On Science Achievement From the Perspective of Different Types of Tests: A Multidimensional Approach to Achievement Validation

CSE Technical Report 572

Carlos Cuauhtémoc Ayala, Yue Yin, and Susan Schultz
Stanford University

Richard Shavelson
CRESST / 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 569, 570, 571, 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.

ON SCIENCE ACHIEVEMENT FROM THE PERSPECTIVE OF DIFFERENT TYPES OF TESTS*

Carlos Cuauhtémoc Ayala, Yue Yin, and Susan Schultz, Stanford University

Richard Shavelson, CRESST/Stanford University

Abstract

Students bring to achievement tests a complex mix of cognitive, motivational, and situational resources to address the tasks at hand. Previous research (Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Hamilton, Nussbaum, & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997) has demonstrated the usefulness of a multidimensional representation of science achievement and, in particular, three reasoning dimensions: basic knowledge and reasoning, spatial-mechanical reasoning, and quantitative science reasoning. Though other authors in this set of reports look at the different patterns of student cognitive, motivational, and situational responses as they predict science achievement, our focus is on the science achievement measures and on their relationships with the three reasoning dimensions. Thirty multiple-choice items, 8 constructed response item and 3 performance assessments, each nominally assigned to one of the reasoning dimensions, were administered to 35 students—a representative subsample of the whole study (N = 341). We found that the different measures of science achievement were moderately correlated with each other, suggesting that these measures tap into somewhat different aspects of science achievement, as expected. We also found that the correlational patterns of student scores on items of like reasoning dimensions did not group as expected, and that student knowledge and experience seemed to suggest how a student solved a problem and not the problem alone. We therefore concluded that the nominal assignment of our items to three reasoning dimensions was problematic.

Dick[1] prepares to complete a statewide science achievement test. He has studied hard, yet little does he know that his performance today reflects more that just his last-minute cramming. He brings to his performance a complex mix of cognitive, motivational, and situational resources to address the task at hand. Snow (1992) believed that these broad aptitude factors reflected students' learning histories—arranged as a collection of mental schemes, response sets, knowledge and skill components, and heuristic problem-solving strategies. These broad aptitudes will be brought to bear on a multidimensional achievement test.

---

*An earlier version of this report was presented at the annual meeting of the American Educational Research Association in Seattle, Washington, in April 2001, under the title *Examining High School Students' Science Achievement With Different Types of Science Assessments: A Perspective From Reasoning.*

[1] In memory of Richard E. Snow.

Hamilton, Nussbaum, and Snow (1997) found three *reasoning dimensions* underlying students' science achievement on the National Education Longitudinal Study of 1988 (NELS:88)—basic knowledge and reasoning, quantitative science reasoning, and spatial-mechanical reasoning—that were confirmed with small-scale interviews (Hamilton et al., 1997; Nussbaum, Hamilton, & Snow, 1997).

The purposes of this study were to determine whether performance assessments could be explicitly designed to tap these three reasoning dimensions; to validate interpretations of science achievement test scores as reflecting these dimensions, including scores on performance tests; and to examine the consistency of student performance across the three achievement measures.

Reasoning Dimensions

We posited three reasoning dimensions following Snow's earlier work. These three dimensions emerged from an analysis of the National Education Longitudinal Study of 1988 (NELS:88) science achievement data (Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995). (See Shavelson et al., 2002, for descriptions and sample items for the three dimensions.) Factor analysis of the NELS:88 science achievement data suggested three reasoning and knowledge dimensions: basic knowledge and reasoning, quantitative science reasoning, and spatial-mechanical reasoning. Corroborating evidence supporting the three reasoning dimensions came from think-aloud protocols, observations and posttest interviews (Hamilton et al., 1997). Furthermore, Hamilton and Snow (1998) identified some of the salient features of multiple-choice and constructed response items that revealed the largest difference in scores. For example the spatial-mechanical dimension, which revealed a gender effect, could be differentiated from the other reasoning dimensions based on students' more frequent use of predictions, gestures, and visualization.

We set out to see whether other multiple-choice, constructed response, and performance assessments nominally fit into the reasoning dimensions. In addition to the set of 30 multiple-choice and 8 constructed response items drawn from NELS, the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS), 3 science performance assessments were selected to reflect one or another of the three reasoning dimensions. We included performance assessments because of their link with science education (inquiry) reform and evidence that they measure a somewhat different aspect of science achievement (procedural and schematic knowledge) than traditional tests (Li & Shavelson, 2001; Shavelson & Ruiz-Primo, 1999).

Performance Assessment Selection

To select the performance assessments, we classified previously published assessments into the three reasoning dimensions (Ayala, Shavelson, & Ayala, 2001). To do this, we examined performance assessment tasks, response demands, and scoring systems and determined which general characteristics of each dimension most closely matched each performance assessment. For example, the Paper Towels investigation (Baxter, Shavelson, Goldman, & Pine, 1992) asked students to determine which of three paper towels absorbed the most/least water with a scoring system focusing on the scientific justifiability of their procedures as well as the accuracy of their inferences. Since Paper Towels involved general science experimentation, general reasoning, and no specific science content (chemistry, biology, physics), we concluded that this assessment fell into the basic knowledge and reasoning category. A total of 27 performance assessments were analyzed by this method, of which 25 assessments were classified as basic knowledge and reasoning, 2 were classified as spatial-mechanical, and none was classified as quantitative science (A. Ruiz-Primo, 1999, personal communication; see Appendix). In order to fill the quantitative science void, a new performance assessment was created that fit the characteristics of the quantitative science category.

In selecting the performance assessments to represent the three reasoning dimensions, we also sought assessments that fell into the content-rich and process-open quadrant of Baxter and Glaser's (1998) Content-Process Space. This quadrant was expected to produce the most scientific reasoning. A performance assessment was content-rich if it required specific content knowledge to succeed. It was process-open if students had to come up with their own procedures for carrying out an investigation rather than follow a procedure or "recipe." And because reasoning demands are related to tasks (Baxter & Glaser, 1998), we selected assessments to represent different task types as defined by Shavelson, Solano-Flores, and Ruiz-Primo (1998): (a) *comparative* investigations, in which students compare two or more objects, and their performance is evaluated for accuracy of procedures and inferences; (b) *component-identification* investigations, in which the task is to decompose a whole (electric mystery box) into its components parts (wire, battery, bulb, etc.) by various procedures (e.g., connecting an external circuit), and performance is evaluated as to confirming and disconfirming evidence; (c) *taxonomic* investigations, in which students construct a taxonomy for a particular purpose such as predicting which objects would sink or float based on volume and mass, with

performance evaluated as the accuracy of the classifications; and (d) *observation* investigations, in which students observe and model a process over time, and their performance is evaluated as to the accuracy of the observations, models and inferences. At a later date, we plan to compare reasoning dimensions and task types.

Ultimately, we selected Electric Mysteries as our basic knowledge and reasoning performance assessment because general knowledge of series circuits and general reasoning could be used to perform the tasks (Shavelson, Baxter, & Pine, 1991). Students are given batteries, bulbs, and wires and asked to connect them to each of six "mystery" boxes to determine the boxes' contents—wire, nothing, two batteries, etc. (see Figure 1). Baxter and Glaser (1998) found Electric Mysteries to be content-rich and process-open because students had to know how electric circuits worked and had to determine their own procedures for finding the contents of the mystery boxes. Shavelson et al. (1998) considered Electric Mysteries to be a component identification investigation task because students had to determine the components in each box.

Electric Mysteries consists of selecting from five choices what circuits are found in six boxes (one is a repeat). Each task uses the same equipment, and no procedures are given to complete the tasks, organize data, or find the correct solution. These six tasks are interchangeable, although some circuits are more complicated to solve than others. The scoring form asks raters to evaluate the circuit used to determine

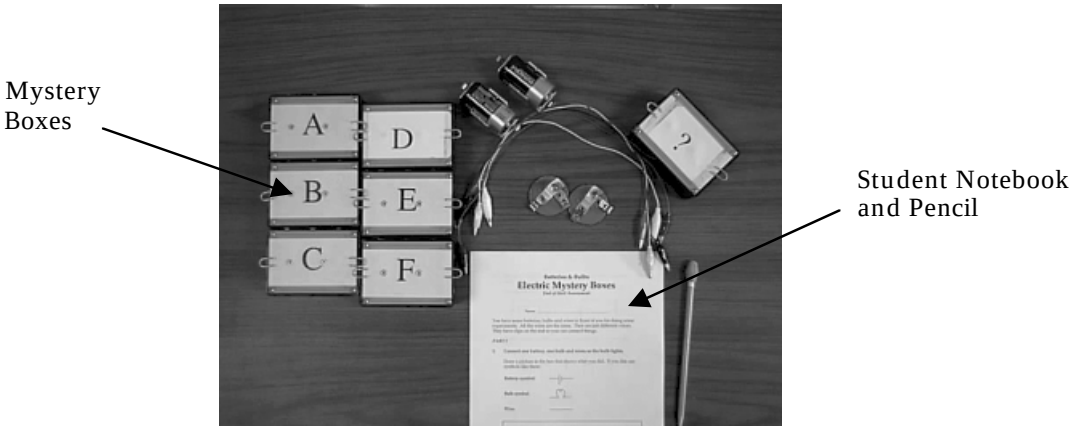

Figure 1. Electric Mysteries performance assessment.

the contents of the box and to determine whether the student did indeed determine the boxes' contents. Scoring was straightforward and reliable (interrater reliability > .90).

Next, we selected Daytime Astronomy as our spatial-mechanical performance assessment because solving it requires spatial observation, modeling, and reasoning, features of the spatial-mechanical reasoning dimension (Solano-Flores et al., 1997; Solano-Flores, Jovanovic, & Shavelson, 1994; Solano-Flores & Shavelson, 1997). Students are given an Earth globe in a box, a flashlight, and a set of "sticky towers" (see Figure 2). Students then use the flashlight as if it were the Sun to project shadows with the towers to determine the time and location of places on Earth. The task requires knowledge of the Sun's position in relation to Earth, Earth's rotation, and the relationship between the position of the Sun and shadows cast on Earth. Consequently, this task is considered content-rich. Because students are not given directions on how to carry out these tasks, the assessment is considered process-open. Because students are asked to model the path of the Sun across the sky and to use shadow direction, length, and angle to solve location problems, Solano-Flores and Shavelson (1997) considered this assessment to be of the observation investigation task type.

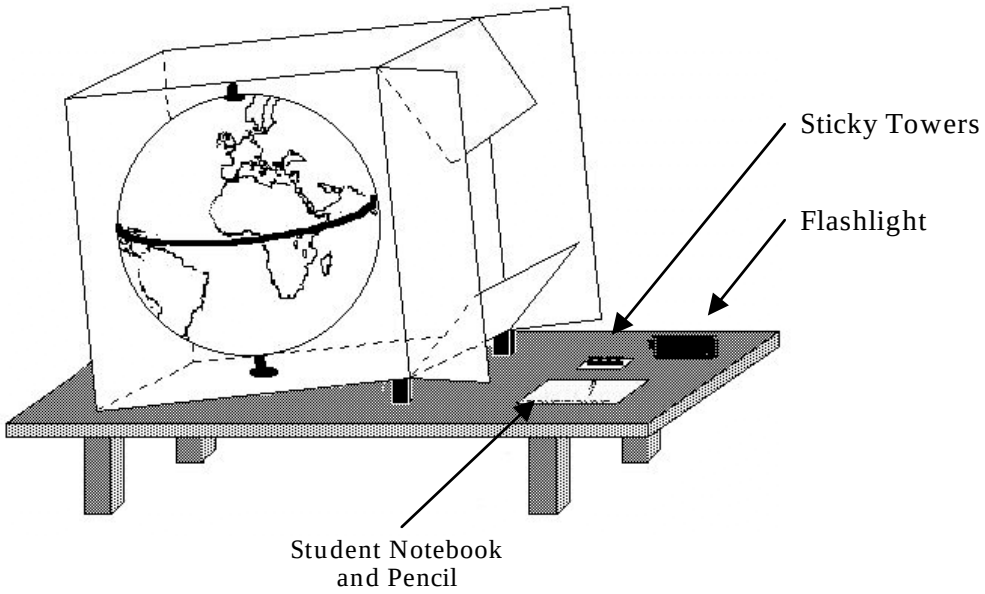

Figure 2. Daytime Astronomy performance assessment.

The Daytime Astronomy performance assessment is divided into six separate tasks, some more closely related than others, but all are designed to tap into a student's understanding of the motion of the Sun in relation to Earth and the shadows that this relationship produces. Although Solano-Flores and Shavelson (1997) said that this assessment could be completed with fewer tasks, we administered all six tasks since our respondents were high school students instead of fifth graders, for whom the assessment was originally created. The Daytime Astronomy scoring form is more complex than the Electric Mysteries scoring form. It consists of rating the accuracy of the student's observations, data gathering and modeling skills, and explanations of the tasks. Solano-Flores and Shavelson reported an interrater reliability of .90.

Finally, we developed a new investigation, Aquacraft, as a quantitative science assessment to match important components of the chemistry curriculum (Ayala et al., 2002). High school chemistry teachers verified that its content was consistent with the students' chemistry curriculum. Students are asked 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 (Figure 3).

Students determine the cause of an explosion using high school chemistry principles and procedures, select the appropriate chemical equations to represent

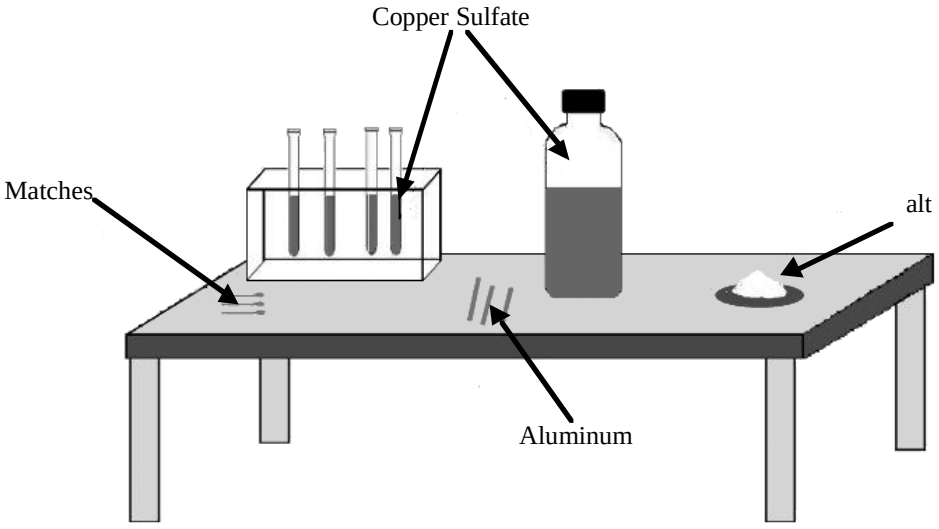

Figure 3. Aquacraft performance assessment.

the reaction, and determine quantitatively the amount of energy released in the explosion. In order to perform the task, students have 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. Because advanced science content knowledge and specialized skills are needed to complete the task, it is considered content-rich. And, because students conduct their own investigations without step-by-step instructions, it is considered process-open. Finally, because students are expected to compare chemical reactions in both fresh and salt water, we considered Aquacraft to be a comparative investigation.

Aquacraft consists of four tasks. Task 1–Chemical Reaction and Task 2–Test the Gas consists of observing and comparing two possible scenarios (copper sulfate in fresh water vs. copper sulfate in salt water), using chemistry procedures and lab techniques. Task 3–Balancing Equations requires students to select the appropriate chemical equations for the reaction. Task 4–Energy Calculations asks students to determine quantitatively whether there was enough energy released in the explosion to cause the reported damage. All four tasks tap into the quantitative science reasoning dimension in different ways, the first two via chemistry content and lab techniques and the last two via chemistry content and quantitative procedures. The scoring form asks raters to evaluate a student's procedures, observations, and conclusions, and for the last two tasks, the student's quantitative steps, explanations and conclusions. Interrater reliability is .97.

Table 1 presents the assessments selected and their classification based on the three frameworks. Once the assessments were selected, we examined their appropriateness for this study by administering them in a pilot study.

Table 1

Performance Assessment Characteristics Based on the Three Frameworks

Performance assessment	Reasoning dimension	Content	Process	Task type
Electric Mysteries	Basic knowledge and reasoning	Rich	Open	Component-identification investigation
Daytime Astronomy	Spatial-mechanical	Rich	Open	Observation investigation
Aquacraft	Quantitative science	Rich	Open	Comparative investigation

Pilot Study

We (Ayala et al., 2001) conducted a pilot study with three teachers ("experts") and three students ("novices") to see whether the three performance assessments did indeed tap the different reasoning dimensions. Each of the performance assessments was administered individually to one expert (a science teacher) and one novice (a high school physics student). Prior research on expertise (e.g., Chi, Glaser, & Farr, 1988) suggested that if the performance-assessment task environment ("nominal task") had an effect on reasoning, then using this extreme group design would allow us to detect the effect. Although every person constructs a somewhat different problem space when confronted with the same nominal task, experts are consistent in their substantive representations of the principle underlying the task, whereas novices are strongly influenced by the specified task features. Hence, a large sample was unnecessary to detect the effect. Of course, the next step in this research would be to confirm systematic effects, if found, with multiple experts and novices—something that we begin to do here.

Expert volunteers were assigned to the performance assessment that most closely matched their teaching expertise. A female chemistry teacher with 4 years of teaching experience was assigned Aquacraft, a female physical science teacher with 7 years of teaching experience was assigned Electric Mysteries, and a male general science teacher with 13 years of experience was assigned Daytime Astronomy. Student volunteers were randomly assigned to each of the different tests. All students were male high school physics students who had completed at least two years of high school science. The student assigned to Electric Mysteries was the only student who had not completed a chemistry course.

Students and teachers were asked to think aloud while they completed the performance assessments. Think-alouds were audiotaped and transcribed. Similar procedures have been used before to investigate cognitive task demands of assessments (Baxter & Glaser, 1998; Ericsson & Simon, 1993; Ruiz-Primo, 1999).

In the pilot study, we developed contemporaneously a process of segmentation of protocols and an encoding system in a manner similar to studies by Ericsson and Simon (1993). The think-alouds were segmented, and iterations of the encoding categories were tried out on the segments. As part of testing of the training and encoding system, two raters classified random segments of the think-alouds independently. The raters then discussed disagreements in coding and ways to make either the segments more identifiable and/or the encoding categories more

explicit. Then comparisons were made between the types of reasoning elicited from performance across novice and experts.

Differences in reasoning demands were evident in the think-aloud data (Figure 4). First, we found that, averaging over experts and novices, all three assessments drew on basic knowledge and reasoning, but less so for Aquacraft than for the other two assessments, as expected. Second, we found clear evidence of spatial-mechanical reasoning with Daytime Astronomy and quantitative science reasoning with Aquacraft, again as expected. And finally, as expected, Electric Mysteries drew heavily on basic knowledge and reasoning.

These data then supported our initial conjecture that Electric Mysteries tapped basic knowledge and reasoning and Aquacraft tapped quantitative science reasoning. However, Daytime Astronomy was not "pure" and elicited spatial-mechanical reasoning and more basic knowledge and reasoning than expected. Aquacraft also was not pure because it tapped into basic knowledge and reasoning as well.

The main purpose of the pilot study was to ascertain whether there were reasoning differences among performance assessments selected to vary in demands on basic knowledge and reasoning, quantitative science reasoning, and spatial-

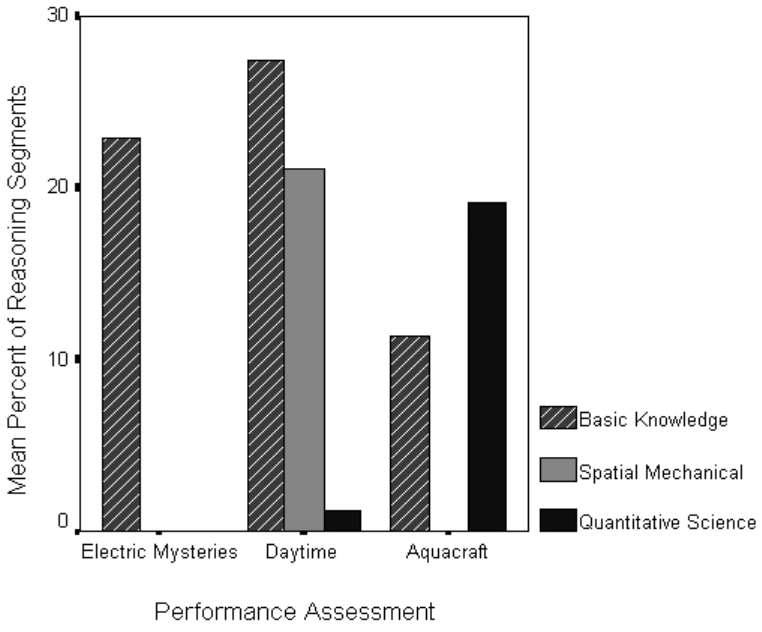

Figure 4. Reasoning demands by performance assessment.

mechanical reasoning. By selecting performance assessments using the general characteristics of the reasoning dimensions, and then collecting think-aloud protocols to study the reasoning these tasks evoked, we found that the different performance assessments did indeed elicit different reasoning patterns albeit for the experts and novices in the pilot study.

Methods

Respondents

In the summer of 2000, following the main study (see Lau, Roeser, & Kupermintz, 2002), a subsample of 35 students completed the three performance assessments while thinking aloud. To find these 35 students, we invited all students who had completed all the other assessments and for whom we had summer contact information (n = 225) to take the performance assessments.

Because many of the 225 students were from the higher academic track, we focused our recruitment on lower performing students. Care was taken also to select some students who had not yet completed chemistry and to select a sample that included girls. It was not easy to get students to take 2 hours of exams during the summer, and many students had to be re-invited before they came. Jokingly, one student remarked, “The hardest test was coming here.” But with all that, students from a variety of grade levels and achievement levels completed the performance assessments (Table 2).

Instrumentation

Three hundred forty-three students completed a series of motivational and cognitive tests and surveys as well as the science achievement measures. Our focus here was on achievement. To measure science achievement, these students took a 30-item multiple-choice test based on items from NELS:88, NAEP, and TIMSS and

Table 2

Characteristics of Performance Assessment Respondents

	Gender		
Chemistry course completion	F	M	Total
Pre chemistry	3	4	7
Post chemistry	12	16	28
Total	15	20	35

an 8-item constructed response test based on TIMSS items, and a subsample of 35 students also took the three different performance assessments described above in the review of the pilot study.

The mean score for the 343 students who completed the multiple-choice test was 16.17 (SD = 5.65) out of 30, whereas the mean score on the constructed responses test was 2.75 (SD = 1.52) out of 8. These mean scores were surprisingly low considering that most of these students had already completed 3 years of high school science. As expected, the correlation between student scores on the multiple-choice and constructed response tests was moderately positive (r = .69, p = .000), indicating that both tests measured somewhat similar, but not identical, aspects of science achievement. Compared with the full sample, our performance assessment subsample (n = 35) had a similar correlation between their multiple-choice scores and their constructed response scores (r = .70, p = .000) and a slightly higher mean score on their multiple-choice test, 18.35 (SD = 4.54) and their constructed response items 3.01 (SD = 1.30). We concluded that the subsample was similar to the students in the main study on these measures.

Performance Assessment Scoring and Reliability

Electric Mysteries. Two trained raters scored each of the Electric Mysteries performance assessments. The original scoring form asked raters to evaluate students based on their drawings of the circuit they used to investigate the contents of the Electric Mystery Box and whether or not the student correctly identified the box's contents. In order to match the other two performance assessments' scoring forms, a more elaborate scoring form was used (Rosenquist, Shavelson, & Ruiz-Primo, 2000). This new scoring form as well as evaluating student drawings and inferences also asked raters to evaluate student observations and explanations used to determine the boxes' contents. Furthermore, our students were not very good at following instructions. Many times our students did not draw the circuit they used for their investigations but their explanations revealed their knowledge about circuits. We used the new scoring form because it allowed us to credit these explanations. The relationship between the original form and new scoring form, excluding those students who did not draw the circuit, was strong (r = .91).

The two raters scored each of the Electric Mysteries performance assessments using the new scoring form. Sixteen of the 35 student notebooks were used for training purposes, and 19 were used for a reliability study. We used more notebooks for training in the Electric Mysteries assessment than were used in training for the

other performance assessments (10-13 notebooks) because of the difficulty in distinguishing scientific explanations from logical reasoning. All students reasoned to identify the contents of the boxes, and we credited students who explained their reasoning with electricity concepts (e.g., current, resistance, circuit). The final interrater reliability for the new scoring form with explanations was high (r = .95). From here forward, we report only Electric Mysteries scores using the new scoring form (with explanations).

Daytime Astronomy. Two trained raters scored the Daytime Astronomy performance assessments. A random sample of 13 of the 35 student notebooks was used for training purposes. Once raters had scored 13 notebooks and had discussed problematic areas, the remaining 22 were scored independently (interrater reliability = .90).

Aquacraft. Two trained raters scored each of the Aquacraft performance assessments. A random sample of 10 of the 35 performance assessments was used for training purposes. Once the raters had reviewed the performance assessments used for training and discussed problematic areas, the remaining 25 assessments were scored independently (interrater reliability = .97).

Results and Discussion

Here we report the results of analyses focused on the consistency of scores across the three types of tests—multiple-choice, constructed response, and performance—and on the extent to which scores on these three measurement methods converged on the three reasoning types—basic knowledge and reasoning, quantitative reasoning, and spatial-mechanical reasoning. Since the multiple-choice items have been the focus of other reports on our study, we begin here by examining the constructed response scores then move to the performance assessment scores. Then we compare all three measures (multiple-choice, constructed response and performance assessments). Finally, we look for the reasoning dimensions in the different measures.

Constructed Response Scores

We examined the constructed response data to see how students responded to individual items (Table 3). We found that students on average did better on some items than on others. As expected, students did better on items drawn from

Table 3

Constructed Response Mean Scores

Constructed response item	Grade level	Complete sample (n = 343) %	Performance assessment subsample (n = 34) %
Pencil Reflection-SM	7–8	79.0	85.3
Tractor Efficiency-QS	7–8	68.0	76.5
Watering Can-SM	7–8	64.8	67.6
Broken Window-BKR	12	8.3	8.8
Electrical Energy-BKR	12	19.8	17.6
Bacteria Growth-QS	12	26.3	35.3
Ice in Aquarium-BKR	12	3.6	2.9
Doppler Effect-BKR	12	4.4	2.9

Note. SM = spatial-mechanical reasoning; QS = quantitative science reasoning; BKR = basic knowledge and reasoning.

the 8th-grade TIMSS than on the 12th-grade TIMSS items. More than 65% of the students answered the 8th-grade items correctly, and less than 35% of the students answered the 12th-grade items correctly. For example, 68% of the students answered Tractor Efficiency (8th grade) correctly, whereas only 26% of students answered Bacteria Growth (12th grade) correctly, even though both of these questions were classified as belonging to the quantitative science reasoning type.

Student also performed very poorly on Doppler Effect (12th grade). This item asked students to explain how the frequency of a blowing car horn changes as the car approaches and passes you. On scoring Doppler Effect, we found that most students held the misconception that as the car approaches, the frequency of the sound continually increases until the car passes you and then continually decreases (like amplitude/volume). It does not.

We also noted our students' lack of reliance on formulas to represent and solve problems. For example on Tractor Efficiency students rarely used a formula for efficiency (hectares per liter), and in Broken Window students did not use Pressure = Force/Area or a similar equation to solve the problem. Rather they relied on general reasoning to solve the problem. In Tractor Efficiency, students were asked to decide which of two tractors is more efficient. One tractor uses [X] liter of gas to complete 2 hectares while the other tractor uses [Y] liter for 1 hectare. To solve this problem, most of our students would equate tractors on the hectares cleared (multiplying second tractor by two) and then compare rather than calculating

efficiencies. In Broken Window, students were asked to explain why a windowpane does not break when a tennis ball strikes it and why it does break when hit by a rock, with both projectiles having equal mass. To solve this problem, our students listed multiple reasons why the tennis ball does not break the window and the rock does (i.e., the ball is soft and bouncy while the rock is hard, or the ball has give and the rock does not). It would appear that our students did not rely on formulas as representations to help them solve problems (Perkins & Unger, 1994).

This lack of formula use might help explain why our students did so poorly on Bacteria Growth. In this problem, students were expected to determine the number of bacteria in a colony given populations at two different times. In order to solve the problem, students should apply an exponential formula or recognize and then extrapolate from the existing numeric sequence. Most students used the second method and made errors in extrapolating the sequence. Again, the students did not rely on formulas to solve problems.

Performance Assessment Scores

Electric Mysteries. The Electric Mysteries mean score for students was 28.1 out of a maximum of 48 points (Table 4). We disaggregated the scores into the four scoring categories to locate the source of students' errors. The new scoring form adjusts for the lack of drawings and credits explanations; however, students did not provide good explanations in their notebooks. The low total score was largely due to the low mean explanation score, 2.39 out of 12 possible points. The drawing, observation and inference subscores were much higher than the explanation subscores.

Table 4

Electric Mysteries Total Score and Subscores Using New Scoring Form

		Electric Mysteries subscores			
	Total	Drawing ($n_i = 6$)	Observation ($n_i = 6$)	Inference ($n_i = 6$)	Explanation ($n_i = 6$)
Mean	28.10	8.55	7.77	9.49	2.39
Standard deviation	10.25	3.84	3.14	3.10	2.29

Daytime Astronomy. The Daytime Astronomy mean score was 27.1 (*SD* = 8.75; Table 5). For comparison, Shavelson et al. (1998) found that the Daytime Astronomy mean score with fifth graders was 14, and Ayala et al.'s (2001) Daytime Astronomy novice scored 34 and the expert scored 60.

Since the Daytime Astronomy scoring form contained three scoring categories, the data were disaggregated. These scores reflected the accuracy of the students' observations, the quality and type of modeling that the students used in answering the questions, and the quality of their explanations.

Aquacraft. The Aquacraft mean score was 15.7 (Table 6). For comparison, in our previous study, the Aquacraft novice scored 17 and the Aquacraft expert scored 32 out of a possible 42 (Ayala et al., 2001). One student (an outlier) scored very high, 35. This was the only student to solve all the problems, yet he did not get a perfect score because he failed to make all the observations to compare the salt and fresh water conditions.

Table 5

Daytime Astronomy Total Score and Subscores

		Daytime Astronomy subscores		
	Total	Results (n_i = 6)	Modeling (n_i = 6)	Explanation (n_i = 6)
Mean	27.10	14.45	5.36	7.29
Standard deviation	8.76	5.35	2.32	3.50

Table 6

Aquacraft Total Score and Subscores

		Aquacraft subscores			
	Total	Task 1 Chemical reaction (n_i = 12)	Task 2 Testing the gas (n_i = 7)	Task 3 Balancing equations (n_i = 7)	Task 4 Energy calculations (n_i = 10)
Mean	15.70	5.27	2.07	6.16	2.20
Standard deviation	6.97	2.43	1.34	2.78	2.53

Further analysis of Aquacraft revealed that students tended to do better on Task 3–Balancing Equations, (6.16 out of 8 maximum) than on Task 4–Energy Calculations (2.20 out of 10 maximum). Both of these tasks required students to use multiple calculations using content knowledge from first-year chemistry. Students clearly knew the process of balancing equations, and many were able to identify elements from memory (i.e., they identified Ba, naming it Barium). However, students were not able to do the energy calculations. Upon reviewing the notebooks, it appeared that they were not able to convert kilograms to moles; that is, kg x 1000 gr / kg x (1 / atomic weight) moles / gr = moles. Though reasoning dimensions might be useful for explaining the types of processes used by students to solve problems, the declarative and procedural knowledge that these students had also played into the mix, and in the case of Task 4–Energy Calculations this lack of procedural (algorithmic) knowledge was limiting. This corresponds to our students' overall lack of formula use to solve science problems as evidenced in the constructed response items.

Performance Assessment Comparison

We postulated that the three performance assessments tapped into procedural knowledge and into different declarative content knowledge. We also postulated that the three performance assessments elicited different reasoning demands from students. We expected from our earlier think-aloud data that Electric Mysteries would be more closely related to Daytime Astronomy than to Aquacraft, that Daytime Astronomy would be more closely related to Electric Mysteries than to Aquacraft, and that Aquacraft would be equally unrelated to Electric Mysteries and Daytime Astronomy (see Figure 4). Table 7 provides correlations for the scores on the three performance assessments, and the correlational pattern suggests just the opposite. The relationship of Aquacraft with Electric Mysteries and of Aquacraft with Daytime Astronomy was stronger than the relationship between Electric Mysteries and Daytime Astronomy.

Investigating the scatter plots, we identified some students who seemed to do much better on one assessment than another. We reviewed their performances and decided whether they should be removed from the comparison analysis—might there be a reason beyond the tasks that somehow affected their scores, like failing to complete test or skipping a section? Removing three students did raise the overall correlations between the three assessments, but the overall pattern remained the same.

Table 7

Correlations Among Scores on Three Performance Assessments

	Electric Mysteries	Daytime Astronomy	Aquacraft
Electric Mysteries	(.95)[a]		
Daytime Astronomy	.19 .20[b]	(.90)[a]	
Aquacraft	.35* .41*[b]	.34* .38*[b]	(.97)[a]

[a] Interrater reliability.

[b] Correlations exclude one student who skipped tasks 1 and 2 in Aquacraft, although this student scored well on tasks 3 and 4.

* Correlation is significant at the 0.05 level (2-tailed).

Consistency of Performance Across Science Achievement Tests

One purpose of this study was to examine the consistency of students' scores across the three different types of science achievement tests. How do these measures compare? How do students perform across the three measures? Because each performance assessment measured procedural knowledge as well as different declarative content knowledge (e.g., Electric Mysteries measured electric circuits and Aquacraft measured chemistry), combining the performance assessment scores would lead to low total performance-assessment internal consistency[2] and consequently did not make sense. Therefore, each performance assessment was compared individually with each of the other measures.

We found, first, that the constructed response score reliability (.42) was too low to permit interpretation of this scale's correlation with the other tests (Table 8). The correlations between multiple-choice and performance assessment scores were positive and of moderate magnitude. We interpreted this as demonstrating consistency in performance. Moreover, these correlations, and their disattenuated counterparts (above the main diagonal in Table 8) indicated that these tests did tap into the science achievement domain, but into somewhat different aspects, as expected. Finally, the multiple-choice scores correlated higher with each of the

[2] The internal consistency of a total performance assessment score with all three measures was low (.56). This low reliability arose because these performance assessments, although they measured some overlapping scientific process skills, individually measured something different (see correlations in Table 8), and three "items" only do not provide a consistent picture.

Table 8

Science Achievement Total Sore Correlations (Observed Correlations Below Main Diagonal and Disattenuated Correlations Above)

	Multiple choice	Constructed response	Electric Mysteries	Daytime Astronomy	Aquacraft
Multiple choice	(.75)[a]	*b*	*.44*	*.83*	*.61*
Constructed response	.65**	(.42)[a]	*.22*	*.65*	*.35*
Electric Mysteries	.38**	.14	(.95)[a]	*.20*	*.36*
Daytime Astronomy	.68**	.40*	.19	(.90)[a]	*.34*
Aquacraft	.52**	.22	.35*	.38*	(.97)[a]

[a] Internal consistency reliability for multiple-choice and constructed response scores; interrater reliability for performance assessment scores.

[b] Exceeded 1.00 due to the low reliability of the constructed response measure.

* .05 level. ** .01 level.

performance scores than the performance scores did with one another. We interpreted this to reflect the broad content coverage of the multiple-choice test compared with the more focused coverage of the performance assessments.

Convergence of the Different Types of Tests on the Reasoning Dimensions

The second purpose of this study was to examine the extent to which our three different achievement test subscores converged on the reasoning dimensions they were intended to measure. We nominally assigned each multiple-choice item to one of the three reasoning dimensions (see Shavelson et al., 2002). The patterns of correlations among the multiple-choice score for each reasoning dimension and for each performance assessment score are shown in Table 9. (In this analysis, we did not use the constructed response scores because of their low internal consistency when disaggregated by reasoning dimension.)

The main diagonal in Table 9 represents the reliabilities (in parentheses) of the corresponding measures. Because the reliabilities for the multiple-choice items were not strong, we corrected the observed correlations and presented the disattenuated correlations above the main diagonal (in italics). It is important to notice that none of these disattenuated correlations is greater than 1, which would have indicated that the results are questionable due to the low reliability of the measures.

Table 9

Multireasoning-Multitest Correlation Matrix (Observed Correlations Below and Disattenuated Correlations Above Main Diagonal)

		Multiple-choice items			Performance assessment		
		Basic knowledge (BKR)	Spatial-mechanical (SM)	Quantitative science (QS)	Electric Mysteries (BKR)	Daytime Astronomy (SM)	Aquacraft (QS)
Multiple-choice items							
Basic Knowledge	(BKR)	(.75)[a]	*.82*	*.85*	*.60*	*.72*	*.54*
Spatial-mechanical	(SM)	.53**	(.56)[a]	*.76*	*.16*	*.48*	*.42*
Quantitative science	(QS)	.65**	.50**	(.78)[a]	*.23*	*.76*	*.53*
Performance assessments							
Electric Mysteries	(BKR)	.51**	.12	.20	(.95)[b]	*.21*	*.36*
Daytime Astronomy	(SM)	.59**	.34*	.64**	.19	(.90)[b]	*.41*
Aquacraft	(QS)	.46**	.31	.46**	.35	.38*	(.97)[b]

[a] Internal consistency reliability for the multiple-choice test subscores (n = 371).

[b] Interrater reliabilities for the performance assessments.

* .05 level. ** .01 level.

Next we looked at the convergent validity diagonal (below the main diagonal and underlined in Table 9) where the correlations between observed scores on one trait as measured by multiple methods were found. In this case, this diagonal represents the correlations between the multiple-choice subscores and the performance assessment that corresponded to one reasoning dimension (e.g., Electric Mysteries, believed to tap basic knowledge and reasoning, correlated with the subscores of the basic knowledge and reasoning multiple-choice items). If the multiple-choice subscores and the performance assessment scores on these reasoning dimensions converge, the correlations in the convergent validity diagonal should be higher than the correlations between other performance assessments and other multiple-choice subscores (i.e., the correlation between Electric Mysteries scores and basic knowledge and reasoning multiple-choice scores should be higher than the correlation between Electric Mysteries and spatial-mechanical reasoning scores or the correlation between Daytime Astronomy and the spatial-mechanical multiple-choice scores).

The empirical evidence did not support our conjecture about the convergence of multiple-choice and performance assessment scores. The validity correlations were hardly larger than the other correlations, and in some cases the validity

diagonal correlations were lower than other correlations between performance assessments and multiple-choice subscores. Either the multiple-choice items or the performance assessments, or both, do not reflect the reasoning dimensions.

In further analyses we noticed that Electric Mysteries scores were most related to the basic knowledge and reasoning multiple-choice subscores (r_{EMbkr} = .51, p = .002) rather than the spatial-mechanical (r_{EMsm} = .12, p = .49) or quantitative science multiple-choice subscores (r_{EMqs} = .20, p = .257). These results corresponded to our nominal analysis placing both the multiple-choice items and the performance assessments into the reasoning dimensions. Moreover, the scores from Daytime Astronomy, our spatial-mechanical performance assessment, were expected to be related only to the spatial-mechanical multiple-choice scores and the basic knowledge and reasoning scores. And, indeed, these scores were so related (r_{DAsm} = .34, p = .05, and r_{DAbkr} = .59, p = .00). However, to our surprise, Daytime Astronomy was also related to the quantitative science multiple-choice subscores (r_{DAqs} = .64, p = .05). Finally, the scores for Aquacraft, our quantitative science performance assessment, were expect to be related to the quantitative science multiple-choice subscore and the basic knowledge and reasoning subscore, and indeed they were (r_{AQqs} = .46, p = .006, and r_{AQbkr} = .46, p = .005, respectively).

Overall, Electric Mysteries scores matched our prediction for their relationship with the reasoning dimensions as defined by the multiple-choice scores being highest in basic knowledge and reasoning and lower in the other two reasoning dimensions. However, Daytime Astronomy and Aquacraft did not match our predictions. We expected Daytime Astronomy to have a strong relationship with basic knowledge and reasoning, to be the only performance assessment related to spatial-mechanical reasoning, and to have the least relationship with quantitative science. This was not the case. Daytime Astronomy scores correlated highest with the quantitative science scores, then with the basic knowledge and reasoning scores, and finally, the correlation was lowest with the spatial-mechanical scores. We also expected that Aquacraft scores would have less relationship with basic knowledge and reasoning, no spatial-mechanical relationship, and be most related to quantitative science. This, too, was not quite the case. Why?

From our observations of the students completing Daytime Astronomy and from reading their notebooks, we found that students were indeed involved in spatial-mechanical activities—shining the flashlight from above, observing the change of shadow length and angle as the globe is rotated. But we found that

students used other content knowledge to perform the assessment, as well as using time zones, geometry, knowledge of meridians, longitude and latitude, and personal travel experience. For example, one student used time zones to place the tower in the correct location: "I know that the time [difference] between the Midwest and Seattle is about 2 hours . . . and so aim tower [shadow] to make sure it was 2 hours behind instead of ahead." This makes the assignment of Daytime Astronomy to any combination of reasoning dimensions questionable because the reasoning used by the student depends on the content knowledge and life history he or she brings to the task.

From our observations of students completing Aquacraft and from their notebooks, we found that students consistently performed well on Task 3–Balancing Equations and poorly on Task 4–Energy Calculations. It may be that students' experience and knowledge in Task 3–Balancing Equations move it from a quantitative science task into a basic knowledge and reasoning task. As for Task 4, it may be that students' lack of experience and knowledge in performing quantitative energy calculations influenced their reasoning. That is, they may have been grasping for straws to answer Task 4–Energy Calculations and using basic knowledge and reasoning rather than the quantitative science reasoning we expected. For example, one student became bogged down in simple metric conversions: "200 kilograms is 2 grams, 20 grams? I don't remember. Oh. I'm thinking it's 2 grams. Let me think. No, it's not. No, it's 2000 grams. Yeah, that's what it is. So 200 kilograms is 2000 grams, I think." We will address these issues in our future work with the think-aloud data.

Conclusions

We set out to determine whether performance assessments could be explicitly designed to tap the three reasoning dimensions found by Snow and colleagues in their analysis of NELS:88 science items; to validate interpretations of science achievement test scores as reflecting these dimensions, including scores on performance tests; and to examine the consistency of student performance across the three achievement measures. We asked, "Can this multidimensionality be found in other types of assessments such as constructed response and performance assessments?" "Can we nominally assign items to these reasoning dimensions?" To answer these questions, we selected 30 multiple-choice items, 6 constructed response items from NELS, NAEP and TIMSS, and 3 performance assessments. We then assigned these items and assessments based on their task characteristics to the three reasoning dimensions: basic knowledge and reasoning, spatial-mechanical

reasoning and quantitative science reasoning. We found the constructed response scores too unreliable to use in our analyses. Consequently, we compared performance assessment scores with the multiple-choice scores and found that not all the multiple-choice and performance assessment scores converged on the reasoning dimensions.

Our nominal assignment of the performance assessments did not always correspond with the empirical findings. Perhaps we should not be surprised. First the good news: The venerable Electric Mysteries behaved as expected, requiring more basic knowledge and reasoning than quantitative science and spatial-mechanical reasoning. Now to reality: Our spatial-mechanical performance assessment, Daytime Astronomy, was correlated with other measures across all the reasoning dimensions but especially with the quantitative science dimension. Our quantitative science performance assessment, Aquacraft, tapped into basic knowledge and reasoning as well as quantitative science reasoning. These results suggest that students reason through the problems in different ways depending on the knowledge that they have (or do not have) about the task at hand. If you know about meridians, you answer Daytime Astronomy using meridians, or if you have traveled to Missouri, you use your travel experience—different reasoning based on you knowledge and experience.

Additionally, when we explicitly developed a quantitative science performance assessment, we found that it tapped into not only quantitative science reasoning but also basic knowledge and reasoning. It may be that the individual tasks in the assessment require quantitative science reasoning, and the actual completion of such tasks as preparing test tubes, making comparisons, observing, reporting findings and concluding may be more closely aligned with basic knowledge and reasoning. Or, it may be that if there are other (easier) ways to complete the tasks, then students use the simpler way. Or, if students do not have the necessary content knowledge to complete s task, then they revert to other ways of solving the task that do not match a nominal analysis of the task. For example, the constructed response item Tractor Efficiency asked students to compare the efficiency of two tractors. We expected students to use quantitative science reasoning to solve this problem, that is, comparing the tractors' efficiency (liters / hectare). When asked to choose the most efficient machine, most students got the right answer. However, in their explanations, we found few students who used the quantitative science reasoning strategies we expected. Most students equated the machine on hectares and then

compared their gasoline usage—a much simpler task. Because of the diverse strategies students may use to solve the same problem, it is problematic to nominally assign tasks into the three reasoning dimensions based on how we, the researchers, solve the tasks.

We believe that the nominal assignment of performance assessments on these three reasoning dimensions is problematic. Though a multiple-choice item may fall neatly into a reasoning dimension, because of the complex nature of performance assessments—the interaction with task and the openness of the responses—these assessments may tap into a variety of reasoning dimensions, especially basic knowledge and reasoning, that all students used to solve problems. Student knowledge and experience seem to suggest how a student solves a problem, not the problem alone.

References

Ayala, C. C., & Shavelson, R. J. & Ayala, M. A. (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.

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.

Baxter, G., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 36-45.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on assessment. *Journal of Educational Measurement, 29*, 1-17.

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

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

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

Hamilton, L. S., & Snow, R. E. (1998). *Exploring differential item functioning on science achievement tests* (CSE Tech. Rep. No. 483). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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.

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.

Perkins, D. N., & Unger, C. (1994). A new look in representations for mathematics and science learning. *Instructional Science, 22*(1), 1-37.

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, M. A. (2000). *On the "exchangability" of hands-on and computer simulated science performance assessments* (CSE Tech. Rep. No. 531). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Ruiz-Primo, M. A. (1999, April). *On the validity of cognitive interpretations of scores from alternative concept-mapping techniques.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Shavelson, R. J., Baxter, G., & 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.

Shavelson, R. J., & Ruiz-Primo, M. A. (1999). On the assessment of science achievement. (English version). *Unterrichts Wissenschaft, 27,* 102-127.

Shavelson, R. J., Solano-Flores, G., & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. *Evaluation and Program Planning, 2,* 171-184.

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

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

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.

APPENDIX

Performance Assessment and Reasoning Dimensions (Ruiz-Primo, 1999)

#	Performance assessment	Task, response and scoring	Classification system	Type of reasoning
1	Daytime Astronomy	Student determines where to place towers on a globe based on the size and direction of their shadows. Students describe the relationship between time and sun location. Scoring is based on observations and modeling.	Observation	SM
2	Electric Mysteries	Student determines what is inside an electric mystery box by constructing and reasoning about circuits. Scoring is evidence based, focusing on evidence and explanation.	Component identification	BKR
3	Friction	Student determines the amount of force needed to drag an object across surfaces of varying roughness. Scoring is procedure based, focusing on how student designs experiments.	Comparative investigation	BKR
4	Paper Towels	Student finds which paper towel absorbs the greatest amount of water. Scoring is procedure based, focusing on the investigation's design.	Comparative investigation	BKR
5	Bottles	Student identifies what makes bottles of different mass and volume sink and float. Scoring focuses on the characteristics and quality of the categorization.	Classification	BKR
6	Bugs	Student determines sow bugs' preferences for light or dark and moist or dry environments. Scoring is procedure based, focusing on the investigation's design.	Comparative investigation	BKR
7	Electric Motors	Student identifies which direction a battery is facing within a mystery box. Scoring is evidence based, focusing on evidence and explanations.	Component identification	BKR
8	Batteries	Student determines which batteries are good or not. Scoring is evidence based, focusing on evidence and explanations.	Component identification	BKR
9	Magnets	Student identifies which magnet is stronger. Scoring is evidence based, focusing on evidence and explanations.	Component identification	BKR
10	Pulse	Student determines how her pulse changes when she climbs up or down a step. Scoring form is based on the observations and modeling.	Observation	BKR
11	Plasticine	Student weighs different amounts of plasticine as carefully as possible. Scoring is evidence based, focusing on evidence and explanations.	Comparative investigation	BKR

Note. SM = spatial-mechanical reasoning; BKR = basic knowledge and reasoning.

(continued)

#	Performance assessment	Task, response and scoring	Classification system	Type of reasoning
12	Shadow	Student finds out the change in size of a shadow made by a card placed between a light and a screen as the card is moved. Scoring form is based on the modeling and explanation.	Observation	SM
13	Solutions	Student determines the effect of temperature on speed of dissolving. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR
14	Rubber Bands	Student determines the length of a rubber band as more and more weight is added. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR
15	Inclined Plane	Student determines the relationship between the angle of inclination and the amount of force needed to move an object up the plane. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR
16	Mystery Powders	Student identifies the components in a mystery powder. Scoring is evidence based, focusing on evidence and explanation.	Component identification	BKR
17	Mystery Powders –6	Student determines the substance contained in each of six bags. Scoring is evidence based, focusing on evidence and explanation.	Component identification	BKR
18	Rocks and Charts	Student identifies the properties of rocks and creates a classification scheme. Scoring focuses on the characteristics and quality of the categorization.	Classification	BKR
19	Saturated Solutions	Student compares the solubility of three powders in water. Scoring is procedure based, focusing design of experiment.	Comparative investigation	BKR
20	Pendulum	Student determines what influences the number of swings of a pendulum. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR
21	Alien	Student determines the acidity of “alien blood” and proposes a remedy. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR
22	Animals	Student creates a two-way classification system. Scoring focuses on the characteristics and quality of the categorization.	Classification	BKR
23	Animals CLAS	Student determines the possible causes of a fish decline. Scoring is evidence based, focusing on evidence and explanation.	Component identification	BKR
24	Chef	Student determines which of three unknowns will neutralize a fourth unknown. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR
25	Critters CLAS	Student classifies 12 rubber insects. Scoring focuses on the characteristics and quality of the categorization.	Classification	BKR
26	Erosion CLAS	Student compares the eroding effects of different solutions on limestone. Scoring is procedure based, focusing on design of experiment.	Comparative investigation	BKR

Note. SM = spatial-mechanical reasoning; BKR = basic knowledge and reasoning.