An Approach to Analyzing the Cognitive Complexity of Science Performance Assessments

CSE Technical Report 452

Gail P. Baxter
University of Michigan

Robert Glaser
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)/
Learning Research and Development Center
University of Pittsburgh

December 1997

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
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA Catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

AN APPROACH TO ANALYZING THE COGNITIVE COMPLEXITY OF SCIENCE PERFORMANCE ASSESSMENTS

Gail P. Baxter
University of Michigan

Robert Glaser
CRESST/Learning Research and Development Center
University of Pittsburgh

Abstract

Psychological theories that describe the development of subject-matter competence in terms of changes in the quality of cognition provide a basis for reconsidering the design and evaluation of alternative assessments. Such an evaluation explicitly considers the thinking and reasoning activities elicited in assessment situations and the extent to which these activities are given preference in defining subject-matter achievement. This paper describes an approach to address this cognitive validity issue that utilizes comparative studies of expertise. Illustrative examples from cognitive analyses of current assessment practices make apparent the utility of this approach for identifying performance objectives and examining the correspondence among test objectives, performance scores, and observed cognitive activity. This approach calls for the integration of assessment practice and knowledge of learning, and challenges the measurement community to (a) reconceptualize achievement test theory and design to systematically incorporate the cognitive aspects of performance, and (b) formulate appropriate methodologies for analyzing the cognitive properties of assessments in various subject-matter areas.

The art and practice of achievement measurement are in transition. Innovative procedures and situations that assess the ability for thinking and high levels of competence realizable in the course of schooling are being introduced. In contrast to selection testing, which developed on the basis of concepts of aptitude and intelligence, the theory underlying changes in assessment of school achievement has been less explicit. Nevertheless, assessment development has proceeded rapidly in response to policy mandates and curricular changes, while at the same time psychological theory has matured

from stimulus-response descriptions of behavioral objectives to more cognitive accounts of the processes of complex human performance involved in thinking, reasoning, and problem solving. This significant theoretical development has laid a foundation that can influence the nature of validity evidence required to support the use of performance assessments. At present, much work is experimental and more study is required to effectively incorporate this theory in the design and evaluation of current forms of assessment.

An overriding requirement for progress along these lines is to conceptualize student competence with respect to the quality of cognition that develops during school learning. Initial efforts in this regard have been guided by the expert-novice literature and its contributions to current understandings of the relationship between competence and quality of cognitive activity (Glaser, 1991). Drawing on these understandings as a framework, the properties and objectives of assessments and scoring systems piloted in a number of prominent state and district testing programs were examined. The objective was to ascertain whether and how these assessments are measuring cognitive capabilities that distinguish various levels of student achievement (e.g., Baxter, Elder, & Glaser, 1994, 1996; Baxter, Glaser, & Raghavan, 1993). Using information obtained through protocol analysis techniques (cf. Chi, 1994; Ericsson & Simon, 1993), observations of student performance, and a review of student written work, researchers matched test objectives and the realization of those objectives to cognitive activities characteristic of effective learning and competence in a subject matter.

From these analyses, three types of assessment situations were identified: (a) those in which tasks elicited appropriate cognitive activity, and the nature and extent of this activity correlated with performance scores; (b) those in which tasks elicited appropriate cognitive activity but the scoring system was not aligned with the task demands, objectives of test developers, or the differential quality of student cognition; and (c) those in which tasks were configured so that cognitive aspects could be bypassed.

The results of these analyses suggest that assessments can vary in terms of stated objectives, the relationship between test objectives and observed performance, and the extent to which students' scores reflect quality of observed performance. Consideration of these aspects of assessment situations within a cognitive framework provides a theoretical and empirical basis to guide

assessment development and evaluation to use the quality of cognition in defining and measuring subject-matter competence. The framework, derived from cognitive studies of the performances of experts and novices in knowledge-rich domains, provides a common language for test developers and users to describe and evaluate student cognition in learning and assessment contexts. This paper describes this cognitive framework and details an approach for gathering evidence of the cognitive complexity of alternative assessments that is guided by this framework. The approach, with examples from empirical studies of current assessment practice, illustrates ways in which issues such as establishing performance objectives, comparing intended objectives with observed performance, and measuring observed performances in terms of the quality of cognition can be articulated and addressed.

Cognitive Components of Competence

Current forms of assessment call attention to the need for additional criteria to establish the validity of score use and interpretation, particularly the quality and nature of the performance that emerges in an assessment situation. "Claims that performance assessments measure higher order thinking skills and deep understanding, for example, require detailed cognitive analysis" (Baker, O'Neil, & Linn, 1993, p. 1216; see also Linn, Baker, & Dunbar, 1991). Detailed cognitive analysis should illustrate the kind of performance actually elicited from students in alternative assessment situations and document the relationship between those performances and the problem-solving activities that contribute to differential performance (Glaser, Raghavan, & Baxter, 1992). That is, "the level and sources of task complexity should match those of the construct being measured and be attuned to the level of developing expertise of the students assessed" (Messick, 1994, p. 21).

The nature of expertise and subject-matter competence has been the focus of numerous studies in human cognition (e.g., Bereiter & Scardamalia, 1987; Charles & Silver, 1988; Chase & Simon, 1973; Gobbo & Chi, 1986; Schoenfeld, 1992). These studies and others (e.g., Chi, Glaser, & Farr, 1988) have examined the differences between people who have learned to be competent in solving problems and performing complex tasks and beginners who are less proficient. Results indicate that when learning a new subject matter, both children and adults develop special features of their knowledge that contribute to their ability

to use it well. Key among them are integrated knowledge, so that students can think and make inferences with what they know, and usable knowledge, knowledge that is not just mere factual information but that allows this information to be used in appropriate situations. Knowledge structures of this sort enable students to accurately represent a problem with respect to underlying principles; select and execute goal-directed solution strategies based on an understanding of the task; monitor and adjust their performance when appropriate; and offer complete, coherent explanations and justifications for problem-solving strategies and changes or adjustments to performance. In contrast, less proficient students are characterized by fragmented knowledge that remains isolated from an understanding of the conditions or situations in which particular conceptual or procedural skills would be appropriately used.

Depending on the experience and degree of learning, this knowledge integration or fragmentation varies as does the nature of cognitive activity. General differences in knowledge structure and cognitive activity are elaborated in Table 1 and below in a heuristic fashion intended solely to frame subsequent discussions. The nuances and complexities of the development of competence in various domains are presented elsewhere (e.g., Chi et al., 1988; Glaser, 1992).

Table 1

Cognitive Activity and Structure of Knowledge

	Structure of knowledge	
Cognitive activity	Fragmented	Meaningfully organized
Problem representation	Surface features and shallow understanding	Underlying principles and relevant concepts
Solution strategies	Undirected trial-and-error and problem solving	Efficient, informative, goal oriented
Self-monitoring	Minimal and sporadic	Frequent and flexible
Explanations	Single statement of fact or description of superficial factors	Principled and coherent

Problem Representation

Competent individuals qualitatively assess the nature of a problem and construct a mental model or internal representation prior to initiating a solution strategy (Gentner & Stevens, 1983; Halford, 1993). These individuals employ a

representation to plan various actions, anticipate alternative outcomes, and generate next steps based on those outcomes. In other words, they perceive a problem in terms of underlying concepts that guide their actions toward problem solution. This is particularly noticeable when students are given problems that are less routine than the ones they are used to. In these situations, competent students tend to use their representation to reduce a problem to a simpler one, or add conditions that make it a familiar problem and then attempt the solution in small steps that increasingly approach the problem's real complexity. This ability to build a meaningful problem representation is not well developed in less competent students. When asked how they will go about solving a problem, these students may name the equipment they will use or offer a simple statement indicating how they will begin without reference to underlying concepts or without anticipating the process, the overall goal, or ways in which the process will lead to a solution.

Solution Strategies

Principled problem solving is characterized by the use of goal-directed, efficient strategies and is reflective of substantial knowledge organization and structure (e.g., Siegler, 1988). Competent students have a repertoire of subject-specific and general problem-solving strategies they employ depending on the particulars of the problem or task. In some situations, the performance of these students may appear algorithmic because of the integration of conceptual knowledge with procedures for its applicability. Strategies are also used effectively and flexibly in response to different situational conditions or problem constraints (e.g., Anderson, 1985). In contrast, the performance of less competent students is characterized by rigid adherence to an initial strategy, rote procedures not informed by details of the problem, or inconsistent application of a potentially effective strategy.

Self-Monitoring

As competent students construct and structure their knowledge, they develop a set of cognitive skills they use to control and regulate their performance (Brown, 1978; Glaser, 1996). They learn to monitor their problem solving by judging the difficulty of problems and allocating their time appropriately; they note their errors and failures to comprehend, and check questionable solutions and answers. With increasing competence, these self-regulatory skills become

well practiced and an integral part of performance. For beginning or poor learners in a subject matter, teachers often explicitly prompt self-regulatory skills, such as restating the problem or anticipating the consequences of a solution step. Although these students may display these behaviors when prompted, they often do not generate them on their own. Rather, once the problem solution is set in motion, these students see it through to the end without checking the utility of their strategy or giving attention to the consequences. This failure to monitor exhibits itself in the form of contradictory or illogical statements that go unnoticed by the individual performing the task.

Explanations

Effective learning of content knowledge enables students to explain concepts and principles underlying their performance, provide justification for their efforts, and draw inferences to other situations. As students acquire knowledge, they become increasingly skilled at justifying what they do and making well-formed explanations for themselves and others about the reasons for their answers and solutions (e.g., Chi, Bassock, Lewis, Reimann, & Glaser, 1989; Fay, 1995). Less knowledgeable or less competent students offer simple assertions to justify particular actions or describe what they did but not why they did it. These students may respond like their more competent peers to simple recall or recognition questions, but their knowledge is not sufficiently structured to sustain thinking and reasoning on complex tasks.

Summary

A well-connected knowledge structure links concepts and processes with conditions under which those concepts and processes should be used. This meaningful organization facilitates thinking and reasoning about a task, planning an appropriate solution strategy, performing in a principled manner, and effectively monitoring one's performance. In contrast, isolated or loosely related pieces of information constrain performance to simple recognition questions, surface-level observations, fragmented explanations, and trial-and-error strategies. An understanding of these features of differential competence provides a useful framework for evaluating the cognitive complexity of alternative assessments (Glaser et al., 1992) because these cognitive features of performance are observable and therefore can be incorporated in subject-matter instruction and assessment.

An Approach to Evaluating Cognitive Complexity

The framework above provides a theoretical and empirical basis for describing test objectives, clarifying situations that elicit student performance consistent with those objectives, and measuring elicited performance so as to give preference to those aspects that reflect competence in a domain. The list of cognitive activities—problem representation, strategy use, monitoring, explanations—focuses attention on the distinguishing features of differential competence and subject-matter achievement and should not be conceived as the necessary requirements for all alternative assessments. As might be expected, assessment tasks will engage students in some cognitive activities (e.g., explanation) but de-emphasize others (e.g., self-monitoring) depending on the objectives of the test developers and the content-process features of the assessment situation. The interrelationship of content and process that influences cognitive complexity is considered below.

Establishing Performance Objectives

The realization of particular cognitive performance objectives stems, in part, from the content and process complexity of the task involved. The nature and extent of content knowledge that assessments demand fall on a continuum from lean to rich. At one extreme, accurate response to a task or successful problem solution is not dependent on prior knowledge or related experiences—that is, necessary knowledge is provided in the task. At the other extreme, in-depth understanding of subject matter—that is, integrated conceptual and procedural knowledge—is requisite for optimal task completion (see Figure 1). Likewise, the task demands for process skills fall on a continuum from constrained to open. Process skills are constrained when student performance is dependent on following detailed directions for task completion and subject-specific procedures given as part of the task. In open situations, explicit directions are minimized as students are expected to generate and carry out appropriate process skills for problem solution.

As shown in Figure 1, assessment tasks can, in theory, involve any of a number of possible combinations of content knowledge and process skills. Science tasks can be explicitly designed so that successful performance is more, or less, dependent on structured content knowledge and student-generated process skills such as observation, classification, or experimental design. The location of

an assessment within this content-process space is related to the nature and extent of cognitive activity underlying competent performance and, as such, provides a useful schema for conceptualizing and clarifying the objectives of assessment design. Four examples from current assessment practice illustrate the correspondence between the content and process demands of the task and the kinds of cognitive activity that are likely to be elicited.

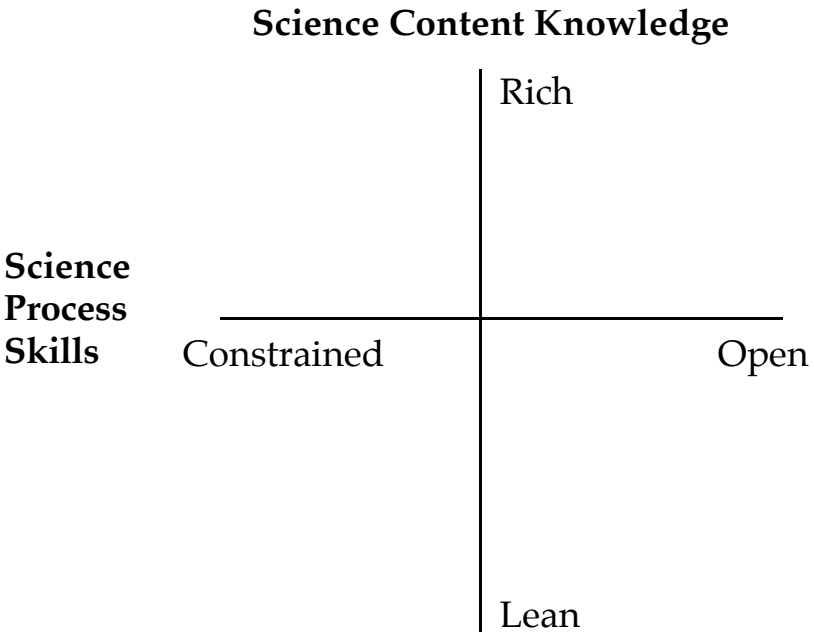

Figure 1. Content-process space of assessment tasks.

Content rich/process open. "Exploring the Maplecopter" is a good example of a content rich/process open task. High school physics students are asked to design and carry out experiments with a maple seed to explain its "flight to a friend who has not studied physics" (Baron, Carlyon, Greig, & Lomask, 1992). The flight of the maple seed "represents a delicate equilibrium between gravity, inertia, and aerodynamic effects" (Seter & Rosen, 1992, p. 196). Consequently, for this task, identification of the causal variables involved requires substantial knowledge of physics concepts of force and motion, the ability to design and carry out controlled experimentation, and the effective employment of model-based reasoning skills. Indeed, understanding how and why the maple seed falls as it does has drawn attention from a broad spectrum of researchers including biologists, biophysicists, and aerospace engineers

because of its complexity and the unresolved controversy over the most appropriate model (e.g., flight of a bird versus helicopter) to adequately describe these phenomena (e.g., Green, 1980; Norberg, 1973; Seter & Rosen, 1992).

Given that the problem does not have a clean, simple solution, the task is rich with opportunities for high school physics students to apply their subject-matter knowledge and in-school experience to understand an everyday phenomenon—the flight of the maple seed. In this context, optimal performance is dependent on an adequate representation of the problem, sustained and systematic exploration strategies (observation and experimentation), monitoring progress toward describing the flight of the maple seed, and explaining the causal relationships observed and tested.

Content lean/process constrained. In contrast to the maplecopter task, tasks that are knowledge lean/process constrained require minimal prior knowledge or school experiences with subject-specific concepts and procedures to successfully complete the task. Rather, students are guided to carry out a set of procedures and then asked to respond to a set of questions about the results of their activities. For tasks of this type, generative opportunities for problem representation, strategy use, and monitoring are precluded by the step-by-step procedures provided by the assessment. Further, knowledge requirements are given in the task such that student responses are independent of the kinds of formal instructional experiences they bring to the situation.

For example, consider a task that asks eighth-grade students to study the effects of a train derailment and the resulting chemical spill on the surrounding environment (California Department of Education, 1993b). As part of their investigation, students replicate potential chemical reactions from that situation. They are explicitly *directed* to add specific amounts of the relevant substances in a *specified* sequence. After following the instructions to set up three chemical reactions, they are *prompted* to observe each of these reactions for temperature, color change, and "other changes observed." A table is provided to *guide* recording of the specified observations. Students are then posed a series of questions that, for the most part, can be answered by *rereading* data from the table of observations or other information provided. In short, the cognitive activities of problem solving discussed above are less relevant in this situation than are the activities involved in reading and following directions.

Content lean/process open. Assessment tasks may require students to coordinate a sequence of process skills with minimal demands for content knowledge. For example, the "Mystery Powders" assessment asks fifth-grade students to identify the substances in each of six bags from a list of five possible alternatives (Baxter, Elder, & Shavelson, 1995). Students are presented with vinegar, iodine, water, and a hand lens to test each substance or combination of substances. Two of the bags have baking soda and cornstarch in them. Each of the others contains either baking soda, baking soda and salt, cornstarch, or cornstarch and sugar. Students are told they can use the equipment in any way they wish to solve the problem.

With instructions of this sort, students structure the problem in terms of actions that follow from what they know about the properties of substances and ways to identify them (i.e., tests and relevant observations). They then implement a strategy, such as adding vinegar to a substance, and revise their strategy, if necessary, based on task feedback (e.g., no fizz, try iodine to test for cornstarch). In carrying out the "Mystery Powders" assessment, students attend to and coordinate multiple pieces of information including knowledge of task constraints, knowledge of critical aspects of their previous investigations, and interpretations of current trials.

In this situation, processes are open in terms of test selection (number and type of test) and test sequence that can be carried out more, or less, efficiently as a function of effective monitoring and students' knowledge of the relationship between substances and their identifying tests that they bring to the situation. The content knowledge requirements for successful task completion are lean relative to, say, the maplecopter task described above; students need to know how to replicate previous investigations and how to match current trials with records of in-class observations of test-substance outcomes.

Content rich/process constrained. Tasks that are content rich/process constrained emphasize knowledge generation or recall—that is, "knowing" science versus "doing" science. For example, high school students were asked to "describe the possible forms of energies and types of materials involved in growing a plant and explain fully how they are related" (Lomask, Baron, Greig, & Harrison, 1992). A comprehensive, coherent explanation revolves around a discussion of inputs, processes, and products (e.g., the plant takes in water, light, and carbon dioxide; through the process of photosynthesis, light energy is

converted into chemical energy used to produce new materials such as sugar needed for plant growth; in addition, oxygen is given off). In developing their explanations, students make decisions about which concepts are important and how these concepts are related, thereby reflecting their conceptual understanding of the topic. Although the opportunities for explanation are apparent, opportunities for other activities, such as planning, selecting and implementing appropriate strategies, and monitoring problem-solving procedures, are not.

Summary. Characterizing assessment situations in terms of components of competence and a content-process space brings specificity to generic assessment objectives such as "higher level thinking and deep understanding" and offers a set of observable cognitive activities as relevant criteria for evaluating student performance and designing instructional environments. Moreover, these sorts of concrete descriptions make apparent the relative merits of selecting a task for purposes consistent with stated content, process, and cognitive performance objectives. Furthermore, with clearly articulated objectives and an understanding of the correspondence between particular task features and cognitive activity, the content and process demands of tasks can be adjusted to align task features with the cognitive performance objectives of test developers.

Relationship Between Objectives and Performances

From a cognitive validity perspective, a critical concern is the translation of performance objectives into assessment situations in ways that ensure the relevant cognitive activities are elicited. In some situations, the expectations and objectives of test developers are realized in observations of student engagement in relevant cognitive activity. In other situations, tasks may be structured in such a way that relevant cognitive skills are bypassed. Mismatches between task expectations and observed performance direct developers to aspects of the task situation, such as instructions, equipment, or response format, that merit reconsideration and possible revision. Information about these matches or mismatches can be obtained most usefully through formative studies carried out during assessment design, but this information can also be obtained following test development (Baxter, Elder, & Glaser, 1994, 1996).

Techniques for obtaining descriptions of various levels of performance have been devised by cognitive psychologists (Chi, 1994; Ericsson & Simon, 1993) and figure prominently in studies of problem solving in knowledge-rich domains (e.g.,

Chi et al., 1988). These techniques, with modifications, can be used to examine the cognitive activity of students while carrying out an assessment task or through stimulated recall and retrospective reports. For each assessment situation, direct questions, prompts, requests for elaboration, and various protocol techniques can be utilized that are appropriate for the age/grade level of participants, the nature of the task situation, and the cognitive activities to be observed. For example, insight into students' problem representations can come from pretask questions such as "How are you going to go about solving this problem?" or "How would you use a model to solve the given problem?" or probes for meaning such as "Which concepts have you learned in physics class that might be helpful in solving this problem?" Evidence for strategy use and self-monitoring can be gathered from a combination of direct observations of student performance, student self-reports or verbalizations, and elaborations of their thoughts, questions, procedures, and corresponding justifications. Explanations can be stimulated by questions about task-related concepts, such as "Can you tell me what a circuit is and how it works?" or "What is the Rh factor and what is its relationship, if any, to blood type?" Information so obtained can be used to examine the relationship between anticipated objectives and observed performance with respect to student cognitive activity.

In studies of the cognitive complexity of science assessments, analyses much like those described above were carried out *after* the assessments were developed and in trial use in state and district testing programs (e.g., Baxter et al., 1993, 1994, 1996; Breen, Baxter, & Glaser, 1995). The sample included tasks that were developed with the express purpose of assessing students' ability to reason with subject-matter knowledge to solve circuit problems (Shavelson, Baxter, & Pine, 1992) or identify unknown substances (Baxter et al., 1995); to design, carry out, and interpret results from an extended inquiry of the flight of a maple seed (e.g., Baron et al., 1992); to generate explanations of photosynthesis, cell respiration, or other major topics in life, earth, and physical science (Lomask et al., 1992); or to bring together their understandings in one or more science domains to make decisions in real-world contexts such as designing a nature walk area (California Department of Education, 1993a) or investigating the impact of an earthquake on the surrounding environment (California Department of Education, 1993b). Development of these assessments was primarily a creative process guided by curricular frameworks, subject-matter specialists, and other available resources.

Explicit or implicit in the design of these assessment tasks was an attempt to assess problem solving and higher level thinking. These efforts to translate goals into practice resulted in varying degrees of success. Three examples are provided below; one is a match between objectives and performance followed by two examples of mismatches.

Match between objectives and performance. The first analysis uses the "Exploring the Maplecopter" task described earlier as an example of a content rich/process open task. In this task, high school students are asked to investigate and then explain the flight of the maple seed. To understand and explain relations among causes and effects requires sustained and broad exploration, beginning with observation and followed by experimentation and interpretation of findings.

Initial observations of the flight of the maple seed are influenced by prior knowledge by which students look at the event. Furthermore, some aspects of the flight of the maple seed are more readily observable than others. For example, even the casual observer would notice that there are two phases—an initial free-fall followed by a spinning stage. Other observations require a more focused and informed look (e.g., rigid edge of the wing is the leading edge). After completing their observations, students work in groups to design and carry out experiments to explain the spinning flight patterns. This provides an opportunity for them to develop explanations and construct paper models for this purpose. Following experimentation, students individually explain the motion of the maple seed.

A review of student responses (n = 6 general physics and n = 8 Advanced Placement physics) showed that all students observed the two phases of free-fall and spinning. In addition, observations varied with respect to the number and combinations of factors observed, regardless of current physics enrollment. That is, some students observed that the velocity of free-fall is greater and the motion is different with different starting positions. Other students observed that the wing spins around a vertical axis of rotation either clockwise or counterclockwise with the rigid edge leading.

When asked which concepts from physics class applied to the current situation, general physics students could list several relevant concepts such as center of mass, air resistance, and gravity. However, when asked to explain the flight of the maple seed using these concepts, their responses consisted primarily

of a list of statements. The following typifies the explanations generated by these students:

> Gravity usually makes an object fall at about 9.8 m/s^2. Because the maple leaf and maplecopters have so much air resistance they didn't fall at even close to that. The center of mass was at the base, and because of this the maplecopter spun around rather than falling straight down. If the mass was in the middle rather than at one end it would have fallen straight down. The smaller the leaf, the faster it would spin. The way it fell depended on its mass, its size, and its wings.

Contrast the quality of this explanation with the following from an Advanced Placement physics student; this explanation more adequately formulates the important forces and factors and their impact on observed flight patterns:

> There are several variables that affect the motion of the winged maple seed; the curve, the weight of the seed with respect to the wing, the weight of the "hard edge," the surface area/air resistance. When the maple seed is initially dropped, it can be seen that the heavier end, the seed, leads the way down. As the light wing of the maple seed accelerates, the air resistance continues to build. As the air begins to build under the wing, it searches for an escape. The curve is concave up to the side without the "hard edge" and so the air escapes to this side. Subsequently, a force is produced that results in a spinning motion in the direction of the "hard edge."

In summary, students were asked to study the motion of maple seeds and design experiments to explain their spinning flight patterns. The intent was to evaluate their understanding of "laws of motion, aerodynamics, air resistance, and the use of models in explaining scientific phenomena" (Lomask et al., 1992). In this situation, performance variation was directly observable in terms of the number and kinds of observations students made and the quality of student explanations. Further, these aspects of performance were directly related to students' experience with the subject matter; students in Advanced Placement physics classes consistently outperformed students in general physics classes. In short, rich knowledge and student-generated process skills define the nature of this task, thereby providing opportunities for students to display their level of understanding of the concepts and processes being tested.

Mismatch (1) between test objectives and observed performance. In a plate tectonics task, eighth-grade students are asked to "examine the process that causes rock layers to fold and twist" (California Department of Education,

1993b). To this end, materials and directions to set up a plate model are given (see Figure 2). Then students are directed to move one plate horizontally and describe what happens, and then move one plate vertically and describe what happens. In the final question, students are given a map that shows the current location of two land formations that were once located next to each other and are asked to "explain how the Pinnacles National Monument and the Tejon Pass were separated from each other." A word bank is provided that lists related vocabulary students can use in their answers.

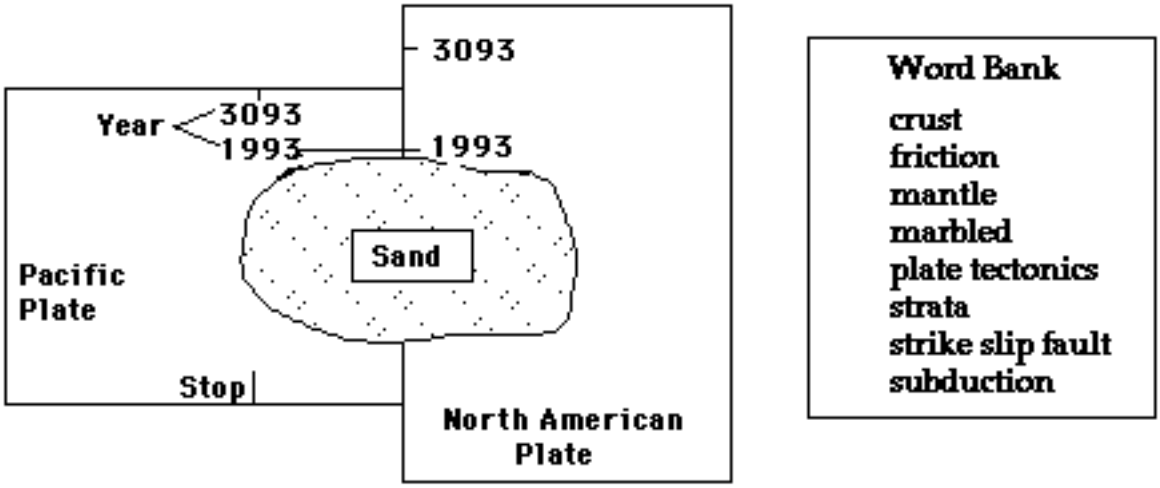

Figure 2. Plate model and word bank for "The Fault Line" assessment (adapted from California Department of Education, 1993b).

When a sample of students (n = 23) were asked to describe what happened when the plates were moved vertically, they all stated that the sand piled up. Likewise, when asked to "describe what happened to the sand at the plate model boundaries," students responded that the sand split, moved, or changed shape. When asked, "Would the direction of the plate movement affect the formation of the mountains?" all students responded "yes," because "either they would get higher or lower than they were" or "because when they came together it made a mountain and if it went the other way it would probably make the mountain disappear." Finally, when asked to explain how two land formations were separated from each other, one-half of the students responded "because of an earthquake" and the other half said because "they moved."

As can be seen from these descriptions of student performance, there was little variation in responses, in part because two-thirds of the students did not see the model with which they worked as a way to simulate how plate tectonics has shaped the evolution of the earth. They responded to the questions based on their observations in this isolated case; they did not indicate that earthquakes, mountains, and plate movement, for example, are manifestations of plate tectonics. Further, students did not connect concepts in the word bank to the model they were using; none of the 23 students used any of the words from the word bank to respond to the questions.

Students' prior experience with the concepts could not be differentiated in their responses; students who studied earthquakes in school could not be distinguished from those who did not study earthquakes or those who studied only the procedures to follow when an earthquake occurs. In this situation, the structure of the task required students to follow directions but did not allow students, as was intended by the test developers, to display differential conceptual understanding of geological and geomorphic processes and how they explain the evolution of the earth.

Mismatch (2) between test objectives and observed performance. In a cell respiration task, 12 high school students were asked to "describe the possible forms of energy and types of materials involved in the digestion of a piece of bread and explain fully how they are related" (Lomask et al., 1992). In responding to this task, students are expected to explain how carbohydrates in bread are converted into usable energy and other byproducts through cellular respiration. This knowledge-rich and process-constrained task emphasizes explanation as a way to evaluate the structure of students' knowledge.

Unlike the test developers, students viewed the task as having multiple possible interpretations: "Do you want me to go through the digestive system or do you want like cell respiration or did you want, you know, how the body uses the energy or what it's used for or . . ." In the end, students made a decision about the interpretation they would respond to, perhaps based on what they felt they knew best: "I don't remember like everything about the cell, you know, and all that." Three-quarters of the students interpreted the question to mean they were supposed to describe the route the bread takes during the course of digestion. The following is typical of the response given by these students: "The food must first be broken down with the help of your teeth. Then once in the stomach the acids

break it down even more. After the stomach it enters the intestines and there are more chemicals that break down the food that is to be digested." Contrary to the specific objectives of this assessment, students' responses did not reflect their understanding of cell respiration; they could interpret the problem in ways that bypassed the knowledge they acquired in classroom instruction.

Summary. Analysis of the correspondence between test objectives and observed performance draws attention to those aspects of the situation that elicit appropriate knowledge and cognitive skills. For some situations, the task is structured in ways that maximize opportunities for students to display differential levels of understanding of the content and process skills that are being tested. In other situations, the task places constraints on student performance that encourage uniform responses and not the display of differences in student understanding.

Relationship Between Measured Performance and the Quality of Cognition

In addition to the match between objectives and the performances (i.e., knowledge and process) elicited in assessment situations, there is a relationship between observed performance and what is scored. The fundamental issue in measurement is reducing a set of performances in theoretically and empirically defensible ways. From a validity perspective, students' scores should reflect their level of proficiency with respect to the knowledge, skills, and processes the task is designed to measure. In particular, evidence for cognitive complexity (or perhaps we can use the term *cognitive validity*) is manifest in positive relationships between the nature and extent of cognitive activity and student performance scores such that the performance of high scorers is characterized by cognition qualitatively different from that of low scorers (Baxter et al., 1993; Messick, 1994). In some situations, performance scores may over- or underestimate student competence and subject-matter achievement by giving preference to surface or other easily quantifiable features rather than to the thinking and reasoning activities that signal differential levels of competence. Illustrative examples follow.

Performance scores corresponding to quality of cognition. The "Electric Mysteries" assessment, characterized as knowledge-lean and process-open, asks students to identify the circuit components enclosed in each of six

black boxes (e.g., Shavelson et al., 1991, 1992). Students are presented with two batteries, two bulbs, and five wires to construct and connect circuits to each of six boxes. Two of the boxes have a wire in them. Each of the others contains either a battery and bulb, two batteries, one bulb, or nothing. Scoring is based on identification of the contents of each box and the legitimacy of the circuit constructed to make this identification. Students are given 1 point for the correct answer (e.g., wire) with a corresponding adequate circuit (e.g., battery and bulb), or 0 points if the answer *or* the circuit is correct but not both. The maximum possible score is 6 (1 point x 6 boxes).

Descriptions of the cognitive activity of fifth-grade students (n = 31) while carrying out the "Electric Mysteries" science performance assessment show a correspondence between quality of cognition and performance score (Baxter et al., 1996). In general, students who scored high (5 or 6) on the assessment described a plan consisting of procedures and interpretation of possible outcomes. For example, one student said, "I'll probably start with a battery, a light bulb, and maybe two wires and put them everywhere and that way. . . if the light just shines regular, that will be [the wire] and if it shines really bright, that will be [two batteries] probably, and if it doesn't shine at all, it will be [nothing]." High-scoring students also expressed an adequate understanding of a circuit by incorporating within their explanation the notion that electricity flows in a circular pathway within a closed system. These students demonstrated an efficient, systematic strategy to solving the problem by testing each box first with a bulb and then with a battery and bulb. Further, they engaged in frequent and flexible monitoring; they compared the results of testing with one circuit to those from testing with another circuit, checked the list of possible options to confirm the legitimacy of their conclusions, and recognized inconsistencies in test results leading them to retest a particular box.

In contrast, students with scores of 0 or 1 offered a hypothesis when asked for a plan (e.g., "I think there is a wire in box A"), provided a factual statement when asked for an explanation (e.g., "battery is a source of energy"), and invoked a trial-and-error strategy of "hook something up and see what happens" to guide their problem solving. In monitoring their performance, they relied primarily on their memory of what had happened with other boxes and not on a set of task-related strategies that would provide appropriate feedback.

Students with scores of 2, 3, or 4 demonstrated some understanding of circuits, but their knowledge was not sufficiently structured to sustain high levels of reasoning and thinking throughout the assessment. These students generated plans and explanations that were accurate but incomplete. For example, one plan consisted of naming the equipment and the sequence in which it would be used: "First, I am going to try the wires and the bulbs in all of them [the boxes] to see if it works and if they light, and if it doesn't, then I'll try the battery and the wires." Their procedural strategies and monitoring were generally informative but insufficient to successfully identify the contents of the boxes. These students did not recognize the utility of systematically testing all boxes with a bulb and then testing with a battery and bulb in circuit, nor were they able to engage in a number of monitoring strategies to inform their problem-solving procedures. Rather, they tended to rely on one type of activity, such as rechecking the boxes, without recognizing that other strategies, such as constructing an external circuit, would have been more informative.

In summary, the nature and extent of cognitive activity observed in the "Electric Mysteries" assessment situation was positively related to students' performance scores. Students with high scores could be distinguished from lower scoring students in terms of their plans, strategies, monitoring, and explanations.

Relevant activities elicited but quality overestimated. In the "Critter Museum" task, fifth-grade students are given a collection of plastic replicas of bugs and asked to organize them for a display at the science museum (see Figure 3). This task was designed "to assess students' ability to sort, classify, and state a rationale for their system of classification, using a variety of insect models" (California Department of Education, 1993a). Students in this assessment situation who understand the distinguishing features of insects should classify according to morphology (form and structure) rather than on the basis of other features such as color or size.

CRITTER MUSEUM

As the director of the new science museum, you have decided to set up a display of animals without backbones that are found in the area. To organize your display you need to sort and classify your collection of animals and provide some information about how they have adapted to the area.

Open **Bag A** and spread the 12 animals on the table. Note that each animal has an Identification Number attached to it. Look at the features of each animal. Sort the animals into groups based on your observations. Form at least 2 groups, but not more than 7 groups. Make sure you put every animal into a group. All the animals in a group should be similar to each other in some way.

Use the chart below to describe your groups. On the left side, list the Identification Numbers of all the animals you put in your first group. On the right side, explain how the animals in this group are similar to each other.

Now do the same thing for each of your other groups. Be sure to draw a line between each group.

When you are finished, put an **A** next to your first group, a **B** next to your second group, a **C** next to your third group, and so on until all your groups have a letter next to them.

Figure 3. "Critter Museum" assessment task (adapted from the California Department of Education, 1993a).

In scoring student performance, an important distinction must be made between what Mayr (1976) calls arbitrary and scientific classification systems. Arbitrary and utilitarian classification systems, such as sorting buttons or cataloguing books in the library, are undertaken with a goal to reduce the heterogeneity of the objects into manageable categories; one scheme is not necessarily any better than another. In contrast, a biological classification is explanatory (i.e., shows evolution of insect or plant) and predictive (i.e., new discoveries are incorporated with little modification to the overall system). More specifically, the California Science Framework, in describing what is important for students to learn, clarifies classification by stating that:

> [A]t each level of classification, distinct characters are used to identify the organisms within each group. Groups of organisms are recognized because they share derived characteristics that appeared in their common ancestor and have been passed on.

> These characteristics serve as the basis for diagnosing and classifying groups of organisms. Within each of these groups are other groups that are distinguished by their own unique characteristics. By identifying these unique characteristics, we discover the evolutionary pattern, which is the basis for classification. (California Department of Education, 1990, p. 122)

An examination of the responses of 20 students showed that in general they used an arbitrary rather than a biological classification system. Six students used no common criteria when forming groups. For example, one student formed three groups of animals; one group contained animals with "a lot of feet," one group contained animals "found around the house," and the last group contained animals that "look same." Three students formed groups by deciding whether or not the animals had backbones; these classifications were incorrect because none of the animals had backbones. Students may have been confused by the task instructions that indicate that they were to "set up a display of animals without backbones that are found in the area" (see Figure 3). Seven students applied relevant criteria such as number of legs across two or more groups but did not apply the criteria across all groups. Many students compared two bugs to each other rather than comparing and contrasting each insect to the other 11 insects; the same characteristics were not used as the basis for comparing all insects. For example, one student sorted the insects into five groups as follows: "they have scissor hands, they have 8 legs, they have 6 legs, is wiggly and light, and has a lot of legs."

In scoring student performance, highest scores are to be given to those responses that classify the 12 bugs into 2 to 7 groups with an accompanying, clearly stated rationale for the groups based on *any* attributes other than color and size (see Figure 4). It is not apparent in looking at the scoring criteria that some attributes (other than color and size) are considered more relevant than others for classifying bugs. In addition, there is no indication in the scoring criteria that students should use the same criteria on all bugs to form mutually exclusive groups that facilitate identification with the classification system. Nor is there any indication that students who classify on four characteristics should receive higher scores than students who sort on 1 or 2 characteristics. For example, a student who received a perfect score categorized the insects as follows: group A, they have 8 legs and they are all medium size; group B, they each have 6 legs and they are all small; and group C, they are each large and shiny. Another student who got a perfect score provided the following response:

group A, they could pinch; group B, they have antenna; group C, they have round bug eyes, they crawl, they have things sticking out of their mouth, and it could fly. Three points were awarded for the following response: group A, they have tentacles; group B, they crawl on grass and floors; and group C, they can all fly or if they can't they still have wings.

Score	Score Criteria
	Attributes such as insects, invertebrates, arachnids, wings/no wings, flies vs. crawls, color/size, antenna shape/length, spiders vs. not, harmful/not, rough legs/smooth, feeds on seeds/plants vs. other animals, camouflaged/not, poisonous/not, number of legs, eyes/eye stem, etc.
4	Uses complex attributes beyond color and size. All sorting rationales are clearly stated with descriptions that match animals in each group. All 12 animals from Bag A are sorted into at least 2 groups but not more than 7 groups. Data may include identification numbers of each animal. May include pictures in addition to written rationale.
3	Uses complex or basic attributes for the sorting. Sorting rationales are clearly stated and describe animals in each grouping. At least 8 animals are sorted into at least 2 groups. Data and/or chart may be incorrect. Animals may or may not be identified by individual numbers.
2	Attributes are limited to color or size of the animals. Sorting rationales are unclear or vague and may not match animals placed in each group. Many animals are not listed in a group or may contain more than 7 groups, including one-animal groupings. Chart is incorrectly labeled.
1	Attempts to sort some animals from Bag A into groups. Attributes are unrelated to any characteristics of the animals. Sortings are inappropriate and don't relate to animal characteristics. Rationales are vague, unclear or don't match animal characteristics. Attempts to sort some animals but many are missing. Chart is incorrectly labeled.
0	No response or attempt to sort/classify animals. Rewrites directions. Inappropriate writing or drawing. Writes off-topic.

Figure 4. Scoring criteria for "Critter Museum" task (adapted from California Department of Education, 1993a).

From these examples and an examination of the scoring criteria, it can be seen that high scores were not dependent on knowledge of the distinguishing features of insects or knowledge of the process of scientific classification. Essentially students were permitted to sort on any basis (i.e., arbitrary). If the goal of assessment is to identify the knowledge and process

differences between scores (i.e., subject-matter competence), the ability of students to sort needs to be distinguished from their ability to develop a scientific classification system.

Quality of performance overestimated. Content-rich and process-constrained tasks often appropriately focus on student explanations. In developing an explanation, students draw on their knowledge to decide which concepts are important and why. The Connecticut Common Core of Learning Assessment Project developed concept maps for scoring student explanations of major topics in life, earth, and physical science. These concept maps provide pictorial representations of core concepts and how they are interrelated (Novak & Gowin, 1984). Concepts were considered core if teachers felt students should reasonably be expected to know them at this point in their schooling (Lomask et al., 1992). In other words, the concept maps focus on a subset of knowledge related to a particular topic and not all the possible concepts that could be mentioned.

An expert's (teacher's) concept map serves as a template against which students' performances are evaluated. Students' explanations for each topic are read, and matches and mismatches with the expert concept map are noted. Thus, the concept map provides a visual display of the correspondence between students' performances and "expert" or expected performance. Scoring focuses on two dimensions of the concept map—size and strength. Size is defined as the percentage of the total number of core concepts in the expert concept map that students included in their explanations. Strength is defined as the percentage of possible connections students used in explaining the relationships between core concepts. The strength thus indicates whether students know the "full story" for the concepts mentioned in their explanation; concepts not mentioned are not considered in determining the strength score.

For example, 12 high school students were asked: "For what would you want your blood checked if you were having a transfusion?" (Lomask et al., 1992). As displayed in the concept map used for scoring, responses should contain references to the possibility of an immune reaction (blood compatibility) and acquiring an infection through blood-transmitted diseases such as HIV, syphilis, and hepatitis B (see Figure 5).

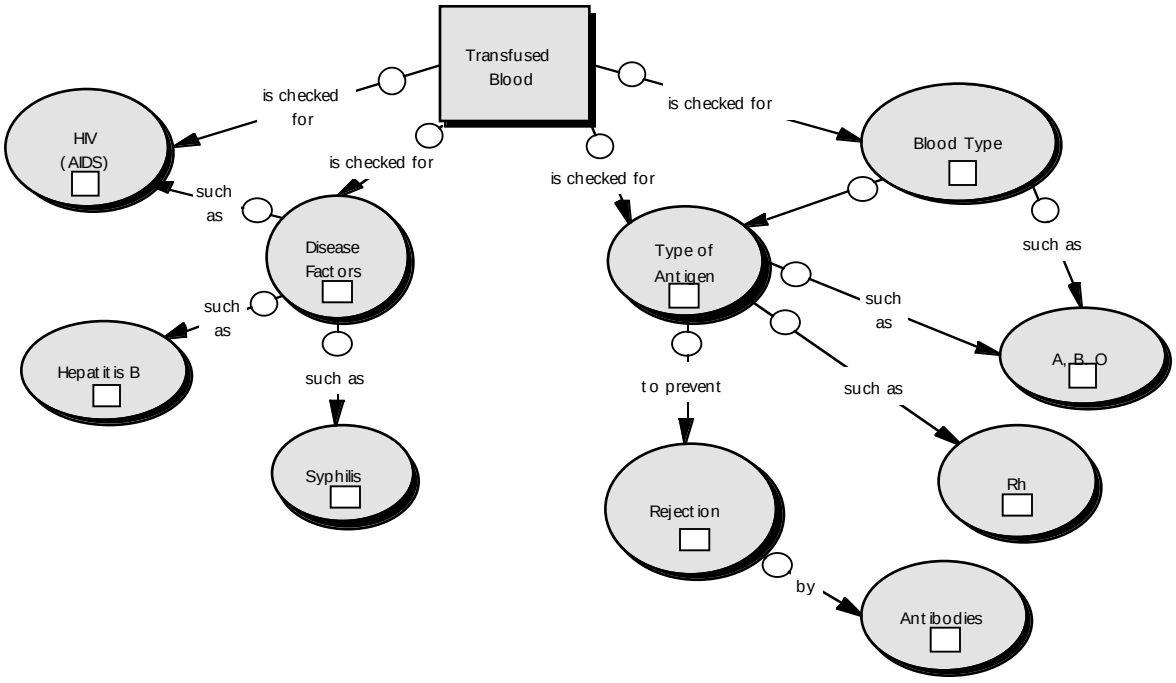

Figure 5. Concept map for scoring "Blood Transfusion" task (adapted from Lomask et al., 1992).

On the surface, concept maps appear to be an excellent way to showcase the differential quality of student responses for teachers and students because they explicitly attend to one of the key distinguishing characteristics of competence: the organization and structure of knowledge. A closer look reveals that this concept map, as currently designed, does not adequately reflect students' understanding of "the reaction between RBC [red blood cell]-surface antigens and naturally circulating antibodies as the basis for blood compatibility" as intended by the test developers (Lomask et al., 1992, p. 11).

For example, one student responded as follows: "Well, first of all I would want it checked to see if the blood was the same blood type as mine was. I would want it checked for diseases." According to the concept map scoring system, this student was credited for 2 of the 10 possible concepts (blood type and disease) and 2 of 2 possible connections ("is checked for"). This student mentioned 20% of the possible concepts and 100% of the connections between these concepts. Consequently, performance was characterized as small (mentions very few concepts) and strong (mentions all connections between them). Although this student provided little more than a list of ideas, the response was characterized

as reflective of strong, interconnected knowledge. Indeed, 80% of the students' responses were judged reflective of strong, interconnected knowledge, despite considerable variation in the number of concepts mentioned in the explanations.

In this assessment situation, there was an apparent discrepancy between test objectives and observed performance with respect to the strength component of the scoring (representing the number of valid connections among core concepts). In part, this overestimate stems from the knowledge assumed in the concept map; one-half of the core concepts are learned in contexts outside science class (HIV, disease, blood type, hepatitis B, and syphilis), and the relations among the concepts are at the level of examples and not processes or underlying causal mechanisms. Notice in Figure 5 that four connections use the term *is checked for* and six use the term *such as*. In the context of this task, students can appear to understand (i.e., integrated knowledge) a considerable amount without ever expressing the interdependent nature of antigens, antibodies, and their relation to the rejection of incompatible blood types (see Mader, 1990; Starr & Taggart, 1989). Unless proficient performance displayed by the concept map requires inferences or reasoning about subject-matter relations or causal mechanisms reflective of principled knowledge, then it serves primarily as a checklist of words and misrepresents (overestimates) students' structure of knowledge.

Improving the Theory and Practice of Achievement Testing

The conceptualization of student competence and subject-matter achievement with respect to the quality of cognition that develops during school learning can now provide a foundation for examining some aspects of the validity of assessment practices. In this context, score use and interpretation should, in part, stem from an analysis of the cognitive activity that is elicited in assessment situations. This paper describes one possible approach to addressing this issue and garnering evidence of the cognitive complexity of current assessment practice.

A framework was presented as a way of discussing levels of performance that make salient the cognitive activities involved. General components of competence in problem solving are described which include problem representation that facilitates planning and anticipation of alternative outcomes, goal-directed strategies that reflect organized knowledge, self-monitoring to control and regulate one's problem-solving activities, and the

explanation of principles underlying performance that can facilitate learning and problem solving. This framework of cognitive activities serves as a guide in evaluating assessments that give primacy to the acquisition and nature of problem-solving competence in subject-matter learning. Further, this framework provides a common language for describing student performance so it can be effectively taught and appropriately assessed.

Because of the special features of science assessments, a content and process space has been introduced that attends to the interaction among content knowledge (declarative and procedural), science process skills, and cognitive activity. This structure represents cognitive task demands in terms of prior knowledge and science process skills requisite for optimal performance in an assessment situation. With this structure, a scaffold is provided that can help anticipate the nature and extent of cognitive activity likely to be observed in particular assessment situations. Recognition of the interrelationships among the subject matter and cognitive features of assessment situations provides a basis for selecting or revising situations to meet specified objectives.

Describing the quality of cognition and content/process demands provides useful tools for identifying performance objectives, examining the correspondence between test objectives and observed cognitive activity, and considering the extent to which performance scores accurately reflect the quality of observed cognitive activity. By focusing on these issues, test developers and users are guided to clarify performance objectives and attend to potential mismatches between what is anticipated and what is observed or between the quality of observed performance and performance scores. The intent is to direct attention to improvements in assessment design.

The approach described here using general features of competence now needs to be refined on the basis of the course of learning in various content domains and an understanding of the goals of assessment practices. The future will require the elaboration of the differential characteristics of developing competence appropriate to various subject matters. Essential in this regard is an iterative working back and forth between theory-based descriptions of developing competence and the art and practice of assessment design. An important outcome of these endeavors will be the improvement of achievement test development and use, and the formulation of alternate conceptions and methodologies for addressing issues of cognitive validity.

This integration of assessment practice and knowledge of learning has been advocated for a long time (e.g., Glaser, 1981). In 1957, Cronbach advised that psychological testing should be brought into closer contact with other areas of psychology. Ten years later, Anastasi (1967) wrote that increasing specialization "has led to a concentration upon techniques for test construction without sufficient consideration of psychological research for the interpretation of test scores" (p. 305). Now, almost 20 years later, it may be possible to remedy this situation.

References

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist*, *22*, 297-306.

Anderson, J. R. (1985). *Cognitive psychology and its implications* (2nd ed.). New York: W. H. Freeman and Company.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*, 1210-1218.

Baron, J. B., Carlyon, E., Greig, J., & Lomask, M. (1992, March). *What do our students know? Assessing students' ability to think and act like scientists through performance assessment.* Paper presented at the annual meeting of the National Science Teachers Association, Boston.

Baxter, G. P., Elder, A. D., & Glaser, R. (1994). *Cognitive analysis of a science performance assessment*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

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., Elder, A. D., & Shavelson, R. J. (1995). *Effect of embedded assessments on performance in elementary science classrooms.* Unpublished manuscript, University of Michigan.

Baxter, G. P., Glaser, R., & Raghavan, K. (1993). *Cognitive analysis of selected alternative science assessments*. Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Breen, T. J., Baxter, G. P., & Glaser, R. (1995, April). *Thinking and reasoning on statewide science assessments: Examples from performance-based assessments in California.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

Brown, A. L. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 1, pp. 77-165). Hillsdale, NJ: Lawrence Erlbaum Associates.

California Department of Education. (1990). *Science framework for California public schools kindergarten through grade twelve*. Sacramento, CA: Author.

California Department of Education. (1993a). *Science grade 5 administration manual*. Sacramento, CA: Author.

California Department of Education. (1993b). *Science grade 8 administration manual*. Sacramento, CA: Author.

Charles, R., & Silver, E. (Eds.). (1988). *The teaching and assessing of mathematical problem solving*. Reston, VA: National Council of Teachers of Mathematics.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Chi, M. T. H. (1994). *Analyzing the content of verbal data to represent knowledge: A practical guide*. Unpublished manuscript, Learning Research and Development Center, University of Pittsburgh.

Chi, M. T. H., Bassock, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*, 145-182.

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

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.

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

Fay, A. L. (1995, March). *Factors affecting the content and structure of children's science explanations*. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Indianapolis, IN.

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

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, *36,* 923-936.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice Hall.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in science and mathematics* (pp. 63-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1996). Changing the agency for learning: Acquiring expert performance. In A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 303-311). Mahwah, NJ: Lawrence Erlbaum Associates.

Glaser, R., Raghavan, K., & Baxter, G. P. (1992). *Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments* (CSE Tech. Rep. No. 349). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Gobbo, C., & Chi, M. T. H. (1986). How knowledge is structured and used by expert and novice children. *Cognitive Development*, *1*, 221-237.

Green, D. (1980). The terminal velocity and dispersal of spinning samaras. *American Journal of Botany*, *67*, 1218-1224.

Halford, G. S. (1993). *Children's understanding. The development of mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 5-21.

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

Mader, S. S. (1990). *Human biology* (2nd ed.). Dubuque, IA: Wm. C. Brown.

Mayr, E. (1976). Evolution and the diversity of life: Selected essays. Cambridge, MA: Belknap Press of Harvard University.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Norberg, R. A. (1973). Autorotation, self-stability, and structure of single-winged fruits and seeds (samaras) with comparative remarks on animal flight. *Biological Review*, *48*, 561-596.

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

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334-370). New York: Macmillan.

Seter, D., & Rosen, A. (1992). A study of the vertical autorotation of a single-winged samara. *Biological Reviews*, *67*, 175-197.

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

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

Siegler, R. S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development*, *59*, 833-851.

Starr, C., & Taggart, R. (1989). *Biology. The unity and diversity of life* (5th ed.). Belmont, CA: Wadsworth.