C-R-E-S-S-T

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

Technical Report

You can view this document on your screen or print a copy.

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

Cognitive Analysis of a Science Performance Assessment

CSE Technical Report 398

Gail P. Baxter and Anastasia D. Elder
University of Michigan/CRESST

Robert Glaser
CRESST/LRDC/University of Pittsburgh

March 1995

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

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

COGNITIVE ANALYSIS OF A SCIENCE PERFORMANCE ASSESSMENT[1]

Gail P. Baxter and Anastasia D. Elder
University of Michigan/CRESST

Robert Glaser
LRDC/University of Pittsburgh/CRESST

Abstract

In this paper we demonstrate an approach for examining the cognitive activity students engage in during a science performance assessment. Thirty-six fourth- and fifth-grade students were interviewed while they conducted an investigation to determine the properties of various powders. Interview protocols and observations were analyzed with respect to several characteristics of proficient performance such as planning, monitoring, solution strategy, and explanations. Students with differing levels of competence (e.g., high and low scorers) could be distinguished in terms of these characteristics. High scorers provided a more complete plan for approaching the task, were more strategic in their problem-solving approach, engaged more frequently in a variety of self-monitoring activities, and generated better explanations of content-related concepts than low scorers. The results suggest the viability of this approach for analyzing the extent to which performance assessments measure higher order thinking.

Alternative forms of assessment have been proposed as a major factor in the current impetus for educational change. These assessments are intended to evaluate student understanding and provide models of performance that educational practice should foster in all students. For example, assessments developed to support and enhance instruction in hands-on science classrooms ask students to reason with subject matter knowledge to solve a problem. Scoring is designed to focus on the thinking and reasoning processes by which the solution is generated and key aspects of the performance drawn from the principles underlying the topic.

---

[1] The Mystery Powders assessment was developed through grant ESI 90-55443 from the National Science Foundation (NSF) to the first author and her colleague, Richard J. Shavelson, University of California, Santa Barbara. Opinions expressed are those of the authors and not necessarily those of NSF. Thanks to Tim Breen, University of Michigan, Jasna Jovanovic, University of Illinois at Urbana-Champaign, and Karen Malhiot, Chicago Public Schools, for their help in data collection.

These complex, open-ended, performance assessments are purportedly designed to engage students in higher order thinking processes as they develop solutions to problems. Moreover, it is assumed that the scoring systems are able to differentiate levels of student competence. Despite widespread support for these kinds of assessments, criteria for evaluating their effectiveness continue to revolve around traditional concerns for reliability and validity. Changes in the assumptions underlying test development and use require changes in the kinds of evidence necessary to support interpretations of student performance (e.g., Linn, Baker, & Dunbar, 1991; Moss, 1992, in press; Shepard, 1993). In particular is concern for an evaluation that documents the nature and extent of student thinking and reasoning required for optimal performance. A strong positive relationship between performance score and quality of thinking and reasoning processes would provide evidence to support claims that these assessments measure higher order thinking (Baxter, Glaser, & Raghavan, 1993). Nevertheless, strategies to guide this type of evaluation are not well established. This paper attempts to address this gap in the literature.

More specifically, we suggest a methodological approach that relies on protocol analysis techniques developed by cognitive psychologists in their studies of problem solving in knowledge-rich domains. We demonstrate the viability of this approach with "Mystery Powders," a classroom-based assessment currently being piloted by several large school districts. Our intent is to provide a working example of one possible approach for carrying out a cognitive analysis of a science performance assessment.

Cognitive Analysis

The shift from multiple-choice to performance-based assessments has called attention to the need for expanded conceptions of validity (e.g., Linn et al., 1991; Messick, 1994; Moss, 1992; Shephard, 1993). The implicit assumption that performance assessments will improve instruction requires explicit attention to the utility and consequences of test use (e.g., Messick, 1994) and what the test claims to do (Shephard, 1993). Linn et al. (1991) further explicate validity criteria by describing several aspects of performance assessments that merit attention. Most notable is a concern for the cognitive complexity of these assessments—an aspect that clearly distinguishes performance assessments from traditional achievement measures. Evidence for the cognitive complexity of performance

assessments has, for the most part, been assumed on the basis of task complexity and/or task difficulty in a psychometric sense. Little attention has been paid to an evaluation of the kind and quality of cognition required for optimal performance (Glaser, Raghavan, & Baxter, 1992) or to procedures for carrying out this type of evaluation.

"Claims that performance assessments measure higher order thinking skills and deep understanding, for example, require detailed cognitive analysis" (Baker, O'Neil, & Linn, 1993, p. 1216). Detailed cognitive analysis should illustrate the kind of performance actually elicited from students in alternative assessment situations and document the relationship between those performances and the problem-solving activities that contribute to optimal performance (Glaser et al., 1992). If the test is intended to tap certain skills and processes, then an individual's score should reflect his or her level of proficiency with respect to those skills and processes. Establishing links between performance scores and processes of thinking and reasoning provides evidence to support inferences that these assessments are cognitively complex.

Characteristics of Proficient Performance

Cognitive studies of expertise suggest some characteristics of student performance that develop as students display increasing proficiency in problem solving and higher order thinking (cf. Chi, Glaser, & Farr, 1988). First, proficient or competent students are characterized by integrated knowledge that fosters their ability to reason, explain, and make inferences with what they know. Second, these students effectively represent the meaning of a problem and plan an approach before employing a solution strategy. Third, their strategy for problem solving is reasoned and efficient, not a trial-and-error process. Finally, competent students have a repertoire of well developed self-regulatory skills that they use to monitor their performance. Because the quality of the aforementioned characteristics indicates relative proficiency with a subject matter, they provide a useful framework for analyzing students' thinking and reasoning in an assessment situation.

To test the viability of this cognitive analysis framework, we document the kind and level of cognitive activity students engage in while conducting a science performance assessment, "Mystery Powders." Specifically, we examine the relationship between students' performance assessment scores and the extent to

which they (a) explain principles underlying task performance, (b) generate a knowledge-based plan for approaching the task, (c) utilize a principled problem-solving strategy, and (d) monitor their performance while carrying out the task.

Mystery Powders Unit and Assessment

The *Mystery Powders unit* engages students in systematic investigation of the properties of substances. Students study five simple, white powders—salt, sugar, baking soda, cornstarch, and plaster of paris. The purpose of the unit is to help students understand (a) that each powder has a unique set of properties, some physical and some chemical; (b) that particular properties are more salient than others and therefore more reliable for identification purposes (e.g., iodine turns black when mixed with cornstarch but not with other powders); and (c) that a combination of confirming and/or disconfirming evidence may be required for the identification of a powder(s).

To develop this understanding, teachers guide students through systematic investigation of each of the powders. Students document the tests (adding water, vinegar, iodine, or heat), observations, and information from sensory input (smell, touch, taste, and observation with a hand lens) for each of the five powders in their science journals. Comparing and contrasting the accumulated information draws attention to the differential reliability of each method for distinguishing one powder from another. For example, iodine turns purple/black with cornstarch but yellow or orange with all the other powders. Vinegar fizzes with baking soda but not with any of the other powders. Sugar melts and becomes caramel-like when heated. Salt has a unique crystal structure (uniform, cube-shaped) when viewed under a hand lens. Plaster of paris becomes hard when mixed with water and, unlike the other powders, will remain hard when water is mixed with it again.

For the *Mystery Powders assessment* students are asked to identify the contents of six bags from a list of five possible options; some bags contain individual powders, some contain two powders. Students engage in an iterative sequence of generating hypotheses, testing out hypotheses, evaluating observations, and drawing conclusions until a solution is reached. They use their journals from science class as a resource when completing the assessment.

Cognitive Expectations

It was reasoned that if the Mystery Powders assessment requires higher order thinking for optimal performance, then students with high scores would plan, explain, problem solve, and monitor their performance on the assessment at a qualitatively different level than students with low or midrange scores. More specifically, students who understand the concepts and processes central to the Mystery Powders unit are expected to (a) generate a coherent explanation that reflects their understanding of the knowledge and processes underlying the unit; (b) provide a plan that guides their solution strategy based on their representation of the task and the principles on which performance is dependent; (c) engage in an efficient, knowledge-based strategy for solving the problem; and (d) monitor their thinking and reasoning, correcting errors and recognizing inconsistent findings from one test to another.

In contrast, students who lack a general understanding of the relationship between the powders and the tests and observations, and how these might be used to identify the contents of each bag, are expected to (a) provide inadequate or fragmented explanations of task-related concepts; (b) begin solving the task without a plan that guides their solution; (c) use a trial-and-error strategy to solve the task; and (d) fail to monitor effectively.

Students with partial knowledge might be expected to vary in the quality of their cognitive activity. For example, these students may plan effectively and conduct a principled investigation, yet provide incomplete explanations of task-related concepts and fail to monitor their work. These students, then, demonstrate some characteristics of proficient performance but not all.

In this paper, we examine student performance on the Mystery Powders assessment. Using the characteristics of proficient performance as a guide, we document the nature and extent of student thinking and reasoning in this assessment situation. Correspondence between performance scores and quality of cognitive activity provides evidence that this assessment requires higher order thinking.

Method

Subjects

The Mystery Powders assessment was administered to 36 fourth- and fifth-grade students within one week of completing an eight-week, activity-based unit of study on the properties of substances. Students represented a diverse range of ethnic/cultural backgrounds, and males (n = 19) and females (n = 17) were equally represented. All students lived in an urban or large suburban school district where they participated in districtwide, inquiry-based science programs for two or more years. Students were chosen by their respective teachers in each of six different classrooms with the expressed purpose of ensuring a range of science ability. Interviewers were unaware of the teachers' rankings of students' science ability.

Instrumentation

Task. Students were asked to identify common white powders such as salt, baking soda, and cornstarch contained in each of six bags—some individually (e.g., baking soda), some in combination (e.g., baking soda and cornstarch). The possible contents of the bags were clearly conveyed to the students—baking soda, cornstarch, baking soda and salt, cornstarch and salt, and baking soda and cornstarch (see Figure 1).

Each of the six bags contained one of the possible options; two of the bags contained the same thing. Students were provided with iodine, water, vinegar, and a hand lens to conduct their investigations. Students could also consult their science journals where they recorded the results of their investigations of each of the powders during science class. After conducting tests of the six bags of powders, making observations, and recording their notes, students were asked to summarize their findings in a table of results and conclusions (see Figure 2).

Students may have referred to the powders by the letters A, B, C, D, E, and F because during the course of instruction the intent was not to identify the powders by name but to focus on observing those properties that distinguish one powder from another. As such, some students did not know the "names" of the powders and referred to them only by a letter (e.g., baking soda is powder D).

Find out what is in each of the bags 1, 2, 3, 4, 5, and 6. Use any of the equipment on the table to help you determine what is in each bag. **Each bag has <u>one</u> of the "Mystery Powders" listed below.**

Baking Soda	*(D)*
Cornstarch	*(A)*
Cornstarch and Baking Soda	*(A and D)*
Baking Soda and Salt	*(D and C)*
Cornstarch and Salt	*(A and C)*

NOTE: Two of the bags will have the same thing. All of the others will have something different.

==========

<u>Keep notes</u> on what test(s) you did and what you observed as you conduct your investigation. You have room on the following pages. Use your notebook from science class to help you determine what each powder is. When you think you know what is in a bag, record your results and conclusions in the table on the last page.

Figure 1. Mystery Powders assessment.

Scoring. Student responses, recorded on the results and conclusions page of the test booklet, were transcribed to a score form (see Figure 3). Points were awarded for identification of the substance(s) in each bag *and* the evidence provided to support the identification. For the identification score, students received one point for correctly identifying the contents of a bag; zero points for incorrect identification. Students did not receive partial credit for identifying one powder out of two powders in a bag. For example, if students indicated that there was cornstarch in bag 1, they received zero points because bag 1 contained cornstarch and baking soda. The points were summed over all six bags yielding a total correct answers score ranging from zero to 6 (see left-hand column of Figure 3).

RESULTS AND CONCLUSIONS

Look at the tests and observations you made today.
Check your observations with the notebook you kept during science class.
Fill in the table below.

Mystery Powder	What's inside the bag?	What test(s) told you?	How did you know? What happened?
1	*Cornstarch & Baking Soda*	*Vinegar and iodine.*	*When I added vinegar it fizzed, and when I added iodine it hardend and turned black.*
2	*Baking Soda & Salt*	*Vinegar and iodine.*	*When I added vinegar it fizzed, and when I added iodine it became two layers.*
3	*~~Cornstarch~~ & Baking Soda salt*	*Vinegar and iodine and water.*	*When I added vinegar it turned cloudy, When I added iodine it also turned cloudy and milky.*
4	*Cornstarch*	*Water, Vinegar, iodine.*	*When I added water it harden, when I added vinegar it really hardened, and when I added iodine it turned black.*
5	*~~Cornstarch~~ & salt Baking Soda*	*Vinegar and iodine.*	*When I added vinegar it started to turn cloudy, when I added iodine it turned black.*
6	*Cornstarch*	*Vinegar, iodine, water.*	*When I added vinegar it fizzed. When I added iodine it turned black. When I added water, it made it easier to compare.*

Figure 2. Results and conclusions page from student test booklet.

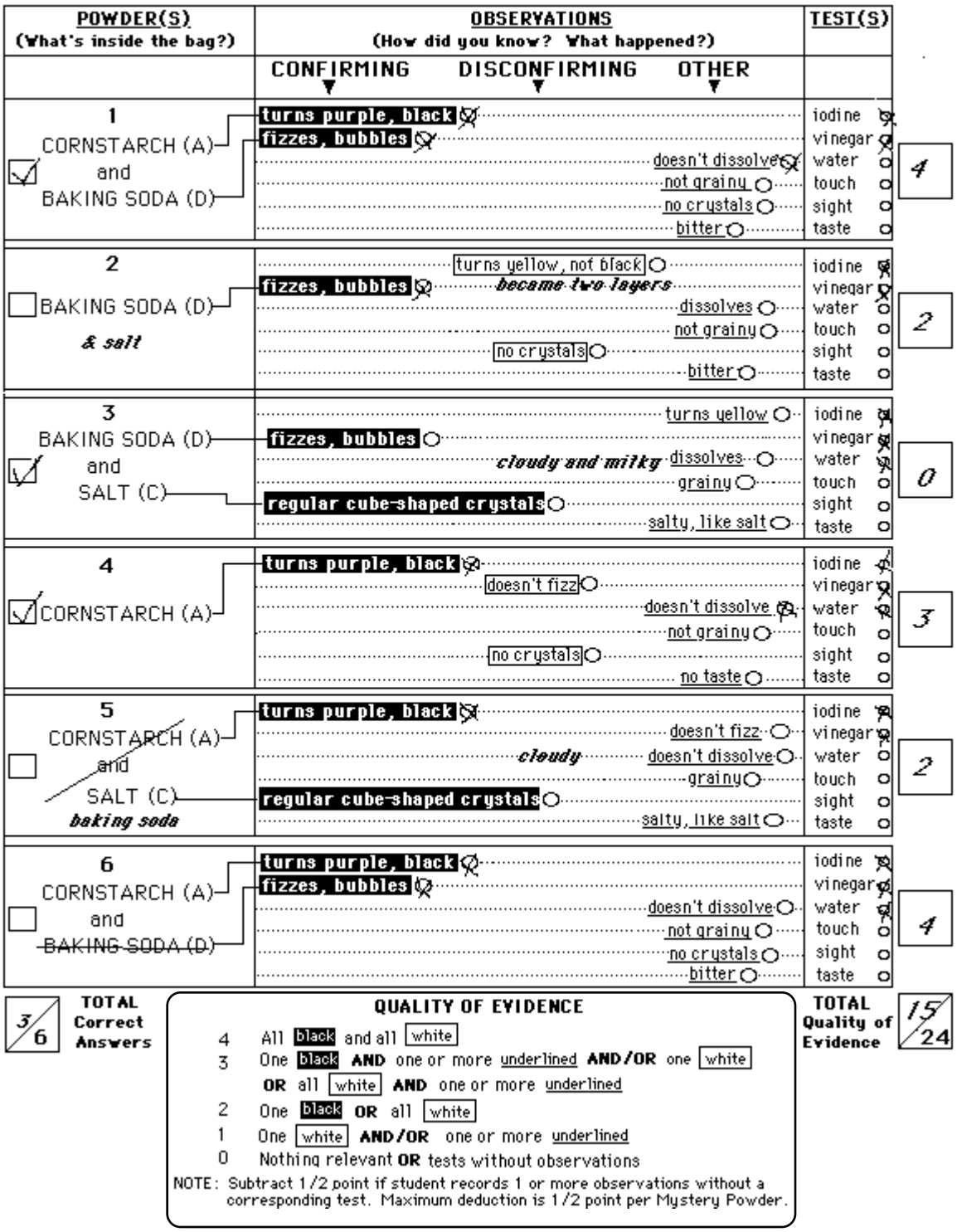

Figure 3. Mystery Powders score form.

For the quality of evidence score, students received 1 to 4 points for providing appropriate support for their answer. Tests and observations constitute evidence, and students were evaluated on the quality of the evidence they provided (see bottom of Figure 3). Quality was dependent to some extent on the contents of the bag. When the "Mystery Bag" contained two powders (e.g., cornstarch and baking soda), students who confirmed or ruled in the presence of each powder (provided complete evidence for both powders) received 4 points. Note that the confirming tests are those indicated with a black box on the score form. Students who provided complete evidence for one powder but incomplete evidence for the other powder received 3 points. Observations that resulted in partial or incomplete evidence are indicated with a line drawn under them. Students who provided complete evidence for one powder but no evidence for the other powder received 2 points. Students who provided incomplete evidence for both powders or inadequate evidence received 1 and zero points, respectively (see Figure 3). For example, if students reported tests (e.g., vinegar, iodine) without corresponding observations (fizzed, turned black) zero points were awarded.

When only one powder was present (e.g., baking soda), students had to confirm or rule in the presence of that one powder and disconfirm or rule out the presence of all other powders which may have been in combination with that powder (cornstarch, salt) to receive 4 points. Note that the disconfirming tests are indicated by white boxes on the score form. As was the case with two powders, points were awarded based on the quality of evidence. The less complete the evidence confirming the presence of one powder and disconfirming the presence of a second, the lower the score (see Figure 3).

If a student provided observations without tests, one-half point was subtracted from his or her evidence score. For example, consider a student who reports testing with vinegar and reports two observations (fizzed and turned black) for bag 1(baking soda and cornstarch). Fizzed with vinegar confirms the presence of baking soda and turned black with iodine confirms the presence of cornstarch. This student would initially receive 4 points. However, the student failed to report the test that led to the observation "turned black." One-half point was subtracted for a final score of $3\frac{1}{2}$ points. In scoring then, reporting observations is more important than reporting tests.

As an example, consider the student responses in Figure 2. For bag 1, the student received 1 point for the answer (identified the powders as cornstarch and baking soda). As evidence, the student reported fizzing with vinegar and turned black with iodine. These tests and observations are considered complete evidence for both powders. Therefore, the student received 4 points for evidence (see Figure 3). As a second example, consider the student's response for bag 2 (see Figure 2). The student incorrectly identified the contents of the bag as baking soda and salt (identification score = zero). As evidence, the student noted fizzing with vinegar and "became two layers" with iodine. The student did not provide evidence to rule out cornstarch (iodine turns it yellow, not black) or salt (no crystals). These tests and observations were required for complete evidence (4 points). The student received 2 points for providing complete evidence for one powder (baking soda; see Figure 3).

Procedure

Students were interviewed and audiotaped, individually, while they conducted the assessment task, that is, while they tried to identify the powder(s) in each of the six bags. Directions were read aloud to all students, and all equipment was introduced (vinegar, iodine, water, hand lens, spoons, stir sticks, cups). Students were encouraged to keep notes and record their observations as they conducted the tests. After hearing the instructions, but before the students began, they were asked whether they understood the task and to explain what they were being asked to do. Next, they were asked about their *plans* for completing the assessment ("Can you tell me how you're going to go about it?").

During the assessment, students were prompted with questions to simulate a think-aloud procedure (see Ericsson & Simon, 1993). The think-aloud procedure was not intrusive; rather, students were encouraged to talk about their procedures and the thinking underlying them while carrying out the assessment (e.g., "Why are you adding iodine?") and when drawing conclusions (e.g., "How do you know it's baking soda and salt?"). Interviewers recorded the *strategies* students used. For example, interviewers noted the sequence of tests conducted on each bag. Also, they noted when students referred to their science journal or the list of possible contents for each of the six bags. This information, in conjunction with student verbal comments, was used as evidence of *monitoring*.

After completing all tests and observations, students were prompted to look at their notes, consult their science journal, and summarize their findings in a table of results and conclusions. Interviewers then asked students if they could determine the contents of each bag by using only water (*explanation).*

Results and Discussion

Transcriptions of the audiotaped interviews, interviewer's written observations of students' strategies and activities (e.g., referred to science journal to check observations, hypotheses, or conclusions), and students' performance (evidence) scores served as data. In the following sections we describe (a) the nature of student thinking and reasoning with respect to four characteristics of proficient performance (explanation, plan, strategy, and monitoring activity), and (b) the correspondence between students' scores and the quality of the cognitive characteristics they display. Measures of association are used to describe the relationship between assessment scores and each characteristic of proficient performance (e.g., Everitt, 1977; Hays, 1981). Distinctions between high- and low-scoring students on each of the aforementioned characteristics combined with strong associations between performance score and this cognitive activity provide evidence to support inferences that this assessment taps relevant higher order thinking.

Performance Scores

Two raters read and scored each student booklet (r = .93). On average, students correctly identified 2.6 of the 6 bags (sd = 2.12). Mean evidence score on the task was 9.28 (sd = 4.92). Although possible evidence scores could range from 0 to 24 (6 bags x 4 points per bag), the maximum score obtained in this sample of students was 18 (see Table 1). Indeed, the scores were quite low, with two thirds of the students scoring 11 or less. Nevertheless, reliability of the Mystery Powders assessment is sufficiently high (α = .82). All results are presented using evidence scores because these scores are based on the procedures students carried out, not just the answer they provided.

Table 1

Distribution of Evidence Scores

Group	Range of evidence scores	Number of students	Percentage of students	Group Mean	Group *SD*
High	20-24	0	0		
	16-19	5	14	16.9	1.1
Medium	12-15	7	19		
	8-11	11	31	10.8	2.6
Low	4 - 7	9	25		
	0 - 3	4	11	4.3	2.4
Total		36	100	9.3	4.9

For the analyses presented in the following sections, we collapse adjacent rows in Table 1 and consider three groups: High (scores from 16-19), Medium (scores from 8-15), and Low (scores from 0-7). The three groups can be distinguished on the basis of mean evidence scores, $F(2, 33) = 56.88$, $p < .001$. Simultaneous pairwise comparisons indicate that mean performance for the High group was, on average, significantly greater than mean performance of the Medium group ($p < .05$), which was significantly greater than mean performance of the Low group ($p < .05$).

Explanation

The purpose of the Mystery Powders unit is to develop students' understanding of the properties of substances and the types of tests that can be used to indicate those properties and thereby identify a substance. After completing the Mystery Powders assessment, students were asked, "Could you identify all the powders by testing with water only?" Students who understand that various tests are differentially effective for the identification of each of the powders are expected to explain that one test would not adequately distinguish among the powders.

Students' responses were evaluated with respect to the quality of their explanation (see Figure 4). Two raters categorized student responses according to

Level 1:	Inadequate. Incorrect response or "I don't know." *"Yes it could get watery and you'd say cornstarch."*
Level 2:	Partial. Correct response with specific example. *"No, because when I put vinegar on powders it did make bubbles, but when I put water on it, it didn't make that many bubbles, it just sank down. So that's why I think no."*
Level 3:	Good. Correct response with general description. *"No, because they're different powders. They don't do the same things. And there are other things that would tell me what they are and water wouldn't.*

Figure 4. Quality of explanations of task-related concepts.

one of three levels: inadequate, partial, good (r = .97). Each level was distinguished by the completeness and coherence of students' explanations: (a) inadequate responses include restating the question, "I don't know" statements or incorrect responses; (b) partial responses describe a particular occurrence or specific example; (c) good responses provide generalized explanations.

As might be expected from the score distribution presented above, the majority of students were unable to offer a generalized explanation that reflected an understanding of the necessary and sufficient evidence to identify each substance. Nevertheless, approximately 20% of the students provided a good quality explanation. One half of the students could articulate an example of an instance when water would be insufficient to identify a substance (i.e., partial). The remaining 30% of the students failed to recognize that water would not be a sufficient and/or necessary test given the substances they were trying to identify—salt, cornstarch, baking soda.

We examined the correspondence between students' evidence scores and the quality of their explanations (see Table 2). Students who scored high provided good explanations that reflected their understanding of the differential reliability of each of the various tests and observations. In contrast, all students who scored from

Table 2

Proportion of Students by Score Level Generating Explanations of Various Quality

Range of evidence scores	Number of students	Quality of explanation		
		Inadequate	Partial	Good
16-24	3	–	–	1.0
8-15	17	.2	.6	.2
0-7	8	.5	.5	–
Total	28[a]			

[a]Data are not available for 8 of the 36 students.

zero to 7 provided partial or inadequate explanations. Students with scores in the middle range (8 to 15) varied in the quality of their explanations from inadequate to good. Our analysis indicates a strong association between quality of explanation and evidence score (γ = .78). The most complete, coherent explanations were offered by students with high scores; inadequate explanations were, for the most part, provided by students with low scores.

Plan

Before beginning their investigation of the six Mystery Powders, students were asked, "Can you tell me how you are going to go about it [your investigation]?" Students who understand the properties of the powders and what constitutes necessary and sufficient evidence for the identification of each powder are expected to articulate a complete plan which will guide their solution strategy. For example, students might say they would test with vinegar to indicate which bags had baking soda, with iodine to indicate which bags had cornstarch, and with a hand lens to indicate which bags had salt.

Student responses were categorized into one of three levels (r = .93). Each level varies with respect to two criteria—the completeness with which the problem is represented and the integrity of the procedures for carrying out the investigation (see Figure 5). Approximately 80% of the students described procedures or materials (e.g., test, use vinegar) without reference to how the information might be used to identify the powders (Level 1 or 2). The remaining students displayed some general understanding that powders have properties and

Level 1:	Restates problem or procedure. *"test them"* or *"check them."*
Level 2:	Names tests. *"By observing it and putting water or vinegar or iodine in it. And that's how I'll find out."*
Level 3:	Names tests and reactions specific to one or more powders. *"Okay I'm going to open it and first try the vinegar on it... I'll know its baking soda because of the bubbles, so and like cornstarch feels soft."*

Figure 5. Quality of plans.

that each of the tests is a differentially effective method for identifying those properties. Their plans (Level 3) explicitly mentioned the relationship between a test/observation and a particular powder.

Proportions of students providing each level of plan (1-3) varied with evidence score (see Table 3). An examination of the correspondence between evidence score and quality of plan indicates that good quality plans are associated with high evidence scores and poor quality plans are associated with low evidence scores (γ = .71). As expected, students with performance scores between 8 and 15 varied in the quality of their plan. Some students in this score range produced a high-quality plan, some students produced a low-quality plan.

Strategy

The Mystery Powders unit is intended to develop students' understanding of the distinguishing properties of powders, which on the surface look quite similar (i.e., white), and the utility of using these properties for identification purposes. This understanding will be reflected in the strategy students use to determine the contents of each bag. Given that all six bags contain one of three substances (cornstarch, baking soda, salt) either individually or in combination, it was expected that competent students would be efficient and test the powders with iodine, vinegar, and a hand lens. Iodine will indicate if cornstarch is in the bag,

Table 3

Proportion of Students by Score Level Generating Various Quality Plans

Range of evidence scores	Number of students	Quality of plan 1 Low	2	3 High
16-24	5	–	.4	.6
8-15	18	.3	.5	.2
0-7	13	.5	.5	–
Total	36			

vinegar will indicate if baking soda is in the bag, and examination of the powder under a hand lens will indicate if salt is present because of the regular, cube-shaped appearance of salt crystals. Carrying out less reliable tests such as mixing with water, tasting, or touching suggests that the student has a limited understanding of the principles that underlie the testing procedures taught as part of the Mystery Powders unit.

Strategy information was obtained from interviewer records of the nature and sequence of tests students carried out during their investigation. We examined the number and type of tests (vinegar, iodine, water, hand lens, taste, touch) students used for each of the six bags. Student strategies were categorized on the basis of the adequacy of the tests they conducted and the systematicity of their approach (see Figure 6). Two raters coded student strategies into four categories: Level 1 (inadequate) to Level 4 (systematic and adequate). Interrater reliability was high (r = .95).

An examination of the data indicated that one half of the students carried out their investigations without sufficient information to draw conclusions. These students (Levels 1 and 2) did not conduct enough tests nor did they test each bag with a reliable test such as vinegar or iodine. The remaining one half of the students were systematic in carrying out their investigations. These students (Levels 3 and 4) appeared to recognize the necessity of conducting the same tests on each bag so as to compare information across bags. Nevertheless, gathering information from multiple tests—some informative, some not informative—does not reflect efficient performance, a characteristic of proficient students.

Level 1:	Inadequate. Students used one uninformative test (e.g., water, touch, taste) per bag.
Level 2:	Trial-and-Error. Students used one or more tests per bag but did not consistently test each bag with a reliable test. The student may have tested bag 1 with vinegar, bag 2 with water, bag 3 with iodine, etc.
Level 3:	Systematic and Inadequate. Students used more than one test per bag, conducted vinegar **or** iodine **and** one or more other tests (e.g., water) on all the bags.
Level 4:	Systematic and Adequate. Students used three or more tests per bag, conducted vinegar **and** iodine on all of the bags **and** one or more unreliable test(s) on all the bags.

Figure 6. Quality of problem-solving strategies.

Results indicate that student strategy varied with evidence score (see Table 4). In general, higher scoring students used a more systematic strategy than their lower scoring peers ($\gamma = .83$). Eighty percent of these students invoked the strategy that you gather all possible evidence before reaching a conclusion (i.e., Level 4). Although this strategy is apparently effective—that is, it results in high scores—it is not as efficient as might be expected from the most proficient students who would exclude unreliable tests.

The trial-and-error strategy of the lowest scoring students indicates their lack of understanding of the differential effectiveness of each of the tests for a given powder. Three of the 13 low-scoring students did not use vinegar, iodine, or the hand lens to test any of the bags. Rather, each of these students relied on one test (i.e., taste, or touch, or water) to identify all the powders.

Table 4

Proportion of Students by Score Level Displaying Various Quality Strategies

Range of evidence scores	Number of students	Quality of strategy			
		1 Low	2	3	4 High
16-24	5	–	–	.2	.8
8-15	18	–	.4	.4	.2
0-7	13	.2	.6	.2	–
Total	36				

Students with scores between 8 and 15 varied in their strategies. Many of these students displayed a similar strategy to that of the highest scoring students by systematically testing bags (i.e., Levels 3 or 4). However, about 40% of these students appeared to randomly test the bags using vinegar or iodine or water on any given bag. Like low-scoring students, they did not conduct the same test on each bag.

In sum, a strong relationship was found between student strategies and their performance score. High scorers could be distinguished from low scorers on the basis of their approaches to the assessment. Nevertheless, all students, regardless of their score, tested the powders with one or more unreliable tests such as water, touch, or taste. Low-scoring students relied largely on these less reliable tests. In contrast, the highest scoring students in this sample used the results of these tests as one of multiple pieces of evidence they gathered before arriving at a conclusion.

Monitoring

In carrying out the Mystery Powders assessment, students should attend to and coordinate multiple pieces of information: knowledge of task constraints, knowledge of critical aspects of their previous investigations, and interpretations of current trials. Simultaneous attention to these pieces of information demands that students apply a range of monitoring skills to check their thinking and reasoning throughout their investigation. It was expected, then, that competent performance would be distinguished by frequent and flexible monitoring.

Transcripts of each audiotaped session were coded for evidence of monitoring. We considered several types of statements as reflecting monitoring activity: (a) Refer to Journal—check their hypotheses, observations or conclusions with results of previous investigations recorded in their science journals; (b) Check Options—look at the options or choices given in the instructions thus indicating an effort to operate within the constraints of the task; (c) Retest Bags—confirm observations or conclusions by retesting a bag; (d) Recognize Problem—recognize that a hypothesis was not confirmed or that a test appeared to duplicate findings of another bag thereby suggesting an error might have been made; and (e) Compare Bags—examine tests and observations across bags noting similarities. A second rater coded 8 of the 36 transcripts (i.e., 22%) for each type of monitoring. Agreement on the monitoring categories ranged from .84 to 1.00.

The monitoring activity of students in the assessment situation can be described in three respects. First, three quarters of the students engaged in at least one of the five forms of monitoring described above. The remainder of the students relied largely on their memory of classroom experiences to inform their current investigations with the powders. Second, on average, students monitored on five separate occasions during their investigation. Third, these students typically used two forms of monitoring—check the list of options (Check Options) and compare the results of the current investigation with results of prior investigations recorded in their science journal (Refer to Journal).

To examine the correspondence between students' scores and their monitoring activity, we first considered the pattern of monitoring students displayed (see Figure 7). Eighty percent or more of the highest scoring students consulted their notebooks to compare their tests and observations and checked the list of possible contents for the six bags to check their conclusions. In addition, approximately 60% of these students compared their tests and observations across the bags, noting similarities and differences, and recognized problematic situations when their hypothesis did not match the evidence from their investigation. These three forms of monitoring yield immediate, adaptive feedback/information to help students operate within the constraints of the task.

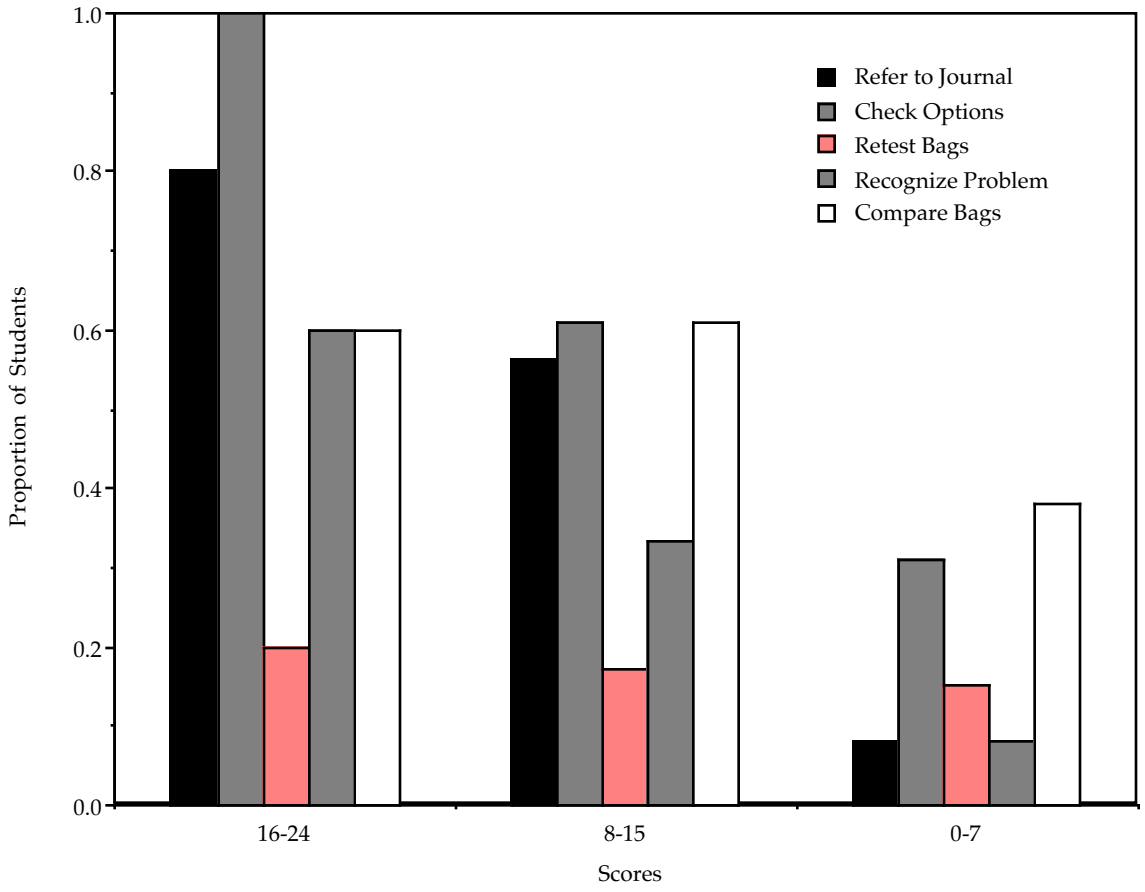

Figure 7. Types of monitoring by score level.

Students who scored in the 8-15 point range displayed a pattern of monitoring similar to that of the highest scoring students; more than half consulted their journals, compared their tests and observations across bags, and referred to the list of options as a way to check their answers. Nevertheless, as a group, proportionately fewer students engaged in four of the five types of monitoring compared to the high-scoring students. Lower scoring students displayed all of the forms of monitoring less frequently than either high- or middle-scoring students. These students relied more on their memory of prior activities (e.g., *"I remember we did this in class"*) than any of the methods of monitoring described above.

Next, we considered the frequency (how many times?) and flexibility (how many different types?) of student monitoring (see Table 5). With respect to

Table 5

Average Frequency and Types of Monitoring by Score Level

Range of evidence scores	Number of students	Total number		Total types	
		Mean	*SD*	Mean	*SD*
16-24	5	9.2	7.6	3.2	0.8
8-15	18	6.2	6.2	2.3	1.2
0-7	13	1.7	2.6	1.0	1.4
Total	36	5.0	5.9	1.9	1.4

frequency, results indicate the groups differed in the number of monitoring statements, $F(2, 33) = 4.35$; $p < .05$. Pairwise comparisons indicated that on average, students who scored high made a significantly greater number of self-monitoring statements (mean = 9.2) than students who scored low (mean = 1.7; $p < .05$). Middle-scoring students (mean = 6.2) could not be distinguished from high- or low-scoring students. In addition to differences in frequency of monitoring, groups differed in the flexibility of their monitoring, $F(2, 33) = 7.03$; $p < .01$. High- and middle-scoring students were on average more flexible in their monitoring than were low-scoring students ($p < .05$). Low-scoring students did not display flexibility in their monitoring, averaging just one form of monitoring in the assessment. Our finding that high-scoring students in this study are characterized by frequent and flexible monitoring is consistent with prior research on competence in subject matter domains (Glaser, 1991).

Summary and Conclusions

Current forms of assessment call attention to the need for additional criteria to establish the validity of score use and interpretation—particularly, the quality and nature of the performance that emerges in an assessment situation. Research in cognition and learning has identified characteristics of proficient problem solvers and has developed protocol analysis techniques for examining these characteristics. If alternative assessments purport to measure problem solving in subject matter domains, then a correspondence should exist between assessment score and the quality of cognitive activity displayed by students.

In this study we examined the correspondence between student score and the quality of cognitive activity required to carry out a science performance

assessment. Student performance was evaluated with respect to four characteristics (explanation, plan, strategy, and self-monitoring skills) derived from cognitive studies of expertise and proficient performance. It was reasoned that if successful task completion is dependent on higher order thinking skills, then students who score high should provide coherent explanations of task-related concepts, generate an initial plan, engage in strategic problem solving, and effectively monitor their performance.

Results indicate that performance on the Mystery Powders assessment was very low. More than one half of the students scored 11 or less out of the 24 possible points. However, even with this distribution of performance, high- and low-scoring students displayed qualitatively different performance characteristics. High-scoring students, in general, (a) demonstrated in their explanations a generalized understanding of the principles underlying the unit; (b) provided an example of a test and corresponding observation when asked for an overall plan; (c) displayed a systematic approach to solving the problem by gathering all possible information before drawing conclusions; and (d) engaged often in effective and flexible monitoring of their performance by referring to their prior investigations (e.g., look at their journal) and operating within the constraints of the task (e.g., check list of options).

In contrast, the low-scoring students believed they could identify all substances with just one test. Their plans consisted of restating the problem or naming the equipment they would use. Their trial-and-error strategy was to do a test and see what happens. In monitoring their performance, they relied primarily on their memory of prior classroom activities or comparison of current observations regardless of their relevance.

The results of the analysis of the Mystery Powders assessment offers evidence to support inferences that this assessment elicits processes that are manifest in proficient problem solving. Moreover, the results also suggest that the methodology employed is viable for evaluating assessments of this kind (i.e., problem-solving tasks). This methodological approach was guided by a particular analytic framework derived from research on expertise and proficient performance. Evaluation of other genres of assessments should be guided by alternative frameworks or sets of cognitive characteristics apropos to the task demands. For example, in other forms of performance-based assessments, attention may focus on the extent to which a student uses knowledge to draw

inferences, to justify and explain a hypothesis, or to apply a principle in a novel situation. As we analyze the properties and objectives of performance-based assessments and match them to our understanding of the cognitive activities involved, we anticipate the development of additional frameworks for carrying out a cognitive analysis. The use of such frameworks should enable us to ascertain whether and how assessments are measuring higher order thinking and the cognitive capabilities that distinguish various levels of student achievement.

References

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*(12), 1210-1218.

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

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.

Everitt, B. S. (1977). *The analysis of contingency tables. Monographs on statistics and applied probability*. New York: Chapman and Hall.

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

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

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt Rinehart, & Winston.

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

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

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P. A. (in press). Rethinking validity: Themes and variation in current theory. *Educational Measurement: Issues and Practice.*

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.