On the Cognitive Interpretation of Performance Assessment Scores

CSE Technical Report 546

Carlos Cuauhtémoc Ayala and Richard Shavelson
CRESST / Stanford University

Mary Ann Ayala
Palo Alto (CA) Unified School District

July 2001

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

Project 3.1 Construct Validity: Understanding Cognitive Processes—Psychometric & Cognitive Modeling, Richard Shavelson, Project Director, CRESST/Stanford University

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

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

ON THE COGNITIVE INTERPRETATION OF PERFORMANCE ASSESSMENT SCORES

Carlos Cuauhtémoc Ayala and Richard Shavelson[1]
CRESST/Stanford University

Mary Ann Ayala
Palo Alto (CA) Unified School District

Abstract

We investigated some aspects of reasoning needed to complete science performance assessments, i.e., students' hands-on investigations scored for the scientific justifiability of the findings. While others have characterized the content demands of science performance assessments as rich or lean, and the processes as constrained or open, or characterized task demands as calling for different cognitive processes, we studied the reasoning demands of science performance assessments on three dimensions based on previous analysis of NELS:88 data: basic knowledge and reasoning, spatial mechanical reasoning, and quantitative science reasoning. In this pilot study, 6 subjects (3 experts and 3 novices) were asked to think aloud (talk aloud) while they completed one of three science performance assessments. The performance assessments were chosen because their tasks appeared to elicit differences in reasoning along these three dimensions. Comparisons were then made across the performance assessments and across the expertise levels. The talk alouds provided evidence of the three reasoning dimensions consistent with our nominal analysis of the performance assessment tasks. Furthermore, experts were more likely to use scientific reasoning in addressing the tasks, while novices verbalized more "doing something" and "monitoring" statements.

If you wanted to know whether a child could tie her shoe, you would probably ask her to show you rather than use a set of multiple-choice items, and your assessment would be based on performance of a criterion task. Although multiple-choice tests are useful for ascertaining a child's conceptual knowledge, an assessment of actual performance maybe more appropriate in some situations. Science education assessment may be one of these situations. Science performance assessments pose a problem and put students in a mini-laboratory to solve it,

[1] We wish to thank Maria Araceli Ruiz Primo, Min Li, Tamara Danoyan, and Angela Haydel for their contributions to this work.

evaluating the solution as to its scientific defensibility (Shavelson, Baxter, & Pine, 1991). These assessments have captured the attention of researchers and policymakers for the last 10 years (Messick, 1994; Ruiz-Primo & Shavelson, 1996). Performance assessments are interpreted as capturing a student's scientific reasoning and procedural skills (California State Board of Education, 1990) and are believed to require the application of scientific knowledge and reasoning in simulated real-world situations as well as in situations similar to what scientists do (National Research Council, 1996). With this study, we further tease out the reasoning needed to complete science performance assessments, testing the validity of these cognitive (reasoning) claims.

Other current research on performance assessment has focused on linking reasoning demands to the characteristics of the tasks in efforts both to assist developers with assessment construction and to validate cognitive interpretations. Baxter and Glaser (1998) for example asked students to talk aloud while doing performance assessments. By observing student performance and analyzing scoring systems, they characterized performance assessments along two continua. The first continuum reflected the task's demands for content knowledge ranging from rich to lean. The second continuum represented the task's demand for process skills ranging from constrained to open. These two continua revealed the "Content-Process Space" of assessment tasks. By analyzing the reasoning and content demands, performance assessments can be located in the Content-Process Space. This framework proved useful because "tasks can be designed with specific cognitive goals in mind, and task quality can be judged in terms of an alignment with the goals and purposes of the developers" (p. 40).

Other research on performance assessments proposed a classification system that distinguished among assessment tasks and linked them to their characteristic scoring systems (Ruiz-Primo & Shavelson, 1996; Shavelson, Solano-Flores, & Ruiz-Primo, 1998; Solano-Flores, Jovanovic, & Shavelson, 1994). Four task types were proposed: comparative investigations, component-identification investigations, classification investigations, and observation investigations (Shavelson et al., 1998). Comparative investigations require students to compare two or more objects on some attribute while controlling other variables, and successful performance is based on the scientific justifiability of the procedures used and the accuracy of the problem solution. Component-identification investigations require students to determine the components that make up a whole; successful performance is based

on the evidence identifying one component and disconfirming the presence of another component as well as the procedures used to collect the evidence. Classification investigations require students to create a classification system for a set of objects based on their characteristics to serve a practical or conceptual purpose; here, successful performance is based whether the student uses characteristics of the objects relevant to the purpose. The observation investigations require students to perform observations on a phenomenon using a model that determines how those data are gathered, and then ask students to describe the results obtained. This classification system proves useful because once a type of assessment "is decided upon, a lot is known about the structure of the task and the nature of the scoring system" (Shavelson et al., 1998, p. 174).

Baxter and Glaser (1998) have developed the Content-Process Space based on the depth of content knowledge elicited and the structure of procedures, and Shavelson et al. (1998) have developed a classification system based on the demands of the assessment tasks and their corresponding scoring systems. In this paper, we propose an additional attribute of performance assessments, "reasoning." This attribute is based on the level and kind of reasoning required to conduct the task at hand. We asked, "What are the reasoning demands needed in order to complete different performance assessments?" and "Can we locate a performance assessment on a particular reasoning dimension based on the characteristics of the content and the task?" Clearly, determining the cognitive validity of performance assessments with respect to scientific reasoning they elicit is paramount, and this pilot study continues to lay this foundation (Messick, 1994). And, since these assessments are touted as tapping higher order thinking skills and as mimicking what scientists do, reasoning may after all be the most important dimension.

Context

This study is a piece of larger study conducted, in part, for the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). In the larger study, 500 high school students from the San Francisco Bay Area were assessed on their cognitive abilities, motivation, attitudes towards science, and science achievement using questionnaires, multiple-choice and constructed-response tests, and performance assessments.

As part of this larger study we posited three reasoning dimensions. 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). Factor analysis of the NELS:88 10th-grade science data suggested three reasoning and knowledge dimensions: basic knowledge and reasoning, quantitative science, and spatial mechanical reasoning. Corroborating evidence supporting the three reasoning dimensions came from talk aloud protocols, observations and posttest interviews (Hamilton, Nussbaum, & Snow, 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, can be differentiated from the other reasoning dimensions based on a student's more frequent use of predictions, gestures, and visualization. Table 1 contains descriptions and sample items for these dimensions.

Purpose

As part of the CRESST study a science achievement test was created selecting NELS:88, National Assessment of Educational Progress (NAEP), and Third International Mathematics and Science Study (TIMSS) multiple-choice and constructed-response items balanced on the three dimensions. Correspondingly, we selected a performance assessment based on its nominal characteristics for each of the reasoning dimensions. Whereas previous research on performance assessments focused on the cognitive and process demands needed to complete an assessment, this study explored whether nominal differences among the three performance assessments could be found in the "problem space" constructed by both experts and novices as they completed these assessments, as revealed by their talk aloud protocols.

Performance Assessment Selection

To select performance assessments, we classified a set of them into the three reasoning dimensions. To do this, we examined performance assessment tasks and scoring systems and determined which general characteristics of each dimension most closely matched the overall task of the performance assessment. For example, a "paper towels" investigation asked students to determine which of three paper towels absorbed the most water with a scoring system focusing on procedures. 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

Table 1

Description of Three Reasoning Dimensions (after Hamilton et al. 1995)

Dimension	Example items
Basic knowledge and reasoning	Choose an improvement for an experiment on mice
General characteristics:	Identify the example of a simple reflex
Reflects general knowledge.	Choose the property used to classify substances
Involves greater use of general reasoning.	Select statement about the process of respiration
Requires more verbal reasoning than quantitative science or spatial mechanical.	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	Read a graph depicting the solubility of chemicals
General characteristics:	Read a graph depicting digestion of protein enzyme
Application of advanced concepts; Manipulation of numerical quantities; Requires specialized course-based knowledge.	Infer from results of experiment using filter
	Explain reason for ocean breezes
Item characteristics:	Interpret symbols describing a chemical reaction
Content includes chemistry and physics content; Numeric Calculations.	Calculate a mass given density and dimensions
	Calculate grams of substance given its half life
	Calculate emissions of radioactive decay
	Choose method of increasing chemical reaction
	Predict path of ball dropped in moving train
Spatial mechanical reasoning	Choose a statement about source of moon's light
General characteristics:	Answer question about earth's orbit
Requires reasoning and interpretation of visual or spatial relationships, motions and / or distances.	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

performance assessments were analyzed by this method. Twenty-five assessments were classified as basic knowledge and reasoning, two were classified as spatial mechanical and none were quantitative science (A. Ruiz-Primo, 1999, personal

communication; see Appendix). In order to fill the quantitative science void, two new performance assessments were created using an iterative process where the performance assessment designer presented a team member with iterations of the performance assessment until a version 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 in order to complete the assessment had to come up with their own procedures rather than follow a procedure. And since reasoning demands are related to tasks (Baxter & Glaser, 1998), we selected assessments to represent different task types as defined by Shavelson et al. (1998). 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 et al., 1991). Students were 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 series 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.

We selected "Daytime Astronomy" as our spatial mechanical performance assessment because to solve it required spatial observation, modeling and reasoning (Solano-Flores & Shavelson, 1997; Solano-Flores et al., 1997). These are features of the spatial mechanical reasoning dimension. Students were given an earth globe in a box, a flashlight, and a set of "sticky towers" (see Figure 2). 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. Since the task requires knowledge about the sun's position in relation to earth, and requires knowledge about the relationship between the position of the sun and shadows cast on earth, this task was considered

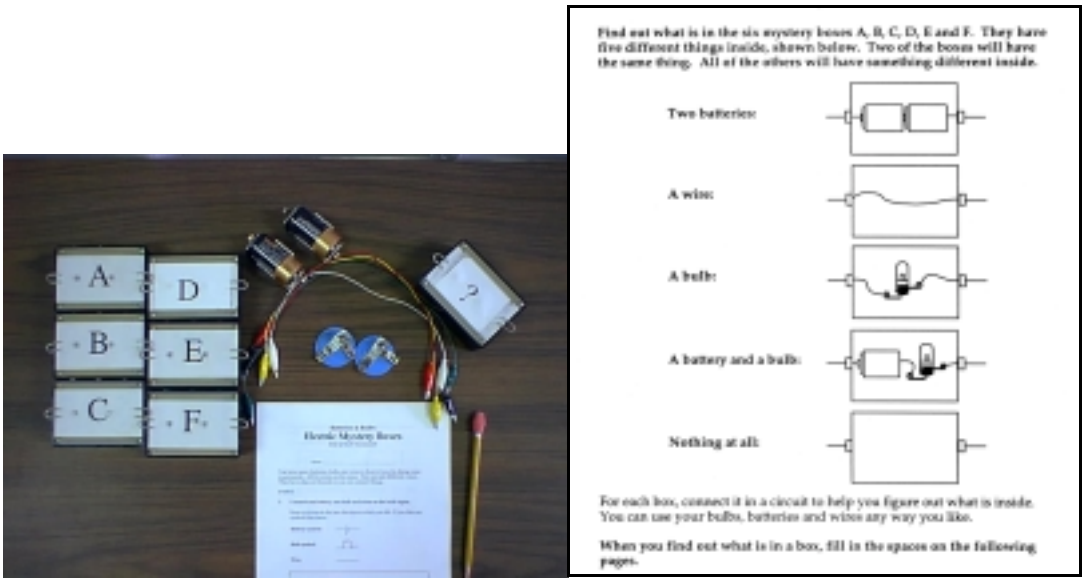

Figure 1. Electric Mysteries performance assessment.

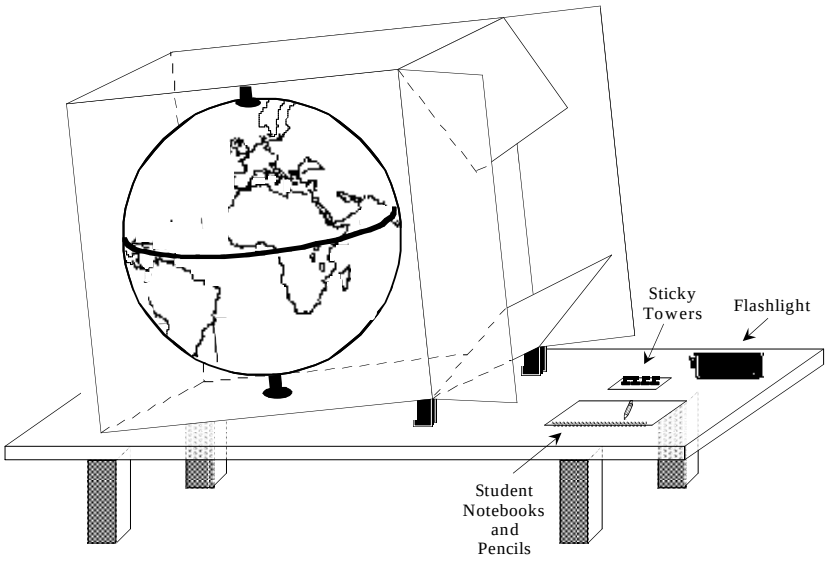

Figure 2. Daytime Astronomy performance assessment.

content rich. Since students were not given directions on how to carry out these tasks, the assessment was considered process open. Since students were asked to model the path of the sun across the sky and to use direction, shadow length, and

shadow angle to solve location problems, Solano-Flores and Shavelson (1997) considered this assessment to be of the observation investigation task type.

We developed a new investigation, "Aquacraft," as a quantitative science performance assessment to match the important components of the chemistry curriculum at our target schools (verified by high school chemistry teachers). Students were asked to determine the cause of an explosion in a submarine by simulating what might have happened in the sub's ballast tanks using glassware, copper sulfate, aluminum, salt and matches (Figure 3).

Students determined the cause of an explosion using high school chemistry principles and procedures, selected the appropriate chemical equations to represent the reaction, and determined quantitatively the amount of energy released in the explosion. In order to perform the task students had to apply advanced science concepts (i.e., testing unknown gases), manipulate numerical quantities and use specialized course-based knowledge—the general characteristics of the quantitative science dimension. Since advanced science content knowledge and specialized skills were needed to complete the task, it was considered content rich. And, since students conducted their own investigations without step-by-step instructions, it was considered process open. Finally, since students were asked to compare chemical reactions in both fresh and salt water, we considered Aquacraft to be a comparative investigation.

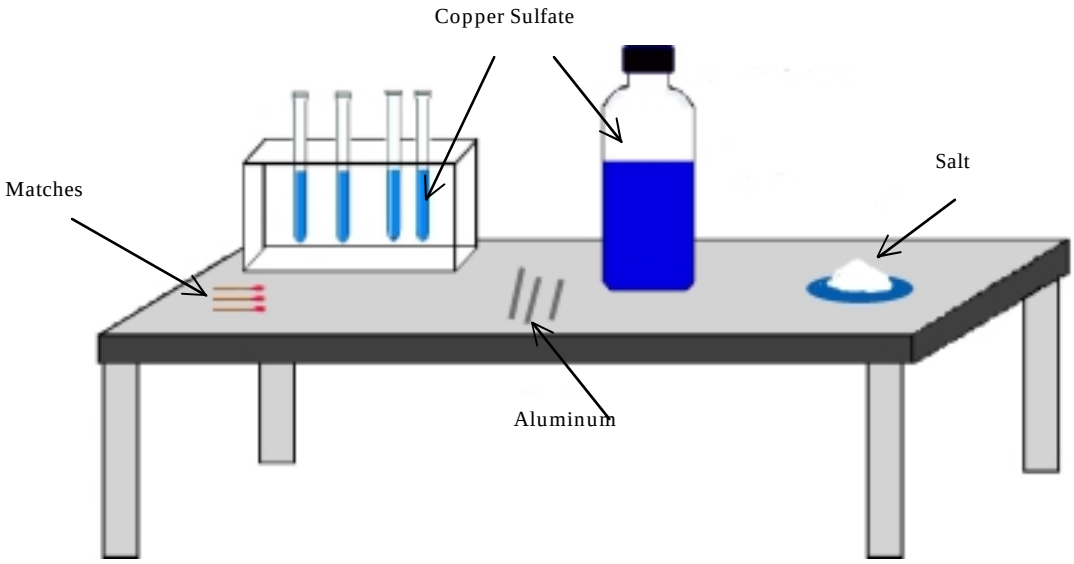

Figure 3. Aquacraft performance assessment.

Table 2 presents the assessments selected and their classification based on the three frameworks. Once the assessments were selected, we administered them to the participants.

Method

Participants

Each of the performance assessments was administered individually to one expert (science teacher) and one novice (physics student). If there was a nominal-task effect on reasoning, prior research on expertise (e.g., Chi, Glaser, & Farr, 1988) suggested that using this extreme group design would allow us to detect the effect. Even though 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; novices are strongly influenced by the specified task features. Hence a "large sample" was unnecessary to detect the effect. Of course, a next step in this research would be to confirm systematic effects if found with multiple experts and novices.

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.

Table 2

Performance Assessment Characteristic 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

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 chemistry.

Talk Aloud Analysis

Students and teachers were asked to concurrently talk aloud while they completed each performance assessment. Talk alouds were audiotaped and transcribed. Similar procedures have been used before to investigate cognitive task demands of assessments (Baxter & Glaser, 1998; Ericsson & Simon, 1993; Hamilton et al., 1997; Ruiz-Primo, 1999).

The most controversial aspects of talk aloud analyses are protocol segmenting and encoding processes (Ericsson & Simon, 1993). Segmenting the protocol is the procedure that is used to divide up a respondent's talk aloud speech into segments. The encoding process is the method that is used to classify each of the segments into categories. These categories are usually based on the research questions asked. In this way, a researcher can make claims about the frequency and kinds of thinking a respondent makes evident while talking aloud.

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

The segmentation protocol involved a two-step process that focused on isolating the smallest meaningful unit of analysis. First, natural pauses at the end of phrases and sentences were used as segmentation markers. And then, since our focus was on reasoning differences, segmentation also attended to maintaining complete meaningful ideas such as "if" statements, numeric calculations, and logical phrases. Finally, extras such as "okay," "dammit," and "Holy cow Batman" were considered as not meaningful and were either kept embedded in their meaningful statements or isolated and scored as extras.

While protocol segmentation was being tried out, encoding categories were developed. These categories emerged from several sources, many of which overlap with each other. The first source for coding categories was the Hamilton et al. (1995) study, which showed differences among the three reasoning dimensions. These categories included calculations, graphing, prediction, making sense, and scientific explanation. The second source of categories came from talk alouds analysis of hands-on performance tasks as described by Hamilton et al. (1997). These included metacognitive skills, application of prior knowledge, expectations, and use of scientific processes. The third source of categories was taken from Baxter and Glaser (1998). These focused on the planning and the knowledge aspects of an assessment. And, finally our own analysis of the verbal data and our understanding of the unique nature of performance assessment suggested others, such as doing something and rationales for actions and conclusions. A blending of all of these sources was the ultimate determinant of the categories that we used to code each segment.

In order to organize the segments, we came up with five super-ordinate categories, which included Assessment Mechanics, Self-Regulation, Scientific Processes, Scientific Reasoning and Concepts, and finally Extras. Each super-ordinate category was further divided into other segment types. There were a total of 16 categories underlying the five super-ordinate categories. These 16 categories were mutually exclusive and comprehensive such that any segment could be placed into one and only one of the 16 categories or into the Extra super-ordinate category.

The Assessment Mechanics super-ordinate category contained segments that reflected the testing situation (such as “Where is the answer sheet?”) and also included segments about accounting for pieces of equipment (surveying materials; Table 3).

The Self-Regulation super-ordinate category captured segments that involved the organization and direction of the execution of the assessment and the monitoring of the execution. It included three types of segments: (a) planning statements, (b) sense making statements and (c) monitoring statements (Table 4).

Table 3

Assessment Mechanics Super-Ordinate Category With Subcategories

Code	Category	Characteristic of segment	Examples from coding
1	Testing mechanics	Segment describes a process of doing the test. Testing taking or organizational segments.	Where is the wire? Here it is. Ok, there, ok number three. Turning the page.
2	Surveying materials	Segment is about materials. Segment describes materials. Segment is description of an apparatus.	There is a wire. The bulb is right here. Got four sticky towers.

Table 4

Self-Regulation Super-Ordinate Category

Code	Category	Characteristic of segment	Examples from coding
3	Planning	Segment describes steps, or describes way-points or goals. Segment is a statement about alternate solution paths.	First I will add copper sulfate to the test tube. I can either add the salt or then the copper sulfate first.
4	Sense making Questioning	Segment is making sense of task A question about what is next A question about what to do.	Is that on there? That's how I did it. How do I do that?
5	Monitoring Checking	Segment is about monitoring progress. Segment is about checking work.	I wonder if this is a proper way of doing this. Let's recheck this; I am going to try this again.

The Scientific Processes super-ordinate category included segments about doing something or reacting to the task. Generally these were observations, but predictions and hypothesis were included here as well. There were five segment types: (a) explorations, (b) doing something, (c) observational outcomes, (d) spatial mechanical observations, and (e) quantifying observations (Table 5).

Table 5

Scientific Processes Super-Ordinate Category

Code	Category	Characteristic of segment	Examples from coding
6	Exploration	Segment is a prediction, hypothesis or a guess. Segment is a statement indicating a test or trial.	I will hook the battery to see what happens. I guess the tower should be in Seattle; I hope this works.
7	Doing something	Segment describes person doing something or something that was done.	I am pouring some out. Measuring 25 grams.
8	Observation of an outcome	Segment is an observation of task. Segment about reporting results. Observation from task, not survey of materials.	There are bubbles in the test tube. This is D bright light.
9	Observation of a spatial mechanical relationship	Segment is a description of the spatial relationship of two objects: location, direction, angle or position.	The sun is directly above the city. The shadow points east.
10	Quantifying	Segment is a numerical response or a measured distance or a time measurement.	The shadow is 12 millimeters long.

The Scientific Reasoning and Concepts super-ordinate category included segments that required some reasoning on behalf of the respondents after they interacted with the task in some way. This category also included segments where subjects brought prior scientific principles or concepts to bear on the task in order to understand or carry it out. This super-ordinate category included 6 segment types: (a) rationales, (b) conditional reasoning statements, (c) basic knowledge conclusions, (d) scientific concepts, (e) spatial mechanical conclusions, and (f) quantitative conclusions and calculations (Table 6).

Furthermore, to look for the basic knowledge and reasoning, spatial mechanical and quantitative science reasoning dimensions across the three performance assessments, the talk aloud segments that could be considered to belong to a particular reasoning dimension were identified. For example, a segment that was coded as a rationale 11 could remain coded 11 if it were a basic knowledge and reasoning rationale or it could be coded 11.1 if it were a spatial mechanical rationale

Table 6

Scientific Reasoning and Concepts Super-Ordinate Category With 6 Subcategories

Code	Category	Characteristic of segment	Examples from coding
11	Rationales	Segment is a rationale for prediction, conclusion or observation. Because…	because they are in my way because it reacts the same way then I would get that kinda of a thing
12	Conditional reasoning statement	Segment is a conditional reasoning statement (if, then) Post outcome.	If the light bulb lights If the gas pops ok so if it is a wire
13	Basic knowledge conclusion	Segment is a concluding statement based on observations or outcomes.	So circuit box D has a light bulb and a battery then it is a battery
14	Scientific concepts	Segment is a scientific principle/concept. Segment is a statement about a concept or idea from memory	Hydrogen atomic weight is one I think that Oxygen weight is 16 Electricity goes down the wire
15	Spatial-mechanical conclusion	Segment is a reasoning about a relationship in area or space. SM conclusions.	then the shadow points east So the angle must be away from the sun
16	Quantitative conclusion Calculation	Segment is a calculation involving numbers, answers included. Or a unit conversion,	Ok, so 64 plus 32 is 96. Or 6 plus 2 is eight Let's quantify this as medium

or it could be coded 11.2 if it were a quantitative rationale. Or, a segment that was a conditional reasoning statement 12 could be further coded as 12.1 if were a spatial mechanical conditional reasoning statement or 12.2 if it were a quantitative conditional reasoning statement. Considering the importance of scientific conclusions, we classified the spatial mechanical conclusions as 15, quantitative conclusions as 16 and basic knowledge conclusions as 13.

Once the segmentation and encoding systems were developed, two raters coded random samples of segments. Initially, agreement between raters ranged from as low as 40% to as high as 100%. Some random samples were more difficult to score than others and some categories were more difficult to distinguish than others. After several days of training and scoring, final agreements averaged 87%.

Results and Discussion

Our respondents' scores on the assessments revealed the expected expert/novice differences (Table 7). The Electric Mysteries novice accurately determined the contents of 2 of the 6 boxes and the Electric Mysteries expert accurately ascertained the contents of all 6 boxes. For comparison, Rosenquist, Shavelson and Ruiz-Primo (2000) found that fifth graders' average score was 3.32 and high school physics students averaged 2.55. The Daytime Astronomy novice scored 34, while our expert scored 60. For comparison Shavelson et al. (1998) found that the Daytime Astronomy mean score with fifth graders was 14. The Aquacraft novice scored 17 and the Aquacraft expert scored 32 out of a possible 42.

These score differences revealed that our students were indeed novices, scoring substantially lower than our teachers. Our teachers were knowledgeable in the subject matter as evidenced by their achieved maximum score in Electric Mysteries, and high scores for Daytime Astronomy. Although the Aquacraft expert score may seem low in comparison to the maximum score, this was more a function of the Aquacraft assessment that was subsequently revised to prompt for more justifications and conclusions. All scoring was done with two raters and any differences in the scoring were discussed until the raters agreed upon a score for each of the six assessments.

Importantly, in all cases, the talk alouds revealed that our experts knew more about the subject matter than was written down on the performance assessment recording forms. For example in Daytime Astronomy our expert's talk aloud

Table 7

Performance Assessment Scores and Number of Segments by Participant Expertise

Performance assessment	Performance assessment score	Number of talk aloud segments
Electric Mysteries		
Expert	6	409
Novice	2	359
Daytime Astronomy		
Expert	34	210
Novice	60	233
Aquacraft		
Expert	32	243
Novice	17	244

revealed substantial knowledge of longitude lines, time zones, and the position of a summer sun that was not expressed in writing on the recording form. Our novice's talk aloud more closely matched what was written down.

To characterize each respondent's talk aloud, we calculated the percent of all segments falling into a category (e.g., conclusions). We did this because although the assessments took similar times to administer, they differed in the number of segments generated (Table 7). The Daytime Astronomy novice talk aloud produced 233 segments, while the Daytime Astronomy expert produced 210 segments. The Electric Mysteries novice generated 359 segments while the Electric Mysteries expert generated 409 segments. The Aquacraft novice generated 244 segments and the Aquacraft expert produced 243 segments. Table 8 contains a section of the 409 segments and codes for the Electric Mysteries expert including *Doing Something* (code 7), *Prediction* (code 6), *Scientific Concepts* (code 14), and *Basic Knowledge and Conclusions* (code 13). Notice how the Electric Mysteries expert makes electrical connections (doing something), then supports these actions with rationales for the actions and then recalls a *scientific concept* to support her thinking and actions.

Table 9 contains two sections of the 244 segments of the Aquacraft novice talk aloud including *Observations* (code 8), *Quantitative Observations* (code 10) and *Quantitative Conclusions* (code 16). Notice how the Aquacraft novice makes observations and conducts trials in the first section. Furthermore, embedded in the statement "second trial" (segment # 224) are assumptions about scientific processes

Table 8

Coding of an Expert's Talk Aloud Segment While Conducting the Electric Mysteries Assessment

#	Code	Category	Segment
103	5	Monitoring	Okay, oops, don't want to lose that part do I,
104	8	Observation	D, let see aha, that is super bright.
105	7	Doing something	I will make a connection,
106	7	Doing something	so run it through here,
107	11	Rationale	so it should still work,
108	11	Rationale	so it still should run through,
109	14	Science concept	so the resistance shouldn't be so much through that wire anyway.
110	6	Prediction	Okay, this will be cool to see if this works anyway. Hot dog.
111	13	Conclusion	Okay, so I think that box D is two batteries,

Table 9

Aquacraft Novice: Segments and Coding

#	Code	Category	Segment
222	8	Observation	Huh, still going…
223	7	Doing something	I am going to do another drawing here.
224	7	Doing something	Second trial.
225	8	Observation	Bum!
272	10.2	Quantitative observation	if there was 675 kg minus 475,
273	16	Quantitative concl. Calculation	Let's see what we have got, 0.0. 200 kg of Aluminum was missing.
274	10.2	Quantitative observation	So if it is 4 to 3, Aluminum to Hydrogen.
275	10.2	Quantitative observation	Now the Aluminum is 4 and Hydrogen is 3.
276	7.2	Doing something quantitative	I am going to do some math here.
277	16	Quantitative concl. Calculation	Divide 200 by 4 equals 50.

and rationales, but since they are not explicit this segment is coded as Doing Something (7). Furthermore, notice how the segment "So if it is 4 to 3, Aluminum to Hydrogen" (# 274) was coded as quantitative observation instead of scientific concepts, this because the assessment states explicitly the ratio of Aluminum to Hydrogen.

Differential Reasoning Evoked by Performance Assessments

We conjectured that the three performance assessments tapped different types of reasoning based on our conceptual analysis of the assessment task demands. That is Electric Mysteries tapped basic knowledge and reasoning, Daytime Astronomy spatial mechanical reasoning, and Aquacraft quantitative science reasoning. In order to bring the talk aloud data to bear on this conjecture, we first totaled the number of segments in each reasoning category for each assessment. Specifically for each assessment separately, we totaled all the statements that were coded as rationales (11), conditional reasoning statements (12), scientific concepts (14) and conclusions (13) into *basic knowledge and reasoning* segments. Similarly, those segments that were coded as rationales–spatial mechanical (11.1), conditional

reasoning–spatial mechanical (12.1), scientific concepts–spatial mechanical (14.1) and conclusions–spatial mechanical (15) were totaled into *spatial mechanical reasoning.* Those segments coded as rationales–quantitative (11.2), conditional reasoning–quantitative (12.2), science concepts–quantitative (14.2) and conclusions–quantitative (16) were totaled together into *quantitative science reasoning.* We then calculated the mean percent of all reasoning segments that fell into each reasoning category for each of the three dimensions (basic knowledge, spatial mechanical, quantitative science) combining the novice and expert talk aloud data.

Differences in reasoning demands were evident in the talk aloud data (Figure 4). First we found that all three assessments drew on basic knowledge and reasoning, 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 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 and more basic knowledge reasoning than expected.

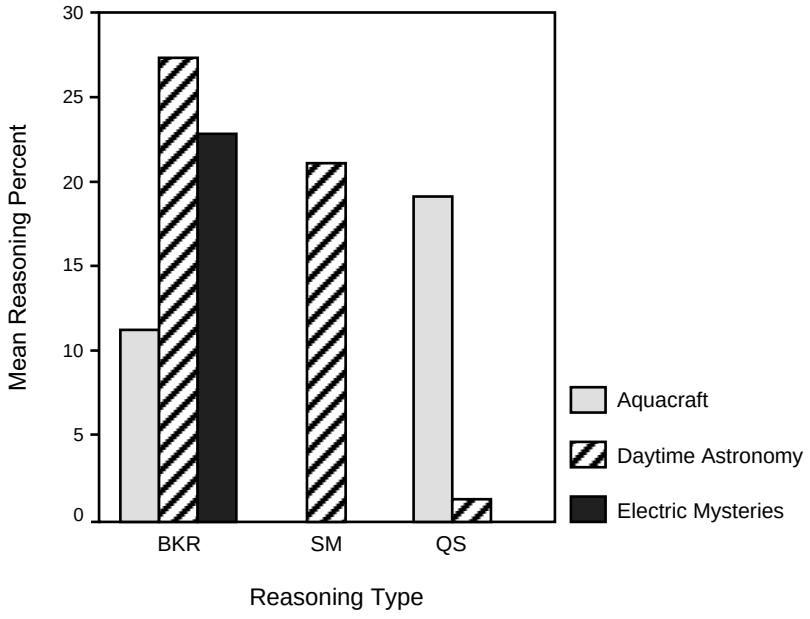

Figure 4. Reasoning demands by performance assessment. BKR is basic knowledge and reasoning, SM is spatial mechanical reasoning, and QS is quantitative science reasoning.

Information in Figure 5 sheds further light on the reasoning differences elicited by the three assessments. Marked differences in the means can be seen in *science concepts*, *observation*, and *monitoring* segments. Electric Mysteries elicited few segments about scientific concepts when compared to Aquacraft or Daytime Astronomy. There were no statements in the Electric Mysteries talk alouds about how to make a circuit. Even so we believe that Electric Mysteries does elicit content knowledge. It seems that our respondents constructed circuits while making statements about connecting this to that. It may be that embedded in these *doing something* segments is knowledge about circuits. This will be investigated further in our large study.

The second important difference across performance assessments arose in the *observation* segments. The percentage of observations elicited by Aquacraft was more than double that of Daytime Astronomy. Our respondents worked out Daytime Astronomy by modeling the rotation of the earth and its relation to the sun and shadows cast and only then by observing to verify model predictions.

Finally, the difference in the mean percentage of *monitoring* segments between Aquacraft and Electric Mysteries suggested another difference in reasoning. Monitoring referred to statements about checking information and about whether respondents were doing something right. The Aquacraft percentage was almost double that of Electric Mysteries. Going back and checking in Aquacraft was done far more often than in Electric Mysteries.

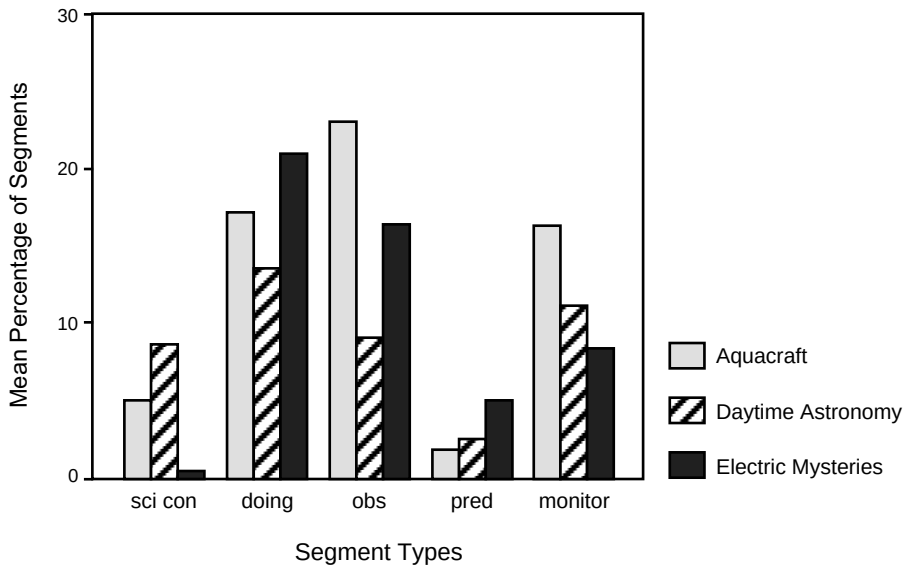

Figure 5. Comparison of mean percent of segment types level by performance assessment.

Expert-Novice Reasoning Differences

We also conjectured that experts' reasoning would differ from that of novices with experts more frequently reasoning when solving scientific problems than novices (Chi et al., 1988; Glaser, 1991; Gobbo & Chi, 1986). In Figure 6 we combined data across assessments to examine the mean percent of the three reasoning dimensions employed by experts and novices. Experts consistently used the three types of reasoning more than novices.

Further analysis of the segments revealed other important differences between experts and novices. Figure 7 shows the mean percentage of segment types for experts and novices across assessments. While percentages of *science concepts*, *observations* and *predictions* are similar for both experts and novices, novices differed from experts in the percent of segments that were *doing something* and *monitoring*. Novices spent more time doing tasks in the performance assessments and then checking and redoing them than did experts, who once they had collected information about the task, reasoned to a solution.

One possible explanation for this difference comes from the expert-novice literature. Experts perceive the principle underlying the observed features of the task and focus on scientific reasoning to reach a conclusion. Novices are attracted by the physical characteristics of the task and "do something" and monitor feedback to guide their problem solving. Of course this interpretation is tentative and awaits evidence from larger numbers of respondents to shed light on its veracity.

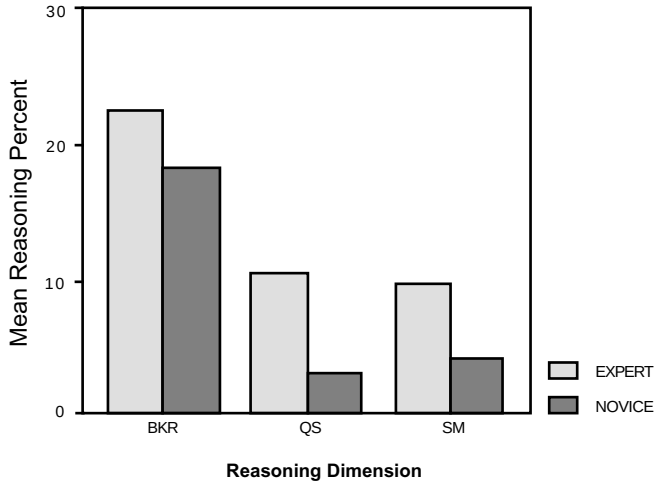

Figure 6. Reasoning elicited from experts and novices.

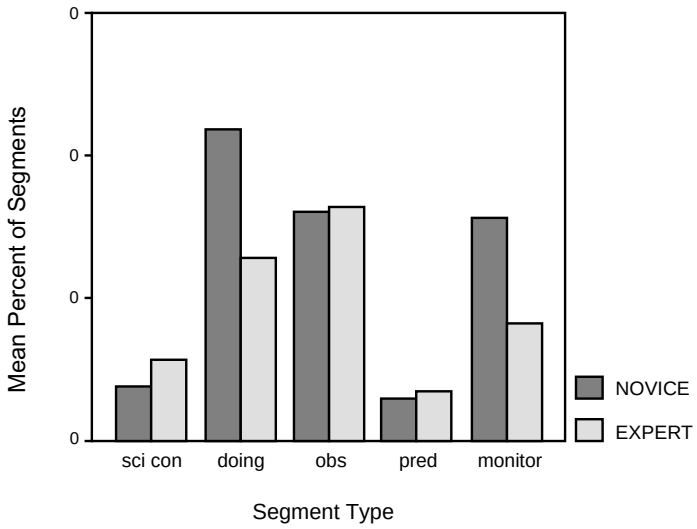

Figure 7. Comparison of mean percent by category elicited from experts and novices.

Conclusions

The main purpose of this 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 mechanical reasoning. By selecting performance assessments using the general characteristics of the reasoning dimensions, and then collecting talk 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 this pilot. Encouraged by these findings we are currently analyzing talk alouds from a group of 35 students across all three performance assessments.

Furthermore, we conjectured that reasoning dimensions generated from multiple-choice tests such as NELS and TIMSS are similar to those elicited by performance assessments. With a larger data set from the full study, we may be able to test the trends found in this small scale study. And, we will be able to connect this reasoning information with the multiple-choice data by linking science achievement scores and science reasoning frequencies.

This study and future studies such as this will shed light on student reasoning in performance assessments. Such studies provide, then, critical tests of the validity of claims made about what performance assessments measure. Furthermore, with

this knowledge in hand, it may be that gaps in student reasoning can be found and then be addressed with instruction that promotes student reasoning. Finally, understanding of the relationship between different performance assessments and the reasoning they elicit can be used by teachers and researchers to design lessons and assessments along specific reasoning dimensions especially since particular content domains may be linked to particular reasoning dimensions.

In this study we did not attempt to investigate Shavelson's task types. Further analysis of this and subsequent data by reformulating categories may lead us to conclusions about the reasoning demands in relation to the Content Process Space and task types. Additionally, since we contended that performance assessments tap reasoning and knowledge in a different way than multiple-choice items, we expect to find in further analyses that the basic knowledge and reasoning, spatial mechanical, and quantitative science dimensions are too limiting.

References

Baxter, G., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17*(3), 37-45.

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

California State Board of Education. (1990). *Science frameworks for California public schools, kindergarten to grade twelve.* Sacramento, CA: California State Board of Education.

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

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

Gobbo, C., & Chi, M. T. H. (1986). How knowledge is structured and used by expert and novice children. *Cognitive Development, 1*, 2221-2237.

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. S. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.

Hamilton, L. S., & Snow, R. E. (1998). *Project 3.1. Construct validity: Understanding cognitive processes- psychometric and cognitive modeling exploring differential item functioning on science achievement tests.* Stanford, CA: CRESST/Stanford University.

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

National Research Council. (1996). *National Science Education Standards.* Washington DC: National Academy of the Sciences.

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, A. (2000). *On the "exchangeability" of hands-on and computer-simulated science performance assessments* (CSE Tech. Rep. No. 531). Los Angeles: University of California, 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.

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

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

Shavelson, R. J., Solano-Flores, G., & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. *Evaluation and Program Planning, 21*, 171-84.

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.

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 evidenced 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 most 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 evidenced based, focusing on evidence and explanations.	Component identification	BKR
8	Batteries	Student determines which batteries are good or not. Scoring is evidenced based, focusing on evidence and explanations.	Component identification	BKR
9	Magnets	Student identifies which magnet is stronger. Scoring is evidenced 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

#	Performance assessment	Task, response and scoring	Classification system	Type of reasoning
11	Plasticine	Student weighs different amounts of plasticine as carefully as possible. Scoring is evidenced based, focusing on evidence and explanations.	Comparative investigation	BKR
12	Shadow	Student finds out the change in size of a shadow made by a card placed between a light and a screen when 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 need 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 evidenced 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 evidenced 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

#	Performance assessment	Task, response and scoring	Classification system	Type of reasoning
23	Animals CLAS	Student determines the possible causes of a fish decline. Scoring is evidenced 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