DOCUMENT RESUME

ED 372 081 TM 021 719

AUTHOR Sugrue, Brenda
TITLE Specifications for the Design of Problem-Solving Assessments in Science. Project 2.1. Designs for Assessing Individual and Group Problem-Solving.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Nov 93
CONTRACT R117G10027
NOTE 63p.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Chemistry; *Educational Assessment; Inferences; Models; Multiple Choice Tests; *Problem Solving; *Research Methodology; *Science Education; *Test Construction; Test Reliability; Test Validity; *Thinking Skills
IDENTIFIERS Hands On Experience; Open Ended Questions; Test Specifications

ABSTRACT

This report describes a methodology for increasing the validity and reliability of inferences made about the problem-solving ability of science students that is based on performance on different kinds of tests. The generalizable cognitive components of problem solving that might be targeted by assessment are described, and specifications are presented for designing multiple-choice, open-ended, and hands-on assessments of problem solving. Some prototype assessments that implement the specifications in the domain of chemistry are presented as examples. Three comprehensive models of the components of problem solving are reviewed to illustrate the range of variables involved and to give a basis for selecting a subset of variables to be targeted by assessment. The specificity of definition of the cognitive constructs of interest in assessment goes beyond vague ideas of understanding and reasoning to identification of specific knowledge types and the links among them and the specific aspects of higher order thinking that have been found to influence problem solving. Nine figures and 17 examples illustrate the discussion. (Contains 138 references.) (SLD)

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – November 1993

Project 2.1 Designs for Assessing Individual
and Group Problem-Solving

Specifications for the Design of Problem-Solving Assessments in Science

Brenda Sugrue, Project Director

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

National Center for Research on
Evaluation, Standards, and Student Testing
Graduate School of Education
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.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Table of Contents

SPECIFICATIONS FOR THE DESIGN OF PROBLEM-SOLVING ASSESSMENTS IN SCIENCE

Brenda Sugrue
CRESST/University of California, Los Angeles

Introduction

This report describes a methodology for increasing the validity and reliability of inferences made about the problem-solving ability of students in the domain of science, based on performance on different kinds of tests. Recent curriculum reform efforts emphasize the development of cognitive skills that generalize across the content of a subject matter domain (California Department of Education, 1990; National Council of Teachers of Mathematics, 1989; National Research Council, 1993). This has necessitated the development of assessments that can tap those generalizable skills. In science, we want to be able to estimate the extent to which a student can engage in higher order thinking across the domain, based on the student's performance on a small sample of activities and content. The movement to hands-on performance-based assessment in science has increased the authenticity or face validity of the tests (Wiggins, 1993), but it has not resulted in reliable identification of a student's ability to engage in generalizable cognitive activities.

Shavelson, Baxter, and Gao (1993) found low correlations between scores on hands-on science assessments that related to different science content. There could be a number of possible reasons for the observed variability in performance across tasks. It is possible that the tasks did not elicit common cognitive skills; or perhaps a particular level of content knowledge is required before higher order cognitive processing can occur; or perhaps the procedure for scoring performance on the tasks was more sensitive to variation in task-specific knowledge and cognitive processes than to variation in task-independent cognitive variables. Baxter, Glaser, and Raghavan (1993) found that, even if a hands-on science assessment task is designed to engage students in higher order thinking, the system for scoring performance may

not be sensitive to differences in understanding and reasoning. Further research is needed to find ways of isolating and scoring the components of performance that generalize across science content and tasks.

Researchers at the Center for Research on Evaluation, Standards, and Student Testing (CRESST) have succeeded in reducing score variability across tasks designed to assess deep understanding of history (Baker, 1992; Baker, Aschbacher, Niemi, & Sato, 1992). This reduction in score variability was accomplished by (a) identifying the critical cognitive dimensions that characterize the generalizable skill of interest (deep understanding) and (b) creating a set of specifications for the design of multiple tasks with common structure and scoring criteria, but different history content.

A similar approach is adopted here to develop a set of specifications for the design of tasks and scoring schemes to measure generalizable aspects of students' ability to solve problems in the domain of science. In addition to supporting the design of multiple tasks with different content, this approach permits diagnosis of the source(s) of poor performance in terms of cognitive weaknesses that can then be targeted by instructional interventions. The creation of multiple tasks that tap the same cognitive structures and processes in the context of different science content will facilitate research on the relative importance and interaction of content-specific and content-independent aspects of problem-solving performance.

A clear conceptualization of the generalizable cognitive constructs to be assessed can be translated into specifications for assessments not only in a variety of content areas, but also in a variety of test formats, from multiple-choice to hands-on (Baker & Herman, 1983; Messick, 1993; Millman & Greene, 1989; Popham, 1993). Shavelson, Baxter, and Pine (1992) found that performance varied across tests that targeted the same science content, but varied in format or method. It may be that if the generalizable cognitive constructs to be targeted by assessment (and antecedent instruction) are operationalized in detailed specifications for assessment design, then students will perform equally well, or equally poorly, across all test formats targeting those constructs.

The creation of multiple tasks in multiple formats, targeting the same underlying cognitive components of problem-solving ability in a domain, will

facilitate research on the validity of the constructs being targeted and the separation of construct-relevant from construct-irrelevant variance (Frederiksen & Collins, 1989; Messick, 1993; Nickerson, 1989; Snow, 1989). Such research could lead to prescriptions for large-scale assessment, classroom diagnostic assessment, and instructional intervention. More efficient test formats, which generate score distributions that correlate highly with performance on more authentic but inefficient formats, could be prescribed for use in large-scale assessment systems, while the more authentic or "benchmark" (Shavelson et al., 1992) formats could be implemented primarily as instructional activities that serve to develop the cognitive components of interest.

This report contains

1. a description of the generalizable cognitive components of problem solving that might be targeted by assessment;
2. specifications for designing multiple-choice, open-ended, and hands-on assessments of problem solving;
3. some prototype assessments that implement the specifications in the domain of chemistry.

The framework presented here is one of a number of possible approaches. This framework emphasizes the diagnostic function of assessment, and the need to define performance in terms of cognitive components that can be measured in a variety of ways.

Cognitive Components of Problem-Solving Performance

A review of the literature on the cognitive variables associated with problem-solving ability was undertaken to identify a set of cognitive components that might be measured to estimate the extent to which a student can solve problems within a subject matter domain, such as science. There is a large body of research on problem solving. The research spans learning and performance in knowledge-rich domains such as school subject matter, medicine, technical occupations, and in knowledge-lean puzzle domains (Anderson, 1993; Greeno & Simon, 1989). Because different studies focus on different variables that might influence problem solving, and because few studies examine the relative importance or interactions among such variables,

it is difficult to piece together a definitive list of the cognitive variables associated with problem solving. However, a number of comprehensive models of the components of problem solving have been proposed; these models are based on review and compilation of results from various strands of research. Three of those models (Glaser, Raghavan, & Baxter, 1992; Schoenfeld, 1985; Smith, 1991) will be described here to illustrate the range of variables involved and also to provide a basis for selecting a subset of variables to be targeted by assessment.

Model 1: Glaser, Raghavan, and Baxter, 1992

This approach represents the latest version of a model that has been suggested and refined by Glaser, Chi and their colleagues over the past decade (Chi & Glaser, 1985; Chi, Glaser, & Rees, 1982; Glaser, 1992; Chi & Glaser, 1988). The model is primarily based on the results of research in the expert-novice paradigm, that is, research that documents differences between the performance of experts and novices on knowledge-rich tasks, such as those in mathematics or physics. Glaser's model describes the following five components of problem solving:

1. **Structured, integrated knowledge:** Good problem solvers use organized information rather than isolated facts. They store coherent chunks of information in memory that enable them to access meaningful patterns and principles rapidly.

2. **Effective problem representation:** Good problem solvers qualitatively assess the nature of a problem and build a mental model or representation from which they can make inferences and add constraints to reduce the problem space.

3. **Proceduralized knowledge:** Good problem solvers know when to use what they know. Their knowledge is bound to conditions of applicability and procedures for use.

4. **Automaticity:** In proficient performance, component skills are rapidly executed, so that more processing can be devoted to decision-making with minimal interference in the overall performance.

5. **Self-regulatory skills:** Good problem solvers develop self-regulatory or executive skills, which they employ to monitor and control their performance.

Model 2: Schoenfeld, 1985

Schoenfeld calls his model a "framework for analysis of complex problem-solving behavior" (1985, p. xii). Schoenfeld's model is supported primarily by results of a series of empirical studies on the effects of instruction (targeted at a variety of cognitive variables thought to influence problem solving) on mathematics learning and performance. However, as Schoenfeld indicates, research from many other fields also supports and informs aspects of the model. Although his model is defined for the domain of mathematics, any other content domain could be substituted. Schoenfeld's model has four general categories of variables that are related to problem solving: resources, heuristics, control, and belief systems, each of which he defines (for the domain of mathematics) as follows:

1. **Resources:** Mathematical knowledge possessed by the individual that can be brought to bear on the problem at hand; intuitions and informal knowledge regarding the domain; facts; algorithmic procedures; "routine" nonalgorithmic procedures; understandings (propositional knowledge) about the agreed-upon rules for working in the domain.

2. **Heuristics:** Strategies and techniques for making progress on unfamiliar or nonstandard problems; rules of thumb for effective problem solving, including: drawing figures; introducing suitable notation; exploiting related problems; reformulating problems; working backwards; testing and verification procedures.

3. **Control:** Global decisions regarding the selection and implementation of resources and strategies; planning; monitoring and assessment; decision-making; conscious metacognitive acts.

4. **Belief systems:** One's "mathematics world view," the set of (not necessarily conscious) determinants of an individual's behavior; about self; about the environment; about the topic; about mathematics.

Model 3: Smith, 1992

Smith's model differentiates between internal and external factors that are thought to affect problem-solving performance, and between good and expert problem solving. The distinction between good and expert problem solving reflects a concern that the conclusions of expert-novice studies are based too heavily on the performance of experts for whom the "problems" solved may not have been novel enough to elicit the kind of problem-solving processes used by less-than-expert, yet successful, performers. Novices often

successfully solve problems, but their solution processes are not the same as those of experts (Smith & Good, 1984). Smith (1992) suggests that expert problem solving is merely a subset of successful problem solving, and that the goal of education in academic settings is "to produce successful problem solvers and not 'experts' as such" (p. 11). Therefore, we need to have a model of the characteristics of good problem solvers who are not highly experienced professionals.

The internal factors included in Smith's model are the most relevant to the present discussion of the cognitive components of problem solving; therefore, only that part of his model is presented here:

> **Affect**. Good problem solving is enhanced by certain affective variables, including self-confidence, perseverance, enjoyment, positive self-talk, motivation, beliefs, and values.
>
> **Experience**. Good problem solving is enhanced by the length of prior successful problem-solving experience (especially in the domain of the problem).
>
> **Domain-specific knowledge**. Good problem solving requires knowledge of the domain from which the problem is drawn. This knowledge is of three types: factual, conceptual or schematic, and procedural. The problem solver's knowledge must be: adequate, organized, accessible, integrated, and accurate (misconception free).
>
> **General problem-solving knowledge**. Good problem solving is enhanced by knowledge of general problem-solving procedures such as means-ends analysis, trial and error, etc.
>
> **Other personal characteristics**. Problem solving success is also affected by the solver's level of cognitive development, relative field dependence, personality, etc.

Smith proposes that good problem solvers (regardless of level of expertise) tend to

1. adapt their knowledge and its organization to facilitate the solution of problems in a domain;
2. apply their knowledge and skills to the problem-solving task;
3. use forward reasoning and domain-specific procedures on standard problems within their domain of expertise, but use "weaker" problem-solving procedures (means-ends analysis, trial-and-error, etc.) on problems outside of their domain of expertise;

4. create an internal "problem space" which incorporates a qualitative representation and redescription of the problem;
5. plan (at least tacitly) the general strategy or approach to be taken (depending on the perceived complexity of the problem);
6. break problems into parts and perform multi-step procedures when necessary, keeping the results of previous steps in mind;
7. employ relevant problem-solving procedures/heuristics—both domain-specific and general;
8. evaluate the solution and the solution procedure; and
9. abstract patterns in their own performance (identify powerful solution strategies) and identify critical similarities among problems (identify useful problem types).

Summary and Implications of the Models

The models just described contain many variables. One way to categorize them would be to use Rumelhart and Norman's (1989) distinction between variables that relate to the structure of knowledge in memory and variables that relate to the cognitive functions that operate on that knowledge to assemble, control and monitor the execution of a solution to meet the demands of an unfamiliar task. It is widely acknowledged that problem solving involves the interaction of knowledge and cognitive function (Alexander & Judy, 1988; Chi, 1985; Peverly, 1991). The addition of a third category of cognitive constructs that relate to motivation/attitudes/beliefs is necessary to account for differences in problem solving that result from perceptions or beliefs about oneself and the task (McCombs, 1988; McLeod, 1985; Snow, 1989).

The assumption made here is that the ability to solve problems in a particular domain results from the complex interaction of knowledge structure, cognitive functions, and beliefs about oneself and about the task. Observed differences in problem solving, from interpretation of the problem to persistence in attempting to solve it, can be attributed to variation in aspects of these three cognitive constructs. Therefore, any attempt to generate a profile of the problem-solving ability of a student would need to include aspects of these three elements of cognition. This three-part model of problem-solving performance reflects Snow's (1989) conclusion that errors in performance occur when a person's "previously stored cognitive components and knowledge

base are inadequate, or poorly applied, the improvisational assembly and action-control devices are weak because they are not geared to the specific task type at hand, or achievement motivation flags prematurely" (pp. 48-49).

It would be impossible to measure at one time all of the variables that relate to each of the three categories of cognition that affect problem solving. The criteria used here for selection of a subset of variables are that research should have indicated that the variables are critical, or, if it is not clear which of a number of variables are critical, that the variables selected should be open to instructional intervention (this criterion was suggested by Snow, 1990). For each of the three categories of cognitive components, the variables selected to be targeted by assessment will now be described. The complete model of the cognitive constructs to be assessed is presented in Figure 1.

Knowledge Structure

Many researchers (for example, Glaser, 1984, 1990, 1992; Marshall, 1988; 1993) describe the structure of good problem solvers' knowledge (sometimes called schemas or mental models or conceptual knowledge) as connected, integrated, coherent, or chunked. In contrast, the knowledge of poor problem solvers is deemed to be fragmented and unconnected. The more connected one's knowledge, the more knowledge is activated when one piece of the network is activated or triggered by information presented in a problem (Anderson, 1983; Gagné, Yekovich, & Yekovich, 1993). The knowledge of good

Knowledge Structure	**Cognitive Functions**	**Beliefs**
1. concepts	1. planning	1. perceived self-efficacy (PSE)
2. principles (links among concepts)	2. monitoring	2. perceived demands of the task (PDT)
3. links from concepts and principles to conditions and procedures for application		3. perceived attraction of the task (PAT)

Figure 1. Cognitive components of problem solving to be assessed.

problem solvers seems to be organized around key principles, and related concepts, which are linked to conditions and procedures for implementation (Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Farr, 1988; Chi, Glaser, & Rees, 1982; Glaser, 1992; Greeno & Simon, 1989; Larkin, 1983; Mestre, Dufresne, Gerace, Hardiman, & Tougher, 1992; Schultz & Lochhead, 1991).

Chi et al. (1981) found that physics experts interpret problems in terms of the principles that would guide their solution, whereas novices focus on superficial features of the problem. Chi et al. (1981) also found that the knowledge structures of these experts contained more physics principles, and more linking of these principles to methods and conditions for applying the principles. A study by Mestre et al. (1992) found positive effects of training in principle-based reasoning on problem-solving performance. Larkin's (1983) research on the cognitive activities of experts and novices in physics indicates that the level of difficulty of a physics problem depends on the number of principles that have to be coordinated in order to interpret and solve it.

The definitions of principles and concepts adopted here are based on the content dimension of the content/performance matrix developed by Merrill (1983) for classifying instructional outcomes. However, these definitions reflect the more recent literature on concept learning (for example, Klausmeier, 1992) and principle-based performance (for example, Larkin, 1983). A principle is defined as a rule, law, formula, or if-then statement that characterizes the relationship (often causal) between two or more concepts. For example, the economic principle governing the relationship between the concepts of supply and demand, or the scientific principle that describes the relationship between the concepts of force and motion. Principles can be used to interpret problems, to guide actions, to troubleshoot systems, to explain why something happened, or to predict the effect a change in some concept(s) will have on other concepts. (de Kleer, 1983; Gentner & Stevens, 1983; Glaser, 1984; Merrill, 1983).

Understanding of a principle assumes understanding of the concepts that are related by the principle. A concept is a category of objects, events, people, symbols, or ideas that share common defining attributes or properties, and are identified by the same name. For example, energy, temperature, heat, and light are scientific concepts, scientist is a concept, assessment is a concept. All concepts have definitions that can be expressed in terms of the attributes or

properties that all instances of the concept share. Understanding of a concept facilitates identification or generation of examples of the concept.

To facilitate problem solving, concepts and principles must be linked to conditions and procedures to facilitate their use in unfamiliar situations. A procedure is a set of steps that can be carried out to achieve some goal. Conditions are aspects of the environment that indicate the existence of an instance of a concept, or that indicate that a principle is operating or might be applied, or that a particular procedure is appropriate. Good problem solvers should be able to recognize situations where a principle is operating; they should also be able to recognize situations where procedures can be performed to identify or generate instances of a concept; and they should be able to carry out those procedures accurately. Good problem solvers should be able to assemble a procedure based on a principle to engineer a desired outcome in an unfamiliar situation.

Diagnostic assessment of problem-solving ability should permit identification of students who understand the concepts but not the principle that links them, students who understand the concepts and principles but lack knowledge of procedures to apply them, and students who can perform procedures correctly but do not know when it is appropriate to apply them. Therefore, three aspects of domain-specific knowledge can be distinguished and targeted by assessment of problem solving: understanding of concepts, understanding of the principles that link concepts, and linking of concepts and principles to application conditions and procedures.

Cognitive Functions

The kinds of cognitive functions that support the flexible adaptation of one's knowledge to meet the demands of an unfamiliar problem are referred to in the literature as metacognitive functions (Brown, Bransford, Ferrara, & Campione, 1983), or higher order thinking processes (Baker, 1990; Kulm, 1990), or assembly and control functions (Snow, 1980; Snow & Lohman, 1984). These cognitive operations are associated with fluid ability (Snow, 1980; Lohman, 1993). Specific aspects of cognitive functioning that have been isolated, and that might be open to assessment, include "planning problem-solving approaches, seeking additional information, searching for and using analogies, and monitoring progress" (Campione & Brown, 1990, p. 148);

planning, monitoring, selecting and connecting (Corno & Mandinach, 1983); planning, monitoring and evaluating (Sternberg, 1985); and "knowing when or what one knows or does not know, predicting the correctness or outcome of one's performance, planning ahead, efficiently apportioning one's time, and checking and monitoring one's thinking and performance" (Glaser, Lesgold, & Lajoie, 1987, p. 49).

Good problem solvers spend a disproportionate amount of time in the initial planning phase of problem solving (Campione & Brown, 1990; Gagné et al., 1993; Glaser, 1992; Voss & Post, 1988), and select and follow one solution path rather than trying out a number of solutions (Larkin, McDermott, Simon, & Simon, 1980). Studies by Schoenfeld (1982, 1985) indicate that problem-solving performance in the domain of mathematics can be improved by acquisition of metacognitive strategies taught in the context of the domain. Cook and Mayer (1988) improved problem-solving performance by training students to select and organize text-based information. Lewis (1989) improved problem-solving performance in mathematics through training students in a strategy for translating sentences into diagrams to represent mathematics word problems.

There is no research on the relative importance of the individual components of cognitive functioning that have been suggested. The components that have most often been singled out for assessment or training are planning and monitoring; therefore these two cognitive functions are included in this model of the set of variables to be targeted in the assessment of problem solving. Planning is defined here as thinking through what one will do before actually doing it. Monitoring is defined, in this model, as keeping track of a number of aspects of one's performance, including time, the effects of one's efforts in relation to the goal and constraints of the problem, and adapting one's strategy if necessary.

The cognitive function of connecting, that is, linking incoming information to familiar information (Corno & Mandinach, 1983), or the creation of new links among existing knowledge structures (Clark & Blake, in press) may be independent of the connectedness of one's existing knowledge, and therefore may warrant separate assessment. However, given the complexity of that cognitive skill, its relationship to analogical reasoning

(Keane, 1988), and the need for more research on this cognitive function, it was decided not to include it in the model presented here.

Beliefs About Self and Task

Although most empirical studies of problem solving have not measured affective/motivational variables, there is increasing acknowledgment of the role of such variables in problem-solving performance (Resnick, 1989; Seegers & Boekaerts, 1993; Silver, 1985; Snow, 1989), and in test performance in general (O'Neil, Sugrue, Abedi, Baker, & Golan, 1992; Snow, 1993). Some theorists combine motivational variables with metacognitive variables in models of "self-regulated learning" (Zimmerman, 1986) or "cognitive engagement" (Corno & Mandinach, 1983), reflecting the complex interaction among all of these variables. The literature on motivation and learning contains a host of psychological constructs (see Snow and Jackson, 1993, for a comprehensive catalogue of these variables).

The set of constructs and model adopted here is that beliefs about one's competence (Bandura, 1982; Boekaerts, 1991) interact with beliefs about the demands (Boekaerts, 1991) and attraction (Boekaerts, 1987; Feather, 1982) of the task to influence effort expenditure (Salomon, 1983) and persistence in the face of difficulty (Dweck & Licht, 1980; Schunk, 1984). Assessment of beliefs about competence or self-efficacy (PSE), demands of the task (PDT), and attraction of the task (PAT) are incorporated in the specifications described later in this document.

Methods for Measuring Cognitive Variables

All of the constructs (in Figure 1) selected to be the focus of problem-solving assessment are "cognitive" in the sense that they exist in people's minds and cannot be measured directly. Indirect methods must be found to indicate a student's knowledge structure, cognitive functioning, and beliefs. There is a growing literature on methodologies for indirectly measuring cognitive constructs (Benton & Kiewra, 1987; Ericsson & Simon, 1984; Garner, 1988; Glaser, Lesgold, & Gott, 1991; Glaser et al., 1987; Lamon & Lesh, 1992; Lohman & Ippel, 1993; Marshall, 1988, 1990, 1993; Royer, Cisero, & Carlo, 1993; Snow & Lohman, 1989; Snow & Jackson, 1993; Tatsuoka, 1990, 1993). The methodologies involve analysis of data based on verbal protocols, videotaped or

computer-generated records of performance, patterns of errors, response patterns across sets of items, teachback protocols, notetaking, eye movements, response latencies, concept maps, sorting and ordering tasks, similarity rating, relative time allocations during task performance, solution time patterns, order of recall, structured interviews, self-assessment questionnaires, summaries, and explanations (written or oral). Many of the methods were developed by psychologists for investigating cognitive theories of learning and memory; the goal now is to adapt these methods for use in educational assessment (Glaser et al., 1987).

Since the validity of many of the methodologies either has not been demonstrated or has been questioned (Garner, 1988; Nisbett & Wilson, 1977; Royer et al., 1993; Siegler, 1989), one should obtain multiple indicators of each construct of interest (Ericsson & Simon, 1984; Norris, 1989; Snow & Jackson, 1993). For each of the cognitive component variables selected for assessment (see Figure 1), a number of methods for measuring those variables will now be described. Some of the methods could be implemented in a variety of test formats; others may require open-ended formats to elicit written or oral responses; yet others may require observation of actual performance on hands-on tasks. Snow (1993) suggests combining scores based on responses to written test formats, such as multiple-choice or open-ended, with data from think-aloud protocols or interviews. A number of the methods described below will be included and elaborated in the set of specifications for designing assessments of the cognitive components of problem solving. However, before describing the methods selected, a brief overview is presented of methods that can be and have been used to measure constructs in each of the three categories included in the model of problem solving adopted here (knowledge structure, cognitive functions, and beliefs).

Assessment of Knowledge Structure

Regardless of the format (multiple-choice, open-ended, or hands-on) used to test knowledge structure, a "problem-solving" test should focus on the extent to which the individual's content knowledge is organized around key concepts and principles that are linked to application conditions and procedures. Knowledge of concepts can be assessed by asking students to classify or generate examples of the concepts (Clark, 1990; Gagné et al., 1993; Hayes-Roth

& Hayes-Roth, 1977; Merrill, 1983; Tennyson & Cocchiarella, 1986); or by noting how many times a student mentions a concept and links it to other concepts during an explanation of some event or process (Baker et al., 1992; Baxter et al., 1993). Methods that have been used to measure knowledge of principles include problem sorting (Chi et al., 1981), and variations of problem sorting where students have to select a problem that involves the same principle as another problem (Hardiman, Dufresne, & Mestre, 1989; Marshall, 1988), or an open-ended format where the student is asked to explain why a number of problems are similar (Chi et al., 1982). Asking students to explain why something occurred, or why they did something during problem solving (Glaser et al., 1991; Marton, 1983), or to predict the outcome of a given situation, are other methods to elicit information from which one can infer students' knowledge of principles (Royer et al., 1993). Accurate interpretation (also called representation or initial understanding) of problems is also an indication of principle-based organization of knowledge in memory (Chi et al., 1981; Chi & Glaser, 1985; Glaser et al., 1987).

Methods for assessing links between concepts or principles and conditions and procedures for applying them include asking students to select or suggest a method for solving a problem (Chi et al., 1981; Chi et al., 1982; Marshall, 1988; Ronan, Anderson, & Talbert, 1976); to debug a solution (Adelson, 1984; Marshall, 1988); to suggest the ordering of procedures or steps in a procedure given a particular set of conditions (Glaser et al., 1991); to think aloud as they attempt solution (Chi et al., 1982; Glaser et al., 1991); or to explain why they used a particular strategy or procedure (Baxter et al., 1993). Procedures for identifying or generating instances of concepts are important in the domain of science, where many tasks involve testing substances or objects in order to classify them. For example, students can be asked to determine the identity of some unknown substance, or students can be asked to create a substance that has a particular set of properties. These kinds of tasks require knowledge of procedures that are linked to the concepts (categories) to which the substances or objects belong. Tasks that require knowledge of procedures that are linked to principles go beyond identification or generation of substances or objects with particular defining properties. Tasks requiring knowledge of principle-related procedures are tasks that require selection or application of procedures

to modify some aspect of a situation that will result in a desired outcome (change in a related concept).

Assessment of Cognitive Functions

There are a number of approaches to assessing cognitive functions. Tests of fluid ability such as Raven's Progressive Matrices provide domain-general estimates of cognitive functions (Snow & Lohman, 1989). Self-assessment questionnaires such as those of Zimmerman and Martinez-Pons (1986), O'Neil et al. (1992) and Pintrich and DeGroot (1990) provide data on the extent to which students perceive themselves to be engaging in a number of distinct metacognitive functions. Think-aloud protocols and retrospective interviews, using videotapes of performance to stimulate recall, have also been used to assess cognitive functions (Gillingham, Garner, Guthrie, & Sawyer, 1989; Peterson, Swing, Breverman, & Buss, 1982; Siegler, 1988; Swing, Stoiber, & Peterson, 1988).

Studies that have attempted to assess students' domain-specific planning skills have used measures such as asking students to demonstrate how they planned to solve particular problems and to justify their plans (Marshall, 1993); to recreate a plan based on a completely executed solution (Gerace & Mestre, 1990); to think aloud as they planned how they would solve a problem (Campione & Brown, 1990; Lesgold, Lajoie, Logan, & Eggan, 1990); and prompting students to describe their plans at different points during problem solving (Glaser et al., 1987). Relative proportions of time devoted to planning and execution have also be used as measures of planning (Chi et al., 1982).

Monitoring has been assessed via think-aloud methods (Hayes & Flower, 1980); observation of student performance to identify extent to which students look back over elements of the material presented or elements of their solutions (Garner & Reis, 1981); comparison of solution speeds under different conditions (Harris, Kruithof, Terwogt, & Visser, 1981); noticing of inadequate instructions (Markman, 1979); the advice a student would give to another student before a test (Smith, 1982); students identifying examples of good and poor monitoring from descriptions of other students' behavior (Snow, 1989); and time allocations during performance (Wagner & Sternberg, 1987). Snow and Jackson (1993) suggest a number of methodologies that might be used for

assessing monitoring skills, including the extent to which students keep track of time remaining and adjust their plans and strategies accordingly.

Assessment of Beliefs

Beliefs about one's competence and about the demands or attractiveness of a task are usually measured via questionnaires or interviews. Numerous interview schedules (for example, Zimmerman & Martinez-Pons, 1986) and self-assessment questionnaires (for example, Bandura, 1989; Boekaerts, 1987; Feather, 1988; O'Neil et al., 1992; Pintrich & De Groot, 1990; Stipek, 1993) have been developed to tap perceptions/attitudes/beliefs. Sometimes students are presented with task scenarios and asked how they would respond in that situation (Zimmerman & Martinez-Pons, 1990). Most of the time, students are asked to rate how well a particular statement reflects their beliefs. For example, Feather (1988) used the following item to measure students' beliefs about their mathematics ability: "In general, how do you rate your ability to do well in mathematics?" Students had to indicate their rating on a 7-point scale ranging from very low on one end to very high on the other end. One of Feather's (1988) items to measure subjective valence of mathematics was "How interested are you in mathematics?" Students responded on a 7-point scale ranging from not interested at all to very interested. Boekaerts (1987, 1991) has developed an instrument that asks students to respond on a 5-point scale to items such as "How much do you like these kinds of tasks?" and "How eager are you to work on this kind of task?" to measure variables such as task attraction, perceived difficulty, and perceived competence.

Few methodologies for eliciting and scoring either open-ended responses to questions about beliefs or behavioral indicators of beliefs have been developed. Snow (1989, 1990) describes an open-ended approach to eliciting beliefs that may be more valid than fixed-format inventories. Chi et al. (1982) used an open-ended technique in which they asked students to indicate the aspects of a task that made them judge it as difficult.

Selection of Assessment Techniques for Measuring Cognitive Components of Problem Solving

Before specifications for the design of assessments to target the selected cognitive components of problem solving (see Figure 1) can be described, a

subset of assessment techniques must be selected that will be used to provide multiple indicators of those components. There is little research to guide the selection of the most appropriate methodology for measuring problem solving. Mathematics word problems or hands-on science tasks have intuitive appeal as authentic measures of problem solving in the domain. However, performance on such tasks has been found to be sensitive to even minor changes in the way a task is presented (context, structure, length, vocabulary, syntax), as well as to changes in type of response required from the student (Bennett & Ward, 1993; Goldin & McClintock, 1984; Messick, 1993; Millman & Greene, 1989; Snow & Lohman, 1989; Webb & Yasui, 1992). In addition, the effects of differences in task stimulus characteristics and response formats are not the same for all students (Kilpatrick, 1985; Snow, 1993).

The methodologies recommended in the assessment design specifications presented in the next section of this report will be categorized according to the type (format) of student response they demand (multiple-choice, open-ended, or hands-on) as opposed to the format in which information is presented to the student. The reason for focusing on response format rather than stimulus format is that it is not at all clear what aspects of stimulus format are most critical in relation to the cognitive constructs to be assessed. Some general recommendations will be made about task structures only as they relate to the elicitation of responses that can be scored and interpreted in terms of the cognitive constructs of interest.

The limitations of each type of test response format (multiple-choice, open-ended, or hands-on) are not clear; most of the arguments for and against different formats are based on their face validity rather than on their construct validity (Bennett, 1993; Bridgeman, 1992; Nickerson, 1989; Snow, 1993). While multiple-choice formats do not provide opportunities for students to explain why they made a particular choice, underlying cognitive processes and structures may be inferred from patterns of responses across sets of items that were constructed to reveal weaknesses in underlying reasoning (Tatsuoka, 1990; 1993) and knowledge schemas (Marshall, 1993).

Baker and Herman (1983) advocate treating format as a separate dimension of assessment, in order to investigate its separate contribution to performance score variance. Messick (1993) suggests using a construct X format matrix to guide test design. The format dimension should represent a

"choice-to-construction continuum" (Messick, 1993, p. 66) based on the amount of constraint or degree of openness entailed in the student's response to the test item. Messick (1993) suggests that the format dimension might also include blends of formats such as multiple-choice with justification of choice or hands-on plus written justification of actions or results of actions. Blends of formats are becoming more prevalent in large-scale testing. Current manifestations of hands-on assessment in science ask students to write answers to questions as they perform the hands-on task, but only the written responses are examined and scored.

Snow (1993) describes a provisional 8-level format continuum that goes from multiple-choice at one end to long essay/demonstration/project and collections of multiple assessments over time at the other end. Messick (1993) proposes that some cells of a construct X format matrix may be empty, indicating constructs that cannot be directly tapped by certain formats, although it may be possible to predict or estimate any construct with any format. The position adopted here is that all formats can be used to provide information on all of the cognitive constructs selected as the components of problem solving to be assessed. However, some formats, particularly hands-on tasks, are more authentic than others or can be used to measure a number of components at one time. The following construct X format matrix (Figure 2) represents the constructs and formats included in the design specifications that are described in the next section of this report. Assessment design specifications will be presented for each cell in this matrix and for combinations involving open-ended with either multiple-choice or hands-on formats.

Specifications for Designing Problem-Solving Assessments in Science

Specifications for the design of multiple assessment strategies to target the constructs identified as critical to problem solving will now be outlined. The function of the specifications is to standardize the behavior of assessment designers, to increase the comparability of tasks and generalizability of performance on them, and ultimately to enhance the validity of test-score inferences (Baker et al., 1992; Millman & Greene, 1989; Popham, 1984). Hively (1974) suggests that specifications should include directions for presenting the

Construct	Format		
	Multiple-choice	Open-ended	Hands-on
Knowledge Structure Concepts			
Principles			
Links from concepts and principles to conditions and procedures for application			
Cognitive Functions Planning			
Monitoring			
Beliefs PSE			
PDT			
PAT			

Figure 2. Construct X format matrix for assessment of problem solving components.

item stimulus, recording the student's response, and deciding how appropriate the response is. Millman and Greene (1989) present a longer list of test attributes to be specified, including external contextual factors such as characteristics of the examinee population or how the test will be administered, as well as internal attributes of the test itself. Baker et al. (1992) include specifications for the training of raters in addition to specifications to control the cognitive demands of the task, the structure of the task, and the generation and application of scoring rubrics.

The specifications presented here will focus on each of the three construct categories (knowledge structure, cognitive functions, and beliefs) separately. Specifications for designing tasks (in each of the three formats) to measure the important components of knowledge structure will be presented first. Then specifications for modifying those tasks or generating additional tasks to target cognitive functions and beliefs will be outlined. Finally, specifications for scoring performance on each type of task will be outlined. The specifications will be illustrated at each point for a small subdomain of chemistry (solution chemistry).

Specifications for Designing Tasks to Assess Knowledge Structure

According to Millman and Greene (1989), the most important attribute of a test to be specified is its content. Content analysis must be driven by some conceptualization of the knowledge and performance to be assessed, such as a set of instructional objectives or a set of cognitive dimensions (Millman & Greene, 1989; Nitko, 1989; Popham, 1993). The specifications for content analysis described here focus on the knowledge structure constructs included in the model of problem solving that was presented earlier in this report; that is, content must be analyzed in terms of concepts, principles, and their links to conditions and procedures for application.

Specifications for Content Analysis for Assessment of Knowledge Structure

If assessment is to focus on knowledge of concepts, principles that link those concepts, and conditions and procedures for applying those concepts and principles, then the test designer's first task is to identify the concepts, principles, and related conditions and procedures to be assessed. The approach advocated here can be applied to content domains of any size, from

narro · science topic areas such as sound or electricity to larger domains of science such as energy or the chemistry of matter. What constitutes the task domain of interest depends on the scale and purpose of the assessment.

The curricular domain of science is a large one. It is generally divided into the subdomains of life sciences, earth sciences, and physical sciences. Each of these subdomains is further subdivided into topic areas. For example, the physical sciences domain is subdivided into the following areas in the California science curriculum framework (California State Department of Education, 1990): matter, reactions and interactions, force and motion, and energy (sources and transformations, heat, electricity and magnetism, light, and sound). In the most recent draft of the National Science Education Standards (National Research Council, 1993), the main categories of "fundamental understandings" in the physical sciences are also the chemical and physical properties of matter, energy, and force and motion. In the California science curriculum framework, teachers are encouraged to focus on themes or unifying ideas that cut across traditional topics.

The slice of science content to be assessed should reflect the combination of topics and themes that students were exposed to in their curriculum. Therefore, a classroom teacher might analyze the content that he or she covered in a week, month, or year, depending on the scope of the assessment. State assessment might focus on the science content emphasized in the state curriculum framework, which presumably is the content that guides the curriculum taught in all schools in the state. The content domains emphasized in state curriculum frameworks presumably reflect the national view of what students should learn about science in school. Before embarking on the detailed analysis of the domain or subdomain of interest, one must decide the grade level to be targeted by the assessment. The grade level will determine the level of technicality of the definitions of concepts and principles. For example, in chemistry, usually it is not until high school that definitions involving chemical formulas or molecular composition are used.

Although the concepts and principles to be assessed at the classroom level may at times be less general than those targeted by statewide or national assessment, the more general concepts and principles should also be assessed by the classroom teacher. However, regardless of the level of generality of the concepts and principles to be assessed, the same approach to content analysis

can be adopted. For larger domains, the results of a content analysis will provide a clear depiction of the domain to which performance should generalize, and will permit sampling of content that relates to each of the cognitive components to be targeted by assessment; for example, if 20 concepts are identified, then a sample of them can be selected for assessment.

The resources one needs to do a content analysis are a set of the most current textbooks on the domain of interest and access to a number of teachers of that subject matter. The assumption here is that the test designer is someone who is not familiar with the content domain and so will be in a better position to extract and categorize the appropriate content from the textbooks and subject matter teachers. This assumption is based on evidence, from cognitive task analysis for job training, that subject matter experts do not develop complete enough descriptions of their own knowledge (Cooke, 1992; Glaser et al., 1991). However, teachers may be trained to do their own content analysis.

The test designer should read the sections of a number of textbooks that relate to the domain to be assessed. As one reads, one should use the following forms (Figures 3, 4, and 5) to compile and categorize the content that will be used to create test items or tasks. A completed set of forms for the narrow domain of solution chemistry follows the blank sample forms. The designer should keep in mind the definitions of concepts, principles and procedures presented earlier in this report. A concept is defined as a category of objects, events, or ideas that share a set of defining attributes. A principle is defined as a rule that specifies the relationship between two or more concepts. A procedure is a set of steps that can be carried out either to classify an instance of a concept (for example, a test to identify the pH level of a liquid) or to change the state of one concept to effect a change in another (for example, a set of actions to change the pH level of a lake in order to counteract the effects of acid rain). Conditions are aspects of the environment that indicate the existence of an instance of a concept, and/or that a principle is operating or can be applied, and/or that a particular procedure is appropriate.

Content/topic area: ID code:

Important concepts (names only):

Procedures/techniques/tests for classifying instances of the concepts (names only):

Procedures/techniques for generating instances of the concepts (names only):

Principles that link any of these concepts (brief statements):

Procedures/techniques for applying any of these principles (names only):

Figure 3. Content analysis: Overview form.

For each concept listed in the overview ID code:

Content/topic area:

Concept name:

Other names:

Concept definition:

Example:

Source(s) for other examples:

Steps in one procedure/technique/test for identifying instances of the concepts:

Names and source(s) of information on alternative procedures:

Other concepts linked to this one:

Principles that link this concept to each of the others:

Figure 4. Content analysis: Concept form.

For each principle listed in the overview ID code:

Content/topic area:

Principle name:

Other name(s):

Principle statement:

Example of a situation where the principle operates:

Source(s) for other examples of situations where the principle operates:

Steps in one procedure/technique for applying the principle:

Names and source(s) of information on alternative procedures:

Figure 5. Content analysis: Principle form.

Example of completed content analysis: Overview form

Content/topic area: solution chemistry

ID code: SC1

Important concepts (names only):

solution; solute; solvent; concentration; evaporation; density; buoyancy; temperature; boiling point; freezing point

Procedures/techniques/tests for identifying instances of the concepts (names only):

Tyndall test; Use balance to measure mass per unit volume (density); Buoyancy test; Glass-tube test; Use of heat and thermometer to find boiling points of the liquids; evaporation

Procedures/techniques for generating instances of the concepts (names only):

stirring; heating; plus procedures for identifying/checking that the product is in fact an instance of the concept

Principles that link any of these concepts (brief statements):

The greater the concentration of a solution, the lower its freezing point.
The greater the concentration of a solution, the higher its boiling point.
The greater the concentration of a solution, the greater its density.
The greater the concentration of a solution, the greater its buoyancy.
The higher the boiling point of a solution, the lower its freezing point.

Procedures/techniques for applying any of these principles (names only):

same as for identifying and generating instances of concepts (goal would be to change state of one concept by manipulating another; therefore one would need to test for changes in concentration, buoyancy etc.)

Example of completed content analysis: Concept form

Content/topic area: solution chemistry ID code: SC1

Concept name: solution

Other names: N/A

Concept definition:

a homogeneous mixture of more than one substance (solid, liquid or gas); particles of solute are spread evenly throughout solvent; liquid solutions are transparent (but can have color)

Example: salt in water; sugar in water; blood

Source(s) for other examples:

Holt, Rinehart and Winston, *SciencePlus*, Red level book
Prentice Hall Science, *Chemistry of Matter* book
Addison-Wesley, *Science Insights*, Blue book

Steps in one procedure/technique/test for identifying instances of the concept:

Tyndall test: ...shine bright light through hole in cardboard into the liquid; if light passes through, leaving no trace, then it is a solution (solutions are transparent)...if particles in the mixture are large enough to scatter the light (the path of light is visible through the liquid), then the liquid exhibits the Tyndall effect; no true solution shows the Tyndall effect.

Names and source(s) of information on alternative procedures:

Evaporation: Holt, Rinehart and Winston, *SciencePlus*, Red level book

Other concepts linked to this one: N/A

Principles that link this concept to each of the others: N/A

Example of completed content analysis: Concept form

Content/topic area: solution chemistry ID code: SC1

Concept name: concentration

Other names: N/A

Concept definition:

the amount of solute per given volume of solvent (often expressed as number of grams of solute per 100 grams of solvent; 100 grams of water = 100 ml of water)

Example: a solution of 20 grams of sugar in 100 ml of water has a higher concentration than 10 grams of sugar in 100 ml of water

Source(s) for other examples:

Holt, Rinehart and Winston, *SciencePlus*, Red level book

Prentice Hall Science, *Chemistry of Matter* book

Steps in one procedure/technique/test for identifying instances of the concept:

Determine the mass of a given volume of a solution

Step 1. Determine the mass in grams (using a balance) of a beaker.

Step 2. Determine mass of beaker with 100 ml of solution in it.

Step 3. Subtract mass of beaker from mass of the solution.

Step 4. Conclude that the mass of 100 ml of solution is ___.

Names and source(s) of information on alternative procedures:

Evaporation: Holt, Rinehart and Winston, *SciencePlus*, Red level book,
Prentice Hall Science, *Chemistry of Matter* book

Other concepts linked to this one:

boiling point, freezing point, density, buoyancy

Principles that link this concept to each of the others:

The greater the concentration of a solution, the lower its freezing point.

The greater the concentration of a solution, the higher its boiling point.

The greater the concentration of a solution, the greater its buoyancy/density.

Example of completed content analysis: Principle form

Content/topic area: solution chemistry ID code: SC1

Principle name: Concentration/buoyancy

Other name(s):

Principle statement: The greater the concentration, the greater the buoyancy (density) of the solution.

Example of a situation where the principle operates:

Dead Sea (high salt content makes it easier to float)

Source(s) for other examples of situations where the principle operates:

Holt, Rinehart and Winston, *SciencePlus*, Red level book

Prentice-Hall Science, *Chemistry of Matter* book

Steps in one procedure/technique for applying the principle:

Test-tube and straw test for relative buoyancy: Pour same volume of each solution into a separate test tube; for each test tube make a "tester" (a primitive hydrometer) by sticking two thumbtacks onto the end of a pencil or a straw; place one "tester" in each test tube; the one that stands the highest indicates the more concentrated the solution.

Names and source(s) of information on alternative procedures:

Holt, Rinehart and Winston, *SciencePlus*, Red level book

Prentice-Hall Science, *Chemistry of Matter* book

Glass tube and food coloring (layering) test for relative densities

Comparing masses (using balance) of equal volumes of two solutions

Egg floating test

When the forms have been completed and edited based on a number of print-based sources, the designer should ask one or more teachers to verify and edit the information on the forms. When the set of forms has been

verified, one is ready to move to the next stage of the design phase, item or task construction. Specifications for task construction will be presented next.

Specifications for Task Construction for Assessment of Knowledge Structure

Specifications will be now be outlined for translating the product of a content analysis into multiple-choice, open-ended, and hands-on assessment tasks. The goal is to present students with multiple opportunities to demonstrate their knowledge of the concepts, principles of interest, and when and how to apply them in unfamiliar situations. It is important that the situations presented to students be unfamiliar to them. Efforts to identify the defining characteristics of “problem-solving tasks” have mostly resulted in the conclusion that the extent to which a task is a “problem” depends on how novel the task is for the solver (Bodner, 1992; Linn, Baker, & Dunbar, 1991; Lohman, 1993; Snow, 1989). If a student can complete a task mindlessly, without understanding the underlying concepts and principles, or without having to assemble a new strategy, then the task is not a problem for that student (Elshout, 1987; Smith, 1991). Once a procedure for a task has been learned, the task can no longer be a problem (Sowder, 1985). Even Larkin's (1983) characterization of the difficulty of a problem in terms of the number of principles that have to be considered in order to solve it does not rule out the fact that, unless the task is unfamiliar to the solver, it may not be a true test of his or her problem-solving ability.

It would be impossible to determine the extent to which a particular task is novel for every student who might be asked to solve it. The most feasible alternative approach is to develop a variety of tasks targeting the same content and cognitive variables and to examine the pattern of performance of a student over multiple tasks. If a student performs well on one task but not on other, comparable tasks, then one might infer that the reason the student did well on one task was because that task was very similar to a task that the student had completed during instruction. The pattern of performance of a student who is truly a good problem solver should be consistent across tasks targeting the same cognitive constructs in the context of same content.

What follows is a set of recommendations for designing, and examples of, items and tasks to measure each knowledge structure component in multiple-choice, open-ended, and hands-on response formats.

Assessment of Concepts (Multiple-choice format)

Present students with opportunities to identify examples of concepts and distinguish between examples that are and are not instances of the concept(s) of interest.

Example (concept of solution):

Identify which of the following liquids is a solution by placing an X beside it if it is a solution:

___ 5 grams of salt in 20 grams of water

___ 10 grams of salt in 20 grams of water

___ 3 drops of food coloring in 10 grams of water

___ 3 drops of food coloring in 20 grams of water

___ maple syrup

___ a cloudy liquid

___ a clear red liquid

___ a clear liquid with no color

___ a clear mixture of baking soda and water

Assessment of Concepts (Open-ended format)

Add a *Why?* question to a multiple-choice concept item; or present students with opportunities to give examples of concepts and explain what makes them examples of the concept; or ask students to give an example of something that is *not* an instance of the concept and explain why it is not an example.

Example (concept of solution):

Give an example of a liquid that is a solution and explain why it is a solution.

Give an example of a liquid that is not a solution and explain why it is not a solution.

Assessment of Concepts (Hands-on format)

Ask students to examine live examples of the concept and separate those that are examples of the concept of interest from those that are not examples of it (do not allow students to actually perform tests on the examples; performing tests on them would involve linking the concepts to procedures which is a separate construct to be assessed). Observe and record students' categorization of the objects; or ask students to record their own categorization of the examples on a form.

> **Example (concept of solution):**
>
> Give students a number of liquids, some obviously clear, some cloudy and ask the students to select (without testing them) the liquids that are most likely to be solutions.

Assessment of Principles (Multiple-choice format)

Present students with opportunities to select problems that are similar and dissimilar; to select the best explanation for a described event; or to select the best prediction for a described situation.

> **Examples (concentration/buoyancy principle):**
>
> One of the following problems is different from the rest. Indicate which one is the odd one out by placing an X beside it.
>
> _____ 1. Fred wanted to display his model boat floating in a basin of water, but when he tried it, the boat sank.
>
> _____ 2. Jamal's vinegar and oil salad dressing would not stay mixed up in the bottle.
>
> _____ 3. Maria wanted to make a sponge that would soak up as much water as possible.
>
> _____ 4. Joan wanted to make a number of colored liquids stay in separate layers in the glass.

Examples (concentration/buoyancy principle) continued:

It is very easy to stay afloat on lakes that have a lot of salt in them. Which of the following statements best explains this?

_____ 1. Our skin is allergic to salt and so we try to keep as much of our bodies out of the water as possible.

_____ 2. Higher salt concentrations cause greater buoyancy.

_____ 3. Higher salt concentrations make these lakes warmer.

_____ 4. Higher salt concentrations cause greater solubility.

John put four spoons of sugar in a cup of coffee. Which of the following is most likely to happen when he adds some whipped cream.

_____ 1. The cream will sink to the bottom of the coffee.

_____ 2. The cream will mix evenly throughout the coffee without even stirring it.

_____ 3. The cream will float on the top of the coffee.

_____ 4. The cream will evaporate.

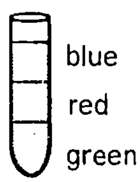
Which of the liquids in the glass tube drawn below is likely to have the highest concentration?

a. the blue liquid

b. the green liquid

c. the red liquid

d. they will all have the same concentration



Assessment of Principles (Open-ended format)

Add a *Why?* question to a multiple-choice item, or ask students to explain given events, or to make predictions about what will happen in a given situation.

> **Example (concentration/buoyancy principle):**
>
> Why is it easy to float on Utah's Great Salt Lake?
>
> A bowl is half-full with a solution of lemonade powder and water. An egg is dropped in the lemonade and it floats. What is likely to happen to the egg if water is added to fill the bowl to the top?

Assessment of Principles (Hands-on format, with open-ended responses in writing)

Ask students to follow some step-by-step instructions to do something, observe the result, and then explain the result.

> **Example (concentration/buoyancy principle):**
>
> Follow the instructions below, observe the result, and then answer the questions at the end.
>
> Instructions:
>
> Pour the maple syrup into the beaker; then pour the colored water in on top of it; then pour the corn syrup on top of that.
>
> Questions:
>
> 1. Draw a picture of how the liquids look in the beaker. Label each liquid.
> 2. Why do the liquids end up in these layers?

Assessment of links from concepts to conditions and procedures for application (Multiple-choice format)

Provide students with opportunities to select the correct procedure for identifying the concept to which a particular substance or object belongs.

> **Example (concept of concentration):**
>
> John needed to test two sodas to see which one contained the most sugar. Some of the following methods will given him the answer, and some of them won't. Put an X beside any method that will give him the answer.
>
> _____ 1. Use a balance to compare the mass of 100 ml of each of the sodas.
>
> _____ 2. Find an object that will float on one, but not on the other.
>
> _____ 3. Compare the amount of each soda it takes to soak a sponge of the same size.
>
> _____ 4. Boil both liquids and compare their boiling points.

Assessment of links from concepts to conditions and procedures for application (Open-ended format)

Ask students to describe a method for determining the identity of a substance or object.

> **Example (concept of concentration):**
>
> Describe how you would determine which of two solutions of sugar and water was the more concentrated (without tasting them).

Assessment of links from concepts to conditions and procedures for application (Hands-on format with observation and/or open-ended responses in writing)

Ask students to actually test some unknown substance or object to identify it. Observe the student's actions and/or ask students to keep a written record of their actions and observations.

> **Example (concept of concentration):**
>
> Use the equipment and materials provided to identify which of the liquids in the cups is more concentrated. (Provide the equipment necessary for more than one kind of test that is appropriate.)

Assessment of links from principles to conditions and procedures for application (Multiple-choice format)

Provide students with opportunities to select the most appropriate procedure to change the state of one concept by manipulating another.

> **Example (concentration/freezing point principle):**
>
> A soda manufacturer wants its sodas not to freeze in very cold weather. Which of the following methods would be most likely to solve this problem?
>
> a. Decrease the amount of gas in the soda.
>
> b. Decrease the amount of sugar in the soda.
>
> c. Increase the amount of sugar in the soda.
>
> d. Store the soda in bottles rather than in cans.

Assessment of links from principles to conditions and procedures for application (Open-ended format)

Add a *Why?* question to a multiple-choice item, or ask students to describe how they would solve given problems involving two or more concepts, where one concept can be manipulated to effect a change in another. Also, ask why they think their solution would work, or ask them to explain why a given solution method would work, or to explain why given solutions would not work.

Example (concentration/freezing point principle):

Customers in Alaska were complaining that cans of diet cola were freezing and bursting in the cold winter temperatures. Cans of regular cola were not freezing. What would you tell the makers of diet cola to do to stop it from freezing and why?

Assessment of links from principles to conditions and procedures for application (Hands-on format, with observation and/or open-ended responses in writing)

Ask students to perform the tests and actions necessary to solve a problem where two or more concepts are affecting each other. Observe the solution process and/or have students answer questions in writing as they work on the problem.

Example (concentration/freezing point principle):

You are going on an expedition/journey to the North Pole. You need to bring a supply of liquid to drink during your expedition/journey. You can choose one of the three liquids on the table. Consider what you know about solution chemistry and perform the tests necessary to select the liquid that will be least likely to freeze during your journey. Then write the answers to the following two questions:

1. What is your conclusion (which liquid is least likely to freeze)?
2. Why?

Summary of Specifications for Designing Tasks to Assess Knowledge Structure

First, carry out an analysis of the subject matter content to be assessed. The goal is to identify key concepts, principles and procedures that are embodied in the content. Second, create a variety of multiple-choice items, open-ended items, and hands-on tasks to measure knowledge of concepts, principles and their links with conditions and procedures for application. Figure 6 summarizes the critical features of items or tasks of each format for each knowledge structure component to be assessed. The trend in science assessment is towards more extended tasks that combine hands-on response with written response to open-ended questions based on the hands-on activity. In the design model presented here, such extended tasks can be compiled by

Construct	Format		
	Multiple-choice	Open-ended	Hands-on
Concepts	select examples	M/C + Why or generate examples (describe orally or in writing)	select live examples
Principles	select similar problems or select best prediction or select best explanation	M/C + Why or make prediction or explain an event (orally or in writing)	Follow instructions, observe and explain result (orally or in writing)
Links from concepts to conditions and procedures for application	select correct procedure for identifying instances	M/C + Why or generate (describe) a procedure for identifying instances	perform procedures (tests) to identify instances
Links from principles to conditions and procedures for application	select most appropriate procedure to change the state of a concept by manipulating another	M/C + Why or generate (describe) a procedure to change the state of one concept by manipulating another	perform procedures to change state of one concept by manipulating another

Figure 6. Critical features of tasks for assessing components of knowledge structure.

piecing together elements from a number of cells in the construct X format matrix. The advantage of this approach is that each component of the extended task is explicitly linked to the cognitive construct it is tapping. In addition, the creation of comparable scoring systems is facilitated (see the section on designing scoring schemes below).

Specifications for Modifying and Generating Tasks to Assess Cognitive Functions

Once a set of items and tasks has been developed to target the knowledge structure constructs that facilitate problem solving, then one can begin to modify those tasks, or add tasks, to generate information on the separate cognitive functions of planning and monitoring. Planning was defined earlier as the ability to think through what one will do before actually doing it. Monitoring was defined as keeping track of a number of aspects of one's performance, including time, the effects of one's efforts in relation to the goal and constraints of the problem, and adapting one's strategy if necessary. As with the assessment of knowledge structure, multiple indicators of planning and monitoring should be obtained for each student. Figure 7 summarizes the critical features of various methods for gathering information on the planning and monitoring abilities of students. All of the methods suggested assume that it is planning in the context of the task domain of interest that is being measured, rather than a domain-independent metacognitive skill.

	Format		
Construct	Multiple-choice	Open-ended	Hands-on
Planning	rate statements about level of planning	describe plan or give advice to others	proportion of time spent planning
Monitoring	rate statements about monitoring activities	describe how checked/kept track of performance or give advice to others	proportion of time spent checking performance

Figure 7. Critical features of methods for assessing components of cognitive functioning.

There are many ways to embed the assessment of planning and monitoring within the assessment of knowledge structure. Planning and monitoring ability can be inferred from time allocation data gathered as students complete sets of items designed to assess components of knowledge structure, or as students complete a hands-on assessment of the link between principles, and conditions and procedures. Questions can be added, to any test, that ask students to describe their plan for completing the test or for completing a hands-on task before they even begin the test. If performance on a hands-on task is observed, one can count the number of times a student looks back over work completed, or refers back to the instructions, or looks ahead to later parts of the task to be completed.

Alternatively, or in addition to embedded assessment of cognitive function, special items/questions can be created to assess planning and monitoring ability. These questions can be presented in multiple-choice or open-ended format. Multiple-choice questions should ask students to evaluate the extent to which descriptions of planning, monitoring, and the lack of them, relate to themselves. For example, students can be asked to indicate on a scale from 1 to 5 how well the following statements reflected their performance on the test they have just completed:

1. I worked out how much time I should spend on each question and I tried to stick to it.
2. I ran out of time at the end of the test.
3. I spent a long time planning how I would answer the questions.
4. I looked at the clock/my watch every few minutes.
5. I got lost in the middle of the problem and had to start over.
6. I did not look at the clock/my watch very much during the test.

In open-ended format, students can be asked to write, at the end of a test, the advice they would give to other students who might have to take the test in the future. Students could also be asked to describe how they allocated their time, how they checked that their answers were correct, or how they kept track of their progress on the test. Students can be given a number of problems and asked to generate plans for solving them but not to execute the plan, or to

suggest ways that they could check that they were on the right track. Students can be given a description of a problem and the procedure another student used to solve it and can be asked to generate the plan that would have guided the solution procedure.

Specifications for Modifying or Designing Tasks to Assess Beliefs

Three aspects of students' beliefs about themselves and the task should be measured: the student's belief about his or her own ability to perform on the test; the student's belief about how attractive the test is; and the student's belief about how difficult or novel the task is. Figure 8 summarizes the critical features of methods that can be used to assess task-specific beliefs about self-efficacy, task difficulty, and task attraction.

Students can be asked to read the test quickly, and then, before they start to work on it, they can answer a number of questions relating to the three belief constructs of interest. The questions should relate to the topic and format of the test that the students are about to take. Questions can be formatted so that students have to indicate, on some numerical scale, how well a particular statement reflects their beliefs, or students can be asked to give open-ended responses to questions about their own ability, and the difficulty and attractiveness of the task. At the end of the test, students can again be asked to answer some questions about how well they expect to score, how difficult the test was, and how they liked the test.

Some hands-on (behavioral) indicators of PSE, PDT and PAT can also be gathered by observing students as they complete a test. PSE, PDT and PAT influence persistence and effort-expenditure (Bandura, 1986; Schunk, 1990); therefore, one can observe the amount of time a student struggles with parts of the test; how far a student goes before giving up (even though there is time left); whether students seem engaged or bored by the task. Students can be asked to view a video recording of their performance and to answer questions about why they were behaving in certain ways during the test (their answers may reveal information on how difficult the task seemed at different points, how able they felt, and how much they were enjoying the task).

Construct	Format Multiple-choice	Open-ended	Hands-on
PSE	rate statements about one's ability to do well on the test	describe how well you are likely to do on the test (before and after the test)	retrospective interview while watching recording of performance or time spent on parts/ questions where more effort required (persistence)
PDT	rate statements about how difficult the test seems	describe difficulty of the test (before and after the test)	time spent on different parts of the test or time spent when solution not apparent (persistence)
PAT	rate statements about liking or enjoyment of the test	describe how much you liked or enjoyed the test	observe and rate student's level of focus on the test (engagement)

Figure 8. Critical features of methods for assessing components of beliefs.

Specifications for Scoring Performance on Assessments of Problem Solving in Science

The approach to scoring recommended here involves examining patterns of performance over sets of items and aspects of task performance that were designed to measure specific cognitive constructs. The goal is to generate a profile of each student's performance in terms of the constructs of interest. Figure 9 summarizes the main features of the scoring of items or tasks that target each of the cognitive constructs that affect problem solving. Only general recommendations are made here. These are being implemented and evaluated in a study currently underway at CRESST. Future reports will have

Construct	Format		
	Multiple-choice	Open-ended	Hands-on
Concepts	proportion of correct selections out of total number of possible selections	proportion of correct instances out of all instances generated	proportion of correct identifications
Principles	proportion of correct selections	proportion of correct predictions, explanations	proportion of correct predictions, explanations
Links from concepts and principles to conditions and procedures	proportion of correct selections	proportion of correct procedures suggested	proportion of correct procedures selected
Planning	average rating of a number of statements	rating of plans generated	proportion of time spent planning
Monitoring	average rating of a number of statements	rating of descriptions of monitoring	proportion of time spent checking work
PSE	average rating of a number of statements	rating of written or oral descriptions of competence	proportion of time spent on correct and incorrect items
PDT	average rating of a number of statements	rating of written or oral descriptions of difficulty	proportion of time spent on correct and incorrect items
PAT	average rating of a number of statements	rating of written or oral descriptions of attraction of the task	proportion of time engaged; proportion of time bored or distracted

Figure 9. Scoring different response formats with respect to the cognitive constructs of interest.

more detailed recommendations on scoring procedures for the different constructs and formats.

For sets of multiple-choice test items targeting any particular knowledge structure construct, a student's score is simply the proportion of correct selections made. For sets of knowledge structure items requiring open-ended

responses, scoring is based on the extent to which a student describes correct examples of the concept(s), uses appropriate principles to generate explanations or to make predictions, and describes appropriate procedures to identify instances of the concepts or to apply the principles of interest. Scoring of such open-ended responses can be dichotomous (either the student did or did not mention a correct example, or did or did not give an explanation that indicated that he or she understands the principle). Scoring of open-ended response can also be more elaborate, each student being assigned to a point on a numerical scale that represents the degree to which the student demonstrates proficiency on the construct of interest. Each point on such a scale should be defined in terms of the specific elements that need to be present in the student's response.

Only elements of hands-on responses that relate to the cognitive constructs of interest should be scored. Since problem solving is the skill being assessed, the accuracy of the procedures performed should *not* be scored; what is important is that the student selected the appropriate procedure, indicating that the student has linked a concept or principle to a procedure. Assessment of the accuracy or speed of performance of procedures might be a peripheral part of an assessment of problem solving, but these were not identified as critical cognitive variables in the model adopted here.

For cognitive functions and beliefs, scores can be allocated based on average ratings of statements describing one's planning, monitoring or beliefs. The rating of open-ended responses to questions about cognitive functions or beliefs is more problematic; plans generated by students can be rated for their completeness and accuracy; responses about monitoring activity can be rated based on how much and at what points students say they checked their work, kept track of time, etc.; open-ended responses to questions about beliefs can be rated in terms of the degree to which the student believes he or she has the ability to do well on the test, the degree to which the test seems difficult, and the degree to which they look forward to doing the test (or the extent to which they enjoyed it).

Observation of performance and retrospective interviews of students while watching a video recording of their performance may lead to more accurate ratings of cognitive functions and beliefs, or at least serve to validate students' self-reports of these variables. However, one must first decide what aspects of

behavior will serve as indirect indicators of cognitive functions or beliefs. Research is needed to isolate and validate such aspects of behavior. Meanwhile, it is recommended here that one look for proportions of time spent planning, checking work, working on correct and incorrect items, and seeming to be engaged in the task (as opposed to bored, distracted, or frustrated).

Conclusion

Measurement theory and assessment practice are moving towards cognitive conceptions of performance. A cognitively-based approach to assessment means that test development begins with a theory about the cognitive structures and processes that underlie or facilitate the skill or ability to be measured. A cognitive conception of skill in a domain can drive the design of test items and tasks, the scoring of performance on those items and tasks, and inferences about the cognitive capabilities of students. More emphasis is put on the design of the test than on psychometric analysis after the test is written (Glaser et al., 1991).

This report has presented specifications for designing tasks to assess problem-solving ability in science, specifications that are clearly grounded in a cognitive definition of problem-solving performance. The specificity of definition of the cognitive constructs of interest goes beyond vague definitions of "understanding" and "reasoning" to identification of the specific knowledge types and links among them, and the specific aspects of "higher order" thinking that have been found to influence problem-solving performance, regardless of content or domain. The more specific the definitions of the cognitive aspects of performance to be targeted by assessment, the more tasks and scoring of performance on them can be rendered consistent with the underlying cognitive dimensions of performance.

Research is underway to empirically evaluate the extent to which the model for assessment design presented here facilitates (a) the assessment of generalizable components of problem-solving ability in the domain of science, (b) the generation of score profiles that remain constant regardless of the format of the test, and (c) the isolation of the cognitive sources of poor problem-solving performance in the domain of science. This research will lead to a refinement of the model of the cognitive components of problem solving

presented here, and to more precise specifications for eliciting and scoring behaviors that reflect those components. This research will also lead to theory-based recommendations for the selection of test formats to match a variety of authenticity and efficiency requirements.

References

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*(3), 483-495.

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*(4), 375-404.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261-295.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*(1), 35-44

Baker, E. L. (1990). Developing comprehensive assessments of higher order thinking. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 7-20). Washington, DC: American Association for the Advancement of Science.

Baker, E. L. (1992). *The role of domain specifications in improving the technical quality of performance* assessment (Report to OERI). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., Aschbacher, P., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., & Herman, J. L. (1983). Task structure design: Beyond linkage. *Journal of Educational Measurement*, *20*(2), 149-164. (ERIC Document Reproduction Service No. EJ 284 872)

Bandura, A. (1982). Self-efficacy mechanisms in human agency. *American Psychologist, 37*, 122-148.

Bandura, A. (1986). *Social foundations of thoughts and actions*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1989). Regulation of cognitive processes through perceived self-efficacy. *Developmental Psychology*, *25*, 729-735.

Baxter, G. P., Glaser, R., & Raghavan, K. (1993). *Analysis of cognitive demand in selected alternative science assessments* (Report to CRESST/OERI). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Bennett, R. E. (1993). On the meanings of constructed response. In *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-28). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Benton, S. L., & Kiewra, K. A. (1987). The assessment of cognitive factors in academic abilities. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 145-189). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bodner, G. M. (1991). In Smith, M. U. (Ed.). *Toward a unified theory of problem solving: Views from the content domains* (pp. 21-34). Hillsdale, NJ: Lawrence Erlbaum.

Boekaerts, M. (1987). Problem-solving and education. In E. De Corte, H. Lodewikjs, R. Parmentier, & P. Span (Eds.), *Learning and instruction: European research in an international context* (Vol. 1, pp. 169-179). Oxford: John Wiley and Sons.

Boekaerts, M. (1991). Subjective competence, appraisals and self-assessment. *Learning and Instruction*, *1*, 1-17.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, *29*(3), 253-271.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Cognitive development* (Vol. 3, pp. 77-166), of P. H. Mussen (Ed.), *Handbook of child psychology* (4th ed.). New York: Wiley.

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

Campione, J. C., & Brown, A. L. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 141-172). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chemistry of matter. (1993). Annotated Teacher's Edition, Prentice Hall Science. Englewood Cliffs, NJ: Prentice Hall.

Chi, M. T. H. (1985). Interactive roles of knowledge and strategies in the development of organized sorting and recall. In S. F. Chipman, J. W.

Segal, & R. Glaser (Eds.), *Thinking and learning skills, Volume 2: Research and open questions* (pp. 457-483). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Chi, M. T. H., & Glaser, R. (1985). Problem-solving ability. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 227-250). New York: W. H. Freeman.

Chi, M. T. H., & Glaser, R. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, R. E. (1990, April). *A cognitive theory of instructional method.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Clark, R. E., & Blake, S. B. (in press). Designing cognitively-based training for the far transfer of problem solving expertise. In S. Dijkstra & N. M. Seel (Eds.), *Instructional design: International perspectives. Volume II: Solving instructional design problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, L. K., & Mayer, R. E. (1988). Teaching readers about the structure of scientific text. *Journal of Educational Psychology*, *80*, 448-456.

Cooke, N. J. (1992). The implications of cognitive task analyses for the revision of the Dictionary of Occupational Titles. In *American Psychological Association, Implications of cognitive psychology and cognitive task analysis for the revision of the Dictionary of Occupational Titles* (Report submitted to the U.S. Department of Labor).

Corno, L., & Mandinach, E. B. (1983). The role of cognitive engagement in classroom learning and motivation. *Educational Psychologist*, *18*(2), 88-108.

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

Dweck, C. S., & Licht, B. G. (1980). Learned helplessness and intellectual achievement. In J. Garber & M. E. P. Seligman (Eds.), *Human helplessness theory and applications* (pp. 197-221). New York: Academic Press.

Elshout, J. J. (1987). Problem-solving and education. In E. De Corte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction: European research in an international context* (Vol. 1, pp. 259-273). Oxford: John Wiley and Sons.

Ericsson, A. K., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data.* Cambridge, MA: MIT Press.

Feather, N. T. (Ed.). (1982). *Expectations and actions: Expectancy value models in psychology.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Feather, N. T. (1988). Values, valences and course enrollment: Testing the role of personal values within an expectancy-valence framework. *Journal of Educational Psychology, 80*(3), 381-391.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning* (2nd ed.). New York: Harper Collins College Publishers.

Garner, R. (1988). *Verbal-report data on cognitive and metacognitive strategies.* San Diego: Academic Press.

Garner, R., & Reis, R. (1981). Monitoring and resolving comprehension obstacles: An investigation of spontaneous text lookbacks among upper-grade good and poor comprehenders. *Reading Research Quarterly, 16*, 569-582.

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

Gerace, W. J., & Mestre, J. P. (1990). *Materials for developing concept-based problem solving skills in physics.* Unpublished manuscript, University of Massachsetts, Amherst.

Gillingham, M. G., Garner, R., Guthrie, J. T., & Sawyer, R. (1989). Children's control of computer-based reading assistance in answering synthesis questions. *Computers in Human Behavior, 5*, 61-75.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*, 93-104.

Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research*, *14*(5), 475-483

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

Glaser, R., Lesgold, A., & Gott, S. (1991). Implications of cognitive psychology for measuring job performance. In A. K. Wigdor & B. F. Green, Jr. (Eds.), *Performance assessment for the workplace. Volume II: Technical issues* (pp. 1-26). Washington, DC: National Academy Press.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing and measurement* (pp. 41-85). Hillsdale, 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.

Goldin, G. A., & McClintock, C. E. (Eds). (1984). *Task variables in mathematical problem solving*. Philadelphia, PA. The Franklin Institute Press.

Greeno, J. G., & Simon, H. A. (1989). Problem solving and reasoning. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology. Volume 2: Learning and cognition* (2nd ed., pp. 589-672). New York: John Wiley and Sons.

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory and Cognition*, *17*(5), 627-638.

Harris, P. L., Kruithof, A., Terwogt, M. M., & Visser, T. (1981). Children's detection and awareness of textual anomaly. *Journal of Experimental Child Psychology*, *31*(2), 212-230.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hayes-Roth, B., & Hayes-Roth, F. (1977). Concept learning and the recognition and classification of exemplars. *Journal of Verbal Learning and Verbal Behavior*, *16*(3), 321-338.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hively, W. (1974, June). Introduction to domain-references testing. *Educational Technology*, *14*(6), 5-10.

Keane, M. T. (1988). *Analogical problem solving*. New York: Halsted Press.

Kilpatrick, J. (1985). A retrospective account of the past 25 years of research on teaching mathematical problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 1-15). Hillsdale, NJ: Lawrence Erlbaum.

Klausmeier, H. J. (1992). Concept learning and concept teaching. *Educational Psychologist*, *27*(3), 267-286.

Kulm, G. (Ed.). (1990). *Assessing higher order thinking in mathematics*. Washington, DC: American Association for the Advancement of Science.

Lamon, S. J., & Lesh, R. (1992). Interpreting responses to problems with several levels and types of correct answers. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 319-342). Washington, DC: American Association for the Advancement of Science Press.

Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & A. L. Stevens, *Mental models* (pp. 75-98). Hillsdale, NJ: Lawrence Erlbaum Associates.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*, 1335-1342.

Lesgold, A., Lajoie, S., Logan, D., & Eggan, G. (1990). Applying cognitive task analysis and research methods to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G., Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 325-350). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, A. B. (1989). Training students to represent arithmetic work problems. *Journal of Educational Psychology*, *81*, 521-531.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher*, *22*(7), 12-23.

Lohman, D. F., & Ippel, M. J. (1993). Cognitive diagnosis: From statistically based assessment toward theory-based assessment. In N. Frederiksen, R. J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50*(3), 643-655.

Marshall, S. P. (1988). Assessing problem solving: A short-term remedy and a long-term solution. In R. I. Charles & E. A. Silver (Eds.), *Research agenda for mathematics education: The teaching and assessment of mathematical problem solving* (Vol. 3, pp. 159-177). Reston, VA: The National Council of Teachers of Mathematics & Lawrence Erlbaum Associates.

Marshall, S. P. (1990). Generating good items for diagnostic tests. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 433-452). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marshall, S. P. (1993). Assessing schema knowledge. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 155-180). Hillsdale, NJ: Erlbaum.

Marton, F. (1983). Beyond individual differences. *Educational Psychology, 3*, 289-304.

McCombs, B. L. (1988). Motivational skills training: Combining metacognitive, cognitive, and affective learning strategies. In C. E. Weinstein, E. T. Goetz, & P. A. Alexander (Eds.), *Learning and study strategies* (pp. 141-169). San Diego: Academic Press.

McLeod, D. B. (1985). Affective issues in research on teaching mathematical problem solving. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 267-280). Hillsdale, NJ: Lawrence Erlbaum Associates.

Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Ed.), *Instructional-design theories and models: An overview of their current status* (pp. 279-333). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 61-74). Hillsdale, NJ: Erlbaum.

Mestre, J. P., Dufresne, R. J., Gerace, W. J., Hardiman, P. T., & Tougher, J. S. (1992). Enhancing higher-order thinking skills in physics. In D. F.

Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 77-94). Hillsdale, NJ: Lawrence Erlbaum Associates.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

National Council of Teachers of Mathematics (1989) *Curriculum and Evaluation standards for school mathematics*. Reston, VA: Author.

National Research Council. (1993, February). *National science education standards: An enhanced sampler* (Working Paper of the National Committee on Science Education Standards and Assessment). Washington, DC. Author.

Nickerson, R. S. (Ed.). (1989). New directions in educational assessment [Special issue]. *Educational Researcher*, *18*(9), 3-7.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447-474). New York: Macmillan.

Norris, S. P. (1989). Can we test validly for critical thinking. *Educational Researcher*, *18*(9), 21-26.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1992). *Final report of experimental studies on motivation and NAEP test performance* (Report to NCES, Contract # RS90159001). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Peterson, P. L., Swing, S. R., Braverman, M. T., & Buss, R. (1982). Students' aptitudes and their reports of cognitive processing during instruction. *Journal of Educational Psychology, 74,* 535-547.

Peverly, S. T. (1991). Problems with the knowledge-based explanation of memory and development. *Review of Educational Research*, *61*(1), 71-93.

Pintrich, P. R., & De Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*, 33-40.

Popham, W. J. (Ed.). (1984). *A guide to criterion-referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Popham, W. J. (1993). Educational testing in America: What's right, what's wrong?: A criterion-referenced perspective *Educational Measurement: Issues and Practices*, *12*(1), 11-14.

Resnick, L. B. (Ed.). (1989). *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ronan, W. W., Anderson, C. L., & Talbert, T. L. (1976). A psychometric approach to job performance: Fire fighters. *Public Personnel Management*, *5*(6), 409-413.

Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research*, *63*(2), 201-243.

Rumelhart, D. E., & Norman, D. A. (1989). Representations in memory. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology. Volume 2: Learning and cognition* (2nd ed., pp. 511-587). New York: John Wiley and Sons.

Salomon, G. (1983). The differential investment of mental effort in learning from different sources. *Educational Psychologist*, *18*, 42-50.

Schoenfeld, A. H. (1982). Some thoughts on problem solving research and mathematics education. In F. Lester & J. Garofalo (Eds.), *Mathematical problem solving: Issues in research*. Philadelphia: Franklin Institute Press.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. San Diego, CA: Academic Press.

Schultz, K., & Lochhead, J. (1991). A view from physics. In M. U. Smith, (Ed.), *Toward a unified theory of problem solving: Views from the content domains* (pp. 99-114) Hillsdale, NJ: Lawrence Erlbaum Associates.

Schunk, D. H. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist*, *19*, 48-58.

Schunk, D. H. (1990). Goal setting and self-efficacy during self-regulated learning. *Educational Psychologist*, *25*, 71-86.

Science Insights. (1994). Blue book. Menlo Park, CA: Addison-Wesley.

SciencePlus. *Technology and society*. (1993). Orlando, FL: Holt, Rinehart and Winston.

Seegers, G., & Boekaerts, M. (1993). Task motivation and mathematics achievement in actual task situations. *Learning and Instruction*, *3*, 133-150.

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

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.

Siegler, R. S. (1989). Strategy, diversity and cognitive assessment. *Educational Researcher*, *18*(9), 15-20.

Silver, E. A. (Ed.). (1985). *Teaching and learning mathematical problem solving: Multiple research perspectives*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, J. K. (1982, March). *Cognitive approaches to test taking strategies*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Smith, M. U. (Ed.). (1991). *Toward a unified theory of problem solving: Views from the content domains*. Hillsdale, NJ: Lawrence Erlbaum.

Smith, M. U., & Good, R. (1984). Problem solving and classical genetics: Successful versus unsuccessful performance. *Journal of Research in Science Teaching*, *21*(9), 895-912.

Snow, R. E. (1980). Aptitude, learner control, and adaptive instruction. *Educational Psychologist*, *15*(3), 151-158.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher*, *18*(9), 8-14.

Snow, R. E. (1990). New approaches to cognitive and conative assessment in education. *International Journal of Educational Research*, *14*(5), pp. 455-473

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45-60). Hillsdale, NJ: Erlbaum.

Snow R. E., & Jackson, III, D. N. (1993). *Assessment of conative constructs for educational research and evaluation: A catalogue* (CSE Tech. Rep. No. 354). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*(3), 347-376.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Sowder, L. K. (1985). Cognitive psychology and mathematical problem solving: A discussion of Mayer's paper. In E. A. Silver (Ed.), *Teaching and learning mathematical problem solving: Multiple research perspectives* (pp. 139-146). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (1985) Cognitive approaches to intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications* (pp. 59-118). New York: John Wiley.

Stipek, D. (1993). *Student attitude survey*. Santa Monica, CA: Galef Institute.

Swing, S. R., Stoiber, K. C., & Peterson, P. L. (1988). Thinking skills vs. learning time: Effects of alternative classroom-based interventions on students' mathematics problem solving. *Cognition and Instruction, 5*(2), 123-191.

Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 107-134). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tennyson, R. D., & Cocchiarella, M. J. (1986). An empirically based instructional design theory for teaching concepts. *Review of Educational Research, 56*(1), 40-71.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 261-285). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wagner, R. K., & Sternberg, R. J. (1987). Executive control in reading comprehension. In B. K. Britton & S. M. Glynn (Eds.), *Executive control process in reading* (pp. 1-22). Hillsdale, NJ: Lawrence Erlbaum Associates.

Webb, N., & Yasui, E. (1992). *The influence of problem context on mathematics performance* (CSE Tech. Rep. No. 346). Los Angeles, University of

California, Center for Research on Evaluation, Standards, and Student Testing.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *75*(3), 200-214.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, *11*, 307-313.

Zimmerman, B. J., & Martinez-Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal, 23*, 614-629.

Zimmerman, B. J., & Martinez-Pons, M. M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology*, *82*, 51-59.