Conceptual Framework and Design of the High School Study: A Multidimensional Approach to Achievement Validation

CSE Technical Report 569

Richard J. Shavelson, CRESST/Stanford University

Robert Roeser, Stanford University

Haggai Kupermintz, University of Colorado at Boulder

Shun Lau, Carlos Ayala, Angela Haydel, and Susan Schultz
Stanford University

July 2002

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

Project 1.1 Models-Based Assessment: Individual and Group Problem Solving in Science
Project 3.1 Construct Validity: Understanding Cognitive Processes—Psychometric and Cognitive Modeling
Richard Shavelson, Project Director, CRESST/Stanford University

The work reported herein was supported in part under the Educational Research and Development Center Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education, and in part by the National Science Foundation (REC9628293).

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

PREFACE

In 1995, Richard E. Snow wrote in CRESST's proposal to the Office of Educational Research and Improvement that his previous work showed that "psychologically meaningful and useful subscores can be obtained from conventional achievement tests" (Baker, Herman, & Linn, 1995, p. 133). He went on to point out that these subscores represented important ability distinctions and showed different patterns of relationships with demographic, "affective" (emotional), "conative" (volitional), and instructional-experience characteristics of students. He concluded that "*a new multidimensional approach to achievement test validation* should include affective and conative as well as cognitive reference constructs" (italics ours, p. 134).

Snow (see Baker et al., 1995) left hints of what he meant by "a new multidimensional approach" when he wrote, "the primary objective of this study is to determine if knowledge and ability distinctions previously found important in high school math and science achievement tests occur also in other multiple-choice and constructed response assessments. . . . A second objective is to examine the cognitive and affective correlates of these distinctions. And a third objective is to examine alternative assessment designs that would sharpen and elaborate such knowledge and ability distinctions in such fields as math, science, and history-geography" (p. 133).

We, as Snow's students and colleagues, have attempted to piece together his thinking about multidimensional validity and herein report our progress on a research program that addresses cognitive and motivational processes in high school science learning and achievement. To be sure, if Dick had been able to see this project through to this point, it might well have turned out differently. Nevertheless, we attempted to be true to his ideas and relied heavily on the theoretical foundation of his work, his conception of aptitude (Snow, 1989, 1992).

Snow called for broadening the concept of aptitude to recognize the complex and dynamic nature of person-situation interactions and to include motivational (affective and conative) processes in explaining individual differences in learning and achievement. Previous results, using a mixed methodology of large-scale statistical analyses and small-scale interview studies, demonstrated the usefulness of a multidimensional representation of high school science achievement. We identified three distinct constructs underlying students' performance on a standardized test and sought validation evidence for the distinctions between "basic knowledge and reasoning," "quantitative science," and "spatial-mechanical ability" (see Hamilton, Nussbaum, & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997). Different patterns of relationships of these dimensions with student background variables, instructional approaches and practices, and out-of-school activities provided the groundwork for understanding the essential characteristics of each dimension. We found, for example, that gender differences in science achievement could be attributed to the spatial-mechanical dimension and not to aspects of quantitative reasoning or basic knowledge and facts.

Our studies, reported in the set of six CSE Technical Reports Nos. 569–574,* extend the groundwork laid down in Snow's past research by introducing an extensive battery of motivational constructs and by using additional assessment formats. This research seeks to enhance our understanding of the cognitive and motivational aspects of student performance on different test formats: multiple-choice, constructed response, and performance assessments. The first report (Shavelson et al., 2002) provides a framework for viewing multidimensional validity, one that incorporates cognitive ability (fluid, quantitative, verbal, and visualization), motivational and achievement constructs. In it we also describe the study design, instrumentation, and data collection procedures. As Dick wished to extend his research on large-scale achievement tests beyond the National Education Longitudinal Study of 1988 (NELS:88), we created a combined multiple-choice and constructed response science achievement test to measure basic knowledge and reasoning, quantitative reasoning, and spatial-mechanical ability from questions found in NELS:88, the National Assessment of Educational Progress (NAEP), and the Third International Mathematics and Science Study (TIMSS). We also explored what science performance assessments (laboratory investigations) added to this achievement mix. And we drew motivational items from instruments measuring competence beliefs, task values, and behavioral engagement in the science classroom. The second report in the set (Lau, Roeser, & Kupermintz, 2002) focuses on cognitive and motivational aptitudes as predictors of science achievement. We ask whether, once students' demographic characteristics and cognitive ability are taken into consideration, motivational variables are implicated in science achievement. In the third report (Kupermintz & Roeser , 2002), we explore in some detail the ways in which students who vary in motivational patterns perform on basic knowledge and reasoning, quantitative reasoning, and spatial-mechanical reasoning subscales. It just might be, as Snow posited, that such patterns interact with reasoning demands of the achievement test and thereby produce different patterns of performance (and possibly different interpretations of achievement). The fourth report (Ayala, Yin, Schultz, & Shavelson, 2002) then explores the link between large-scale achievement measures and measures of students' performance in laboratory investigations ("performance assessments"). The fifth report in the set (Haydel & Roeser, 2002) explores, in some detail, the relation between varying motivational patterns and performance on different measurement methods. Again, following Snow's notion of a transaction between (motivational) aptitude and situations created by different test formats, different patterns of performance might be produced. Finally, in the last report (Shavelson & Lau, 2002), we summarize the major findings and suggest future work on Snow's notion of multidimensional achievement test validation.

---

* This report and its companions (CSE Technical Reports 570, 571, 572, 573, and 574) present a group of papers that describe some of Snow's "big ideas" with regard to issues of aptitude, person-situation transactions, and test validity in relation to the design of a study (the "High School Study") undertaken after Snow's death in 1997 to explore some of these ideas further. A revised version of these papers is scheduled to appear in *Educational Assessment* (Vol. 8, No. 2). A book based on Snow's work, *Remaking the Concept of Aptitude: Extending the Legacy of Richard E. Snow*, was prepared by the Stanford Aptitude Seminar and published in 2002 by Lawrence Erlbaum Associates.

CONCEPTUAL FRAMEWORK AND DESIGN OF THE HIGH SCHOOL STUDY A MULTIDIMENSIONAL APPROACH TO ACHIEVEMENT VALIDATION

Richard J. Shavelson, CRESST/Stanford University

Robert Roeser, Stanford University

Haggai Kupermintz, University of Colorado at Boulder

Shun Lau, Carlos Ayala, Angela Haydel, and Susan Schultz
Stanford University

Abstract

Richard E. Snow conceived of an individual's performance as the result of a transaction between her aptitudes and the particular characteristics of the situation in which that performance occurs over time. By aptitudes he meant the cognitive and motivational resources that an individual brings to the situation. By situation he meant the characteristics of a particular environment that "afforded or impeded"—that assisted or constrained—certain courses of goal-related action. When Snow applied his ideas to achievement testing situations, he recognized that test performance resulted from a student's background and intellectual history, as well as the cognitive and motivational resources that the student cobbled together to respond to a series of situation-embedded test tasks (e.g., multiple-choice items or performance assessments). In essence, Snow's idea was that individuals' achievement test performance depended not just on their knowledge and abilities, but rather on a full spectrum of interrelated, situation-relevant cognitive and motivational resources that transacted with the affordance, constraint, and demand structures of the testing tasks themselves. From this reasoning and from empirical findings, he concluded that a new multivariate approach to validating interpretations of achievement test scores was needed. We first set forth in more detail Snow's new aptitude theory as it applies to multivariate test validity and then describe the design of the "High School Study" we conducted as a step in exploring his ideas about validity.

In 1995, Richard E. Snow wrote in a proposal to the Office of Educational Research and Improvement (Baker, Linn, & Herman, 1995) that his previous work (e.g., Hamilton, Nussbaum & Snow, 1997; Nussbaum, Hamilton, Snow, 1997) showed that "psychologically meaningful and useful subscores can be obtained from conventional achievement tests" (Baker et al., p. 133). He went on to point out that these subscores represented important ability distinctions and showed different

patterns of relationships with demographic, "affection" (temperamental-emotional), "conative" (motivational-volitional), and instructional-experience characteristics of students. He concluded that "a *new multidimensional approach to achievement test validation* should include affective and conative as well as cognitive reference constructs" (italics ours; Baker et al., p. 134). We begin by expanding on these ideas and building a framework for conceiving multidimensional validity and then describe our "High School Study."

Framework

Snow called for broadening the concept of aptitude to recognize the complex and dynamic nature of person-situation interactions and to include affective and conative processes in explaining individual differences in learning and achievement. He posited that a person's performance was a function of a broad set of aptitudes and the affordances and constraints of a particular situation, and that two general pathways could describe the manner in which these resources played out (Snow, 1994; see also Stanford Aptitude Seminar, 2002).

The first was what he called a "performance pathway"—a concept that denoted the dynamic process by which cognitive resources are activated, retrieved, assembled, and executed in the service of accomplishing particular tasks in particular situations. The other, parallel, hypothesized pathway described by Snow was the commitment pathway—a concept that denoted the process by which motivational resources are activated in the service of energizing and guiding behavior toward particular goals in a given situation. In this person-situation transaction, a person cobbles together a combination of cognitive and motivational aptitudes—an "aptitude complex"—for addressing relevant task and situation-specific goals (e.g., performance).

More specifically, Snow viewed an individual's performance (e.g., test performance) as a transaction between his aptitudes and the particular situation in which that performance takes place. By aptitudes Snow meant all those characteristics (e.g., experience, ability, motivation, beliefs) that an individual brings to and cobbles together to perform in a particular situation. He called this situation-elicited set of aptitudes an aptitude complex. If the performance were requested at another time, the person might attend to different aspects of the situation (test) and bring a somewhat different aptitude complex to bear. By situation Snow meant the external environment that "affords" and "constrains" a particular individual's

performance, as does a test by the very nature of its items, including their content and format. And by transaction he meant that an individual interacts with an environment iteratively to produce an observable performance. To see how this works, consider an individual attempting to solve a physics test problem on force and motion. She brings prior experience, prior learning opportunities, knowledge, reasoning ability, personal goals, a sense of competency and an interest in physics (and so on) to bear in this situation. At the outset, motivation may shape the likelihood of her even attempting to solve the problem. If she isn't interested or feels incompetent, she may never fully engage, or she may approach the problem in a cursory way, giving up easily. If she were emotionally engaged or felt competent enough to approach the task, her sense of expertise and her interest would most likely influence her performance.

If she were a physics novice, the surface features of the task—whether the task involves a pendulum or an object moving on a rough surface—would attract her. That is, given the aptitude complex brought to bear, the situation affords certain information for the problem solver and constrains other information due to lack of expertise. However, if expert, she would be attracted by the particular physical representation (the problem) of an underlying law of force and motion, and the surface features would not afford and constrain her performance in the same way as they do for the novice (e.g., Chi, Feltovich, & Glaser, 1981). Moreover, her interest in physics and concomitant outside-of-school exposure to science might affect her problem-solving performance at this point. Relationships between the current problem and other memories (e.g., seeing a pendulum at the Smithsonian) may provide additional resources for her to draw on.

Finally, problem-solving strategies are also part of her aptitude complex. Regardless of whether she is expert or novice, the interaction between her and the situation becomes an evolving transaction over time until the problem is solved (or she quits!). In short, her actions create new task environments continuously, and such environments "feed back" to her, modifying the kinds of cognitive and motivational resources that she needs to invest next. In this way, the interaction between person and situation becomes an evolving transaction over time until the problem is exited.

As we shall see, over a century of research has characterized the aptitude side of the transaction in great detail; we struggle to characterize the situation side. We turn first, then, to aptitudes. A sketch of our framework is shown in Figure 1.

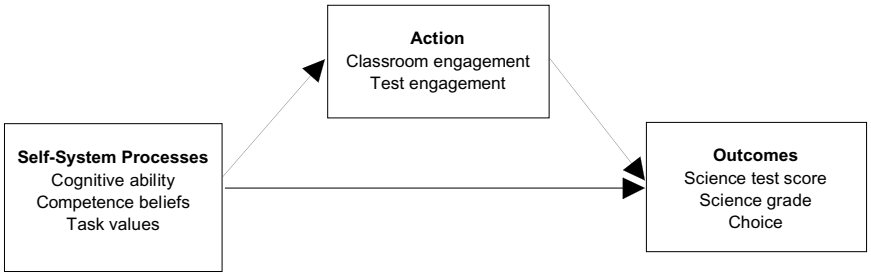

Figure 1. Conceptual model.

Aptitudes

For simplicity, we speak of cognitive and motivational aptitudes. We recognize that this simplicity masks the complexity of what is known about variability among individuals and the detail in which this variability has been characterized.

Cognitive aptitudes. Our conceptual framework for the cognitive domain is derived from Carroll's three-stratum model of human cognitive ability (Carroll, 1993). We focus on three broad abilities—fluid ability, crystallized (verbal and quantitative) ability, and visualization—because they are most relevant to performance on the kinds of science tasks (situations) we examine in this research. These broad, cognitive factors are considered as important aptitudes that reflect students' native abilities and learning histories. Fluid ability refers to reasoning ability that generalizes across a vast number of situations. It is often measured by the speed and accuracy with which an individual can trace through a maze or find a figure hidden in another figure. Crystallized abilities are built up by experience, especially formal schooling. Verbal and quantitative ability are familiar to anyone who has taken the SAT or the GRE. And visualization refers to the ability to mentally visualize and manipulate objects. Visualization tests often ask a person to predict what a figure would look like if rotated 45 degrees, or what a folded piece of paper with a hole punched in it would look like if it were opened. These cognitive aptitudes are organized as a repertoire of mental frameworks, response sets, knowledge and skill components, and heuristic problem-solving strategies (see

Snow, 1992). During performance, different mixes of these cognitive resources are activated and coordinated to perform under particular situational task demands.

Motivational aptitudes. Our conceptual framework for motivational aptitudes focuses on two basic sets of processes that help individuals to evaluate and prepare for actions in specific situations: task-related competence beliefs and task-related values and goals. These constructs are at the heart of several contemporary social-cognitive views of motivation, including expectancy-value theory, self-efficacy theory, and goal theory (see Eccles, Wigfield, & Schiefele, 1997).

Social-cognitive theories of motivation posit that students' self-perceptions of competence, their values and goals, and their emotional experiences learning in a specific academic domain (in conjunction with domain-relevant cognitive abilities) influence their performance and quality of engagement in that domain (Ford, 1992; Snow, 1989). Competence beliefs are most closely tied to performance, whereas goals, emotions, and values are most closely tied to behavioral choice and engagement (Eccles et al., 1997).

Individuals' task-related expectancies for success and values serve the function of preparing and energizing them to engage with a task, to seek out task challenges, to persist at particular tasks, and to choose certain activities in their free time (Eccles-Parsons et al., 1983).

Eccles-Parsons et al. (1983) defined expectancy as individuals' beliefs about how well they would perform on future tasks in a given domain. A closely related construct is Bandura's notion of self-efficacy. Bandura (1997) defined self-efficacy as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (p. 3). In a review of the contribution of perceived self-efficacy to cognitive functioning, Bandura (1993) explicated diverse pathways through which self-efficacy exerts its impact. For example, in a study of mathematics skills development, self-efficacy was found to enhance the mastery of mathematics skills directly by affecting the quality of thinking and use of acquired knowledge and skills, and indirectly by increasing persistence in the search for task solutions (Schunk, 1984).

In contrast to beliefs in which individuals evaluate "Can I do this task?" are beliefs about the values or goals that guide action, the "Why am I doing this task (or not)?" beliefs. Eccles-Parsons et al. (1983) defined achievement task values as individuals' perceived importance of and intrinsic interest in certain tasks, their

perceived utility of a given task in relation to the attainment of other desired goals, and the perceived cost of engaging in a particular task. In general, the motivational dictum "We devote attention, time and resources to that which we value" is a truism and includes the academic domains, like math and science. Eccles-Parsons et al. (1983) showed that, whereas expectations for success were most closely related to high school students' performance in mathematics, their valuing of math was most closely tied to the subsequent enrollments in math-related courses. Values tend to be most strongly tied to patterns of behavioral choice.

In one model of motivated action, called self-system theory, Connell and Wellborn (1991) proposed that motivational processes indirectly, rather than directly, affect performance outcomes through their influence on whether or not an individual attends to, persists in, and engages a task. That is, these researchers posited that an individual's motivational beliefs (perceived competence and values) affect the individual's willingness to cognitively, attentionally, and emotionally engage with a task, with consequent influences on task performance. In their terminology, motivational processes of the self-system are related to patterns of action, which in turn affect outcomes (see Figure 1).

Multidimensional Situations: Science Achievement Tests

Snow's research, using a mixed methodology of large-scale statistical analyses and small-scale interview studies, demonstrated the usefulness of a multidimensional representation of high school science achievement. Specifically, he and his students identified three distinct constructs underlying students' performance on a standardized test and sought validation evidence for the distinctions between "basic knowledge and reasoning," "quantitative science" reasoning, and "spatial-mechanical" reasoning (see Hamilton et al., 1997; Nussbaum et al., 1997). Basic knowledge and reasoning draws on general verbal reasoning using declarative (facts, concepts) knowledge. Quantitative science reasoning involves the manipulation of numerical quantities and requires specialized classroom-based knowledge. And spatial-mechanical reasoning requires reasoning about visual or spatial relations, motions, distances, or some combination of these. (See Table 1 for a description, with examples, of the three constructs.) Of particular note was the finding that these reasoning dimensions showed different patterns of relationships with student background, instructional practices, and out-of-school activities. These different relationships provided the groundwork for understanding the essential characteristics of each dimension. For example, gender differences in

Table 1

Description of Three Reasoning Dimensions (from Schultz, Ayala, & Shavelson, 2001, after Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995)

Dimension	Example items
Basic knowledge and reasoning	
General characteristics	Choose an improvement for an experiment on mice
Reflects general knowledge	Identify the example of a simple reflex
Involves greater use of general reasoning	Choose the property used to classify substances
Requires more verbal reasoning than quantitative science or spatial-mechanical	Select statement about the process of respiration
	Explain the location of marine algae
Item characteristics	Choose best indication of an approaching storm
Content areas include biology, astronomy, and chemistry. General themes in science are also included. For example, experimental design or the difference between a model and an observation	Choose alternative that is not chemical change
	Select basis for statement about food chains
	Distinguish model from observation
	Read population graph: identify equilibrium point
	Identify cause of fire from overloaded circuit
	Explain the harmful effect of sewage on fish
Quantitative science reasoning	
General characteristics	Read a graph depicting the solubility of chemicals
Application of advanced concepts	Read a graph depicting digestion of protein enzyme
Manipulation of numerical quantities	Infer from results of experiment using filter
Requires specialized course-based knowledge	Explain reason for ocean breezes
Item characteristics	Interpret symbols describing a chemical reaction
Content includes chemistry and physics content	Calculate a mass given density and dimensions
	Calculate grams of substance given its half life
Numeric calculations	Calculate emissions of radioactive decay
	Choose method of increasing chemical reaction
	Predict path of ball dropped in moving train
Spatial-mechanical reasoning	
General characteristics	Choose a statement about source of moon's light
Requires reasoning and interpretation of visual or spatial relationships, motions and / or distances	Answer question about Earth's orbit
	Locate the balance point of a weighted lever
	Interpret a contour map
Item characteristics	Identify diagram depicting light through a lens
Content includes astronomy, optics and levers	Predict how to increase period of pendulum

science achievement could be attributed to spatial-mechanical reasoning but not to quantitative science reasoning or basic knowledge and reasoning.

Though Snow and others (e.g., Li & Shavelson, 2001) have made inroads on characterizing the multidimensional cognitive nature of achievement test situations, we do not know of a parallel characterization for the motivational aptitudes. In the following description of our High School Study, we provide a brief sketch of how the motivational aptitudes just described might characterize an achievement test setting and leave the explanation of the link to Haydel and Roeser (2002).

High School Study Design

Snow left hints of what he meant by "a new multidimensional approach" when he wrote in the same proposal, "the primary objective of this study is to determine if knowledge and ability distinctions previously found important in high school math and science achievement tests occur also in other multiple-choice and constructed response assessments. . . . A second objective is to examine the cognitive and affective correlates of these distinctions. And a third objective is to examine alternative assessment designs that would sharpen and elaborate such knowledge and ability distinctions in such fields as math, science, and history-geography" (Baker et al., 1995, p. 133). Accordingly, the High School Study focused on 10th- and 11th-grade students. The study design was correlational in nature and included biographical and motivational surveys and three types of science achievement tests: multiple choice, constructed response, and performance assessment.

Participants and Procedures

Four hundred ninety-one 10th-grade (53%) and 11th-grade (47%) students born between 1982 and 1984 participated. They were enrolled in Earth science, chemistry or biology classes in a northern California high school. The sample was half female (51%) and ethnically mixed: 49% European-American, 27% Latino, 8% African American, 8% Asian American, and 8% other. Of the non-native English speakers, 80% reported that they understood English very well; only two students reported their ability to understand English as "not very well." Most students came from homes with well-educated parents; approximately two thirds of the students' parents attended 4 or more years of college. Though all students participated in almost all aspects of the study, for any randomly selected variable in the survey, we had roughly 10% missing data.

Survey measures of students' general motivational orientation in science and achievement test measures of their mathematical and verbal ability were collected in students' science classrooms in the 1999-2000 school year. Trained research assistants administered the surveys and tests during separate class periods. The researchers returned about one month later to administer measures of aptitude and science achievement. Moreover, to measure test engagement, a survey (Post-Science Test Survey) was administered immediately after students took the science achievement test. Finally, performance assessments (see Ayala, Yin, Schultz, & Shavelson, 2002) were given to a subsample of 35 students in the summer of the academic year.

Instrumentation

The instrumentation followed the conceptual framework in collecting information on students' background (demographic), aptitudes (both cognitive and motivational), and test performance (multiple choice, constructed response, performance). We briefly describe each in turn.

Demographic survey. The demographic survey collected information on age, gender, ethnicity, grade level, course taking, and language background.

Cognitive aptitude tests. Four measures were used to tap students' fluid, crystallized (verbal and quantitative), and spatial abilities. Two tests from the Educational Testing Services Kit (French, Ekstrom, & Price, 1963) were administered. One, the Hidden Figures Test (internal consistency reliability = .66), measured fluid ability, and the other, the Cube Comparison Test (reliability = .72), measured spatial-visualization ability. The other two tests measured crystallized abilities. The quantitative ability test was composed of items from the NELS:88 mathematics test (reliability = .83) that had been analyzed earlier (Kupermintz & Snow, 1997). The verbal ability test was composed of items from a practice Standardized Achievement Test (reliability = .68). Finally, a composite of these cognitive aptitude tests was formed based on a principal components analysis that showed that a single dimension could capture the important covariation among the four aptitude tests. The composite was standardized with mean 0, standard deviation 1, with a reliability of .67.

Motivational aptitudes. Information about students' motivation was collected with two questionnaires, the Beliefs and Attitudes Toward Science Survey (N_i = 147) and the Post-Science Test Survey (N_i = 15). Of particular importance are the

measures we constructed to tap motivational patterns, processes, and students' patterns of "action" (e.g., engagement or disaffection). Here we give an overview of these measures and leave the details to subsequent, relevant reports. The study's Construct Codebook provided details linking survey questions to these and other measures along with their reliability.

Motivational processes were of two types: competency beliefs (what Bandura, 1997, called efficacy beliefs) and task values. We formed a composite measure of competency beliefs from questionnaire items dealing with students' beliefs about their ability to master science content, their ability to perform well on different types of science assessments (Bandura, 1997), and their confidence in their abilities in the domain of science (Dweck, 1986). We formed a task value composite from questionnaire items inquiring into students' values about science, including interest, usefulness, and importance (Eccles & Wigfield, 1995).

Following self-system theory, we also characterized students' patterns of action in relation to science learning. Measures included students' self-reported activities both during class and during our science achievement test. Classroom engagement was measured by students' self-reports of how much they paid attention in class and participated in science activities, by how much homework they completed, and by how much they were involved in self-regulated learning activities. To measure test engagement, the Post-Science Test Survey was administered immediately after students took the science achievement test. Students were asked about their use of cognitive strategies, mood, energy level, and effort during the science test.

Types of Achievement Tests

Snow and colleagues had thoroughly examined the structure of the science test used in the National Education Longitudinal Study of 1988 (Hamilton et al., 1997; Nussbaum et al., 1997). They found three reasoning dimensions underlying the pattern of students' scores: basic knowledge and reasoning, quantitative science reasoning and spatial-mechanical reasoning. Snow wondered whether this structure was unique to the NELS:88 multiple-choice science test or might be characteristic of other large-scale tests. To satisfy Snow's curiosity and to follow his study plan, we built a science achievement test for this study with questions from the NELS:88, the Third International Study of Mathematics and Science Achievement (TIMSS) and the National Assessment of Educational Progress (NAEP). We specifically selected items, on the basis of their content and format, to fall within one or another of the

three reasoning dimensions (see Table 2 for a summary of the multiple-choice test item allocations).

Items were both multiple choice and constructed response in nature. The multiple-choice test items were drawn from NELS:88, NAEP, and TIMSS. An example multiple-choice item is:

1. What does a mitochondrion do in a cell?

   A. It controls the transport of substances leaving and entering the cell.

   B. It contains the information to control the cell.

   C. It produces a form of energy that the cell can use.

   D. It breaks down waste products in the cell.

The constructed response items were drawn from TIMSS. They also were selected to reflect the three reasoning dimensions (see Table 3).

An example constructed response question is:

31. The sketch below shows two windows. The left window has been cracked by a flying stone. A tennis ball, with the same mass and speed as the stone, strikes the adjacent similar window, but does not crack it. . . .

What is one important reason why the impact of the stone cracks the window but the impact of the tennis ball does not?

[sketch of windows]

Table 2

Source and Description of Science Multiple-Choice Items

Item number	Source	Dimension	Description
S01	NAEP8	BKR	What does a mitochondrion do in a cell
S02	NAEP8	BKR	How insulated bottle keeps a liquid cold
S03	NAEP8	BKR	Force responsible for Solar System formations
S04	NAEP8	BKR	Input/output energy forms for a stereo system
S05	NAEP12	BKR	Considerations in planning a nuclear power facility
S06	TIMSS8	QS	Ammeter measurements in circuit
S07	TIMSS8	QS	Balance weights on seesaw
S08	TIMSS8	SM	Diagram of projected ball move from a curved groove
S09	NELS12	BKR	Explain harmful effects of sewage on fish
S10	NELS12	BKR	Explain location of marine algae
S11	NELS12	QS	Calculate grams of substance given its half life
S12	NELS12	BKR	Identify the example of a simple reflex
S13	NELS12	QS	Calculate emission of radioactive decay
S14	NELS12	BKR	Recognize picture of tissue
S15	NELS12	SM	Locate the balance point of a weighted lever
S16	NELS12	QS	Calculate a mass given density and dimensions
S17	NELS12	BKR	Select statement about the process of respiration
S18	NELS12	QS	Read graph depicting digestion of protein by enzyme
S19	TIMSS12	BKR	Statements about liquid evaporation
S20	NELS12	QS	Choose method for increasing chemical reaction
S21	TIMSS12	BKR	Why steam produced when water boils
S22	TIMSS8	QS	Describe pattern in a table
S23	TIMSS12	BKR	Properties of molecules of different gasses
S24	TIMSS12	QS	Graph representing block oscillations on end of spring
S25	TIMSS12	BKR	Fusion in nuclear energy generation
S26	NELS12	SM	Interpret a contour map
S27	TIMSS12	QS	Diagram of gas pressure against temperature
S28	TIMSS12	BKR	Use of CFC
S29	NAEP12	QS	Calculate grams of oxygen and carbon reaction
S30	NELS12	SM	Identify diagram depicting light through a lens

Note. BKR = basic knowledge and reasoning; QS = quantitative science reasoning; SM = spatial-mechanical reasoning; NAEP = National Assessment of Educational Progress; NELS = National Education Longitudinal Study; TIMSS = Third International Mathematics and Science Study.

Table 3

Source and Description of Science Constructed Response Items

Item number	Source	Dimension	Description
S31	TIMSS	BKR?	Why did the impact of the stone and not the tennis ball break the window?
S32	TIMSS	SM	Draw a picture of the pencil as you would see it in the mirror.
S33	TIMSS	QS	Given data, which machine is more efficient?
S34	TIMSS	BKR?	Is the amount of light energy more than, less than, or the same as the amount of electrical energy used? Why?
S35	TIMSS	QS	The number of bacteria was growing exponentially from 1PM, how many at 6PM?
S36	TIMSS	BKR?	What will happen to the water level when an ice cube melts? Explain.
S37	TIMSS	BKR?	A car moving at constant speed comes toward you and then passes. Describe the change in the frequency of sound you hear.
S38	TIMSS	SM	Draw a line in the diagram (watering can) where the surface of the water is now.

Note. BKR = basic knowledge and reasoning; QS = quantitative science reasoning; SM = spatial-mechanical reasoning; TIMSS = Third International Mathematics and Science Study.

Finally, following Snow's design calling for performance assessment, we included three performance assessments. The Electric Mysteries assessment (Shavelson, Baxter, & Pine, 1991) asked students to determine the contents of six mystery boxes (e.g., wire, battery and bulb, two batteries) by hooking up an external circuit to each box. This assessment was selected to measure basic knowledge (e.g., series circuits) and reasoning (ruling in and ruling out the boxes' possible contents by using different tests).

The Acquacraft assessment (Ayala, Ayala, & Shavelson, 2001) asked students to determine the cause of an explosion aboard a submarine by simulating what might have happened when copper sulfate was added to aluminum ballast tanks using glassware, copper sulfate, aluminum, salt, and matches. In order to perform the task, students had to apply advanced science procedures (i.e., testing unknown gases), manipulate numerical quantities, and use specialized course-based knowledge—the general characteristics of the quantitative science dimension. This assessment was used to tap into students' quantitative science reasoning.

And finally, Daytime Astronomy gave students an Earth globe in a box, a flashlight, and a set of "sticky towers." Students then used the flashlight as if it were the Sun to project shadows with the towers to determine the time and location of places on Earth. To solve Daytime Astronomy problems, students had to use spatial observation, modeling, and reasoning, all features of the spatial-mechanical reasoning dimension (Solano-Flores, Jovanovic, & Shavelson, 1994; Solano-Flores & Shavelson, 1997; Solano-Flores et al., 1997). In another report (Ayala et al., 2002), we provide rather extensive descriptions of these assessments because of their relative uniqueness compared with the other two science test formats.

References

Ayala, C. C., Ayala, M. A., & Shavelson, R. J. (2001). *On the cognitive interpretation of performance assessment scores* (CSE Tech. Rep. No. 546). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Ayala, C. C., Yin, Y., Schultz, S., & Shavelson, S. (2002). *On science achievement from the perspective of different types of tests: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 572). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baker, E. L., Linn, R. L., & Herman, J. L. (1995). *Institutional grant proposal for OERI Center on Improving Student Assessment and Educational Accountability: Integrated assessment systems for policy and practice: Validity, fairness, credibility, and utility*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28,* 117-148.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

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

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Self-processes in development: Minnesota Symposium on Child Psychology* (Vol. 23, pp. 43-77). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 40,* 1040-1048.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the achiever: The structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin, 21,* 215-225.

Eccles, J. S., Wigfield, A., & Schiefele, U. (1997). Motivation to succeed. In N. Eisenberg (Series Ed.) & W. Damon (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 1017-1095). New York: Wiley.

Eccles-Parsons, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75-146). San Francisco: Freeman.

Ford, M. (1992). *Motivating humans.* Newbury Park, CA: Sage Publications.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Kits of reference tests for cognitive factors.* Princeton, NJ: Educational Testing Service.

Hamilton, L., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I. M., & Snow, R. E. (1995). Enhancing the validity and usefulness of large scale educational assessments: II. NELS:88 science achievement. *American Education Research Journal, 32,* 555-581.

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

Haydel, A. M., & Roeser, R. W. (2002). *On the links between students' motivational patterns and their perceptions of, beliefs about, and performance on different types of science assessments: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 573). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kupermintz, H., & Roeser, R. (2002). *Another look at cognitive abilities and motivational processes in science achievement: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 571). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS:88 math achievement to 12th grade. *American Educational Research Journal, 34,* 123-149.

Lau, S., Roeser, R. W., & Kupermintz, H. (2002). *On cognitive abilities and motivational processes in students' science engagement and achievement: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 570). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments. IV. NELS:88 science achievement to 12th grade. *American Educational Research Journal, 34,* 151-173.

Schultz, S., Ayala, C. C., & Shavelson, S., and the SA Project. (2001, April). Examining high school students' science achievement with different types of science assessments. In H. Kupermintz (Chair), *Integrating the study of cognitive abilities and motivational processes: High school students' science engagement and achievement*. Symposium presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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

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

Shavelson, R., & Lau, S. (2002). *Multidimensional validity revisited* (CSE Tech. Rep. No. 574). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Shavelson, R., Roeser, R., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., & Schultz, S. (2002). *Conceptual framework and design of the High School Study: A multidimensional approach to achievement validation* (CSE Tech. Rep. No. 569). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Snow, R. E. (1989). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology* (pp. 435-474). Hillsdale, NJ: Lawrence Erlbaum Associates.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27*, 5-32.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionist perspectives on human intelligence* (pp. 3-37). Cambridge: Cambridge University Press.

Solano-Flores, G., Jovanovic, J., & Shavelson, R. J. (1994, April). *Development of an item shell for the generation of performance assessments in physics.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: conceptual, practical and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16-25.

Solano-Flores, G., Shavelson, R. J., Ruiz-Primo, M. A., Schultz, S. E., Wiley, E., & Brown, J. H. (1997, March). *On the development and scoring of observation and classification science assessments.* Paper presented at the annual meeting of American Educational Research Association, Chicago.

Stanford Aptitude Seminar [Corno, L., Cronbach, L. J. (Ed.), Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E.]. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ: Lawrence Erlbaum Associates.