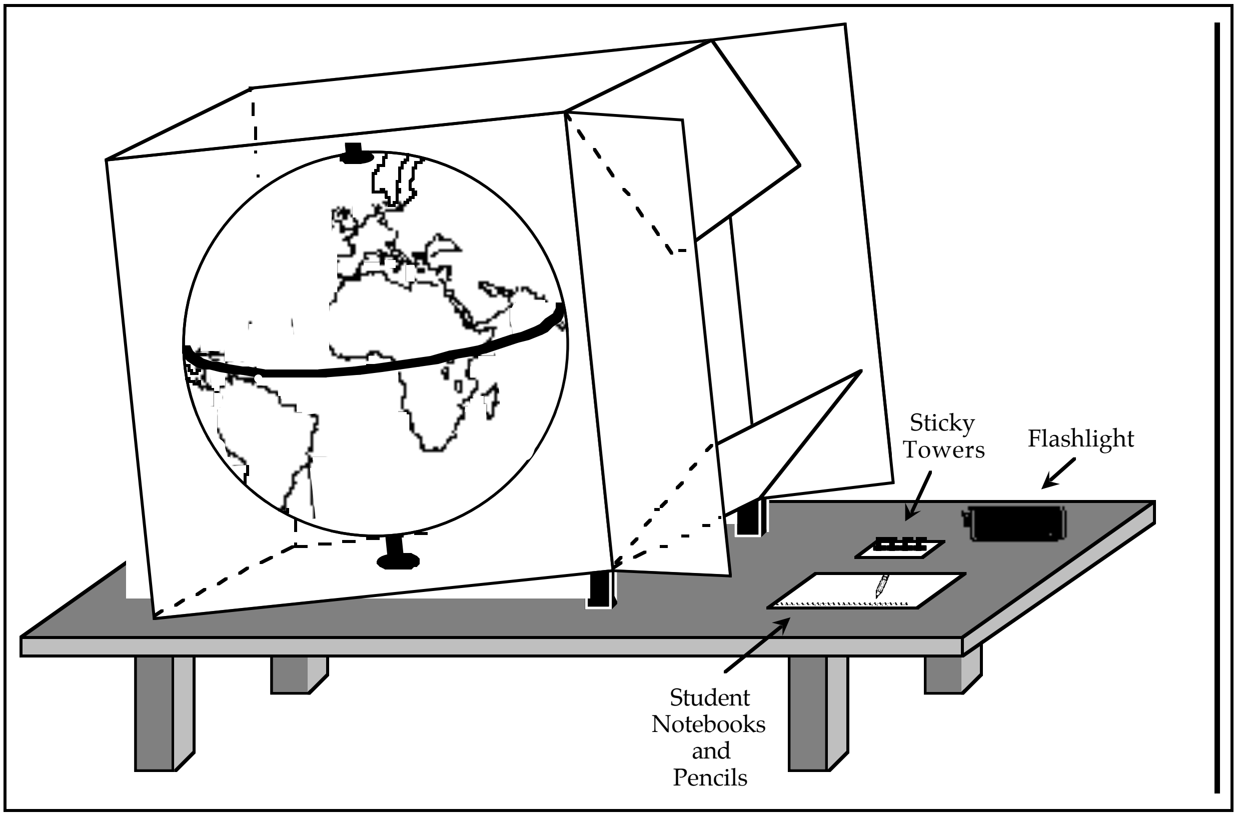

Figure 7. The Daytime Astronomy investigation.

scoring system is *data-accuracy-based*. Scoring focuses on whether students produce accurate results based on both observing natural phenomena and developing models to explain their observations. The scoring system for the "Daytime Astronomy" investigation focuses on the accuracy of the observations made and the procedure used to model sun shadows with a flashlight and an Earth globe.

Performance assessment methods. Students' performances have been measured using different methods: *direct observation*—an observer records a student's performance as the student proceeds with the investigation; *notebook*—a student records his procedures and conclusions in a notebook; *computer simulation*—a student conducts an investigation on a computer; and *paper and pencil*—a student works problems in planning, designing, and/or interpreting a hypothetical hands-on investigation (i.e., short-answer and multiple-choice tests; Baxter & Shavelson, 1994; Shavelson et al., 1991; Shavelson & Baxter, 1992). Direct observation is considered to be the ideal measurement method or "*benchmark*." The other methods are considered to be *surrogates*. Because direct observation is costly in time and human resources, the surrogate methods are used more widely in classrooms and large-scale assessments.

Technical quality. Initially greatest concern about performance assessment was attached to rater sampling variability; complex behavior was assumed to be too difficult to judge either in real time or from a written record. Research is quite clear on this issue: Raters can be trained to evaluate complex performance reliably (e.g., Shavelson et al., 1993). Nevertheless, not all individuals can be trained to score performance consistently, and raters must be continually checked and re-calibrated (Wigdor & Green, 1991).

The findings on task sampling variability are remarkably consistent across diverse domains such as writing, mathematics, and science achievement (Baxter et al., 1993; Dunbar, Koretz, & Hoover, 1991; Shavelson et al., 1993) and performance of military personnel (Wigdor & Green, 1991). Task sampling variability is large. A large number of tasks is needed to get a generalizable measure of student performance.

One study, and perhaps the only study, of occasion sampling variability with science performance assessments indicates that this source of variability may also be large (Ruiz-Primo, Baxter, & Shavelson, 1993; Shavelson et al., 1993). Indeed, occasion sampling variability is often confounded with task sampling variability because assessments are given only at one point in time (Cronbach, Linn, Brennan & Haertel, 1997). We have found that both task sampling and occasion sampling variability combined give rise to the major source of measurement error in performance assessment (Shavelson & Ruiz-Primo, in press).

Finally, method sampling variability is sufficiently great to suggest that different methods may tap into different aspects of science achievement (Baxter & Shavelson, 1994). A student's score depends on the particular task sampled and the particular method used to assess performance (see Baxter & Shavelson, 1994). Research suggests that the paper-and-pencil method is less exchangeable with direct observation ($r < .30$) than other methods. Direct observation, notebooks and computer simulations—all methods that react to the actions taken by the students in conducting an investigation—seem to be more exchangeable ($r \sim .50$; Shavelson & Ruiz-Primo, in press). The important lesson to learn from this research is that performance assessment scores are sensitive to the method used to assess performance.

Issues in the use of performance assessments. Performance assessments face challenges that, at present, limit their practicality in large-scale testing programs. For example, performance assessments are costly and time consuming to develop and administer (Stecher & Klein, 1997). The technology available is limited, and does not approach the efficiency of multiple-choice tests. Although some attempts are being made to develop a technology (Solano-Flores, Jovanovic, & Shavelson, 1997; Solano-Flores & Shavelson, 1997), we are not close to an off-the-self, high-quality performance assessment similar to that associated with multiple-choice tests.

The quality of a performance assessment determines whether or not it taps procedural knowledge. Glaser and Baxter (1997) found that the task, response format, and scoring system hold the key for tapping what performance assessments intend to measure, procedural and strategic knowledge. Assessment tasks that provide step-by-step instructions for conducting an investigation may prohibit students from demonstrating how they applied their knowledge to solve the problem. Indeed, this type of task may only show that a student is able to follow directions. Also, scoring systems inconsistent with the task do not involve students' meaningful use of knowledge and problem-solving procedures. Performance assessments need to focus not only on content but also on the process-demands of the assessment task (Glaser & Baxter, 1997).

A major issue in using performance assessments is that a substantial number of tasks is needed to reliably estimate student- and school-level performance due to task sampling variability. At the individual level, 8 to 23 tasks may be needed to reach reliability .80 (Gao, Shavelson, & Baxter, 1994; Shavelson et al., 1993). At the school level, a trade-off between the number of students tested and number of tasks in an assessment should be considered. For example, to achieve reliability .80, 50 students and 15 tasks or 100 students and 12 tasks may be needed. Gao et al. (1994) found that as few as 7 tasks may be needed for a sample of 25 students if matrix sampling is used. The impact of task sampling on time, cost, and human resources is substantial. It may take about 2.25 hours of testing time to obtain a generalizable measure of student achievement if we consider 7 tasks of 20 minutes each.

The brief review of the state of the art in science performance assessments presented above makes clear that high-quality performance assessments are costly to produce, administer and score. A long-term research agenda should (a) develop a high-quality performance assessment technology for use in curriculum, and in state and national examinations; (b) examine and learn how to reduce task/occasion sampling variability; and (c) explore which measurement methods are the most appropriate for testing students.

Concluding Comments on Directions for Science Achievement Measurement

We have set forth an incomplete working definition of science achievement. We have conceived of three types of knowledge that need to be included in such a definition. The first type of knowledge is declarative knowledge, knowing that something is true. This knowledge includes facts, concepts and principles; paper-and-pencil achievement tests do a reasonably good job of measuring important aspects of declarative knowledge. The second type of knowledge is procedural knowledge, knowing how to do something. This kind of knowledge includes knowledge of procedures for carrying out a scientific investigation (e.g., controlling some variables, manipulating others, and using the appropriate measurement). Performance assessments do a reasonably good job of measuring important aspects of this knowledge. The third type of knowledge is strategic knowledge, knowing the which, when and why of doing something. Strategic knowledge is organized in a mental model that represents a student's understanding of the phenomenon being dealt with and helps her integrate declarative and procedural knowledge so as to bring them to bear in specific situations.

In addition to mapping out three areas of knowledge to be considered in achievement testing, the working definition identified several characteristics of knowledge that warrant consideration in achievement testing. The first characteristic of knowledge (declarative, procedural and strategic) is that it is structured; experts have highly organized knowledge, novices do not. Concept maps do a reasonably good job of measuring aspects of the structure of declarative knowledge, although a great deal of research remains to be done with this measurement technique. However, there has been little research on the measurement of the structural aspects of procedural or strategic knowledge. Further achievement testing research should address this gap.

Another characteristic of knowledge is "metacognition." Metacognition involves an individual monitoring how she accesses and uses knowledge. Metacognition also involves use of heuristic strategies for searching knowledge, and for checking to see whether the search produced reasonable, reliable results. Although there has been considerable research on metacognition, this construct has not been integrated into achievement testing. Again, research is needed to fill this gap.

At this point, it should be clear as to why we consider our definition of achievement an incomplete, working definition. The definition is incomplete because it does not include some aspects of knowledge such as scientists' tacit knowledge of norms in a laboratory culture or the conventions they use to represent ideas when working together in the laboratory (e.g., Bleicher, 1996; Greeno, 1998; Kozma, Chin, Russell, & Marx, 1997). Nor does it include an adequate array of knowledge characteristics such as structure and metacognition. The definition is a working definition in that although it helps to guide the development and interpretation of science achievement assessments, the definition will change as we gather new information from achievement assessments themselves.

References

Baxter, G. P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*, 133-140.

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

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal of Research in Mathematics Education, 24*, 41-53.

Bleicher, R. E. (1996). High school students learning science in university research laboratories. *Journal of Research in Science Teaching, 10*, 1115-1133.

Brennan, R. L. (1992). *Elements of generalizability theory* (2nd ed.). Iowa City, IA: ACT.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4*, 289-303.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 237-270). Washington, DC: American Council on Education.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7*, 323-342.

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

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-39). Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R., & Baxter, G. P. (1997, February). *Improving the theory and practice of achievement testing*. Paper presented at the conference "Science Education Standards: The Assessment of Science Meets the Science of Assessment," National Academy of Sciences/National Research Council, Washington, DC.

Glaser, R., Lesgold, A., & Gott, S. (1991). Implications of cognitive psychology for measuring job performance. In A. K. Wigdor & B. F. Green (Eds.), *Performance assessments in the work place* (Vol. II, pp. 1-26). Washington, DC: National Academy Press.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology*, *83*, 88-96.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, *53*(1), 5-26.

Hein, G. E., & Price, S. (1994*). Active assessment for active science: A guide for elementary school teachers*. Portsmouth, NH: Heinemann.

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement, 6,* 126-160.

Kozma, R., Chin, E., Russell, J., & Marx, N. (1997). *The roles of representations and tools in the chemistry laboratory and their implications for chemistry instruction* (SRI Project No. 5871). Menlo Park, CA: SRI International.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, *30*, 41-53.

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

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

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

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

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

Schau, C., Mattern, N., Weber, R. W., Minnick, K., & Witt, C. (1997, March). *Use of fill-in concept maps to assess middle school students' connected understanding of science*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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). Methods for examining representations of a subject-matter structure in a student's memory. *Journal of Research in Science Teaching, 11,* 231-249.

Shavelson, R. J., & Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49,* 20-25.

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

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessments in science. *Applied Measurement in Education, 4,* 347-362.

Shavelson, R. J., & Ruiz-Primo, M. A. (in press). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*.

Shavelson, R. J., Solano-Flores, G., & Ruiz-Primo, M. A. (in press). Toward a science performance assessment technology. *Evaluation and Program Planning*.

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

Solano-Flores, G., Jovanovic, J., & Shavelson, R. J. (1997). *On the development and evaluation of a shell for generating science performance assessments*. Manuscript submitted for publication, WestEd, San Francisco, CA.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice*, *16*, 16-25.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis*, *19*, 1-14.

Wigdor, A. K., & Green, B. F. (Eds.). (1991). *Performance assessments in the work place* (Vol. 1). Washington, DC: National Academy Press.