Informing the Design of Performance Assessments Using a Content-Process Analysis of Two NAEP Science Tasks

CSE Technical Report 564

Kristin M. Bass
University of Michigan

Maria E. Magone
Learning Research and Development Center
University of Pittsburgh

Robert Glaser
CRESST / Learning Research and Development Center
University of Pittsburgh

July 2002

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

Project 1.1 Models-Based Assessment: Individual and Group Problem Solving—Theory into Practice
Robert Glaser, CRESST/Learning Research and Development Center, University of Pittsburgh, Project Director

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.

INFORMING THE DESIGN OF PERFORMANCE ASSESSMENTS USING A CONTENT-PROCESS ANALYSIS OF TWO NAEP SCIENCE TASKS[1]

Kristin M. Bass
University of Michigan

Maria E. Magone
Learning Research and Development Center
University of Pittsburgh

Robert Glaser
CRESST/Learning Research and Development Center
University of Pittsburgh

Abstract

Modern conceptions of knowledge have spurred efforts for assessments of thinking and reasoning and a principled approach to task design. This study uses a content-process framework to examine the cognitive demands of two science performance assessments in order to begin to articulate heuristics for assessment design. Think-aloud protocol techniques were used to examine the thinking and reasoning of fourth and eighth graders engaged in two separate National Assessment of Educational Progress (NAEP) hands-on science tasks about estimating the concentration of an unknown salt solution. The quality of observed performance and the task scores demonstrate that in accordance with the test developers' intentions, the performance assessments and scoring systems discriminated between students on the basis of their data collection and interpretation skills. However, performance also was influenced by details of item presentation (e.g., wording). These findings continue to make apparent the difficulty of item design and the necessity of creating rules that guide the development of performance assessments.

The design of assessments of thinking and reasoning in science has been more of an intuitive art than a systematic practice. Contemporary views of cognition imply that national and classroom assessment programs should be based on modern conceptions of knowledge with respect to applying organized, structured subject matter knowledge to problems and explanations. Such theoretical understanding has influenced the design of assessments in reading and mathematics (Glaser &

[1] We gratefully acknowledge Ben Sayler for his assistance with data collection and Corinne Zimmerman for her comments on an earlier version of this manuscript. A previous version of this paper was presented at the 2000 annual meeting of the American Educational Research Association, New Orleans, LA.

Silver, 1994). In science, however, this development is more limited. This paper applies a framework of cognitive activity and assessment characteristics to the evaluation of two science performance assessments used by the National Assessment of Educational Progress (NAEP). The results demonstrate how a cognitive framework can be used to inform and improve the design of assessment situations in science.

A particular development in assessment design has been the use of frameworks describing performance that facilitate the creation of items and the assessment of complex cognitive activity (Baxter & Glaser, 1998). This work emphasizes and displays the match of observed student performance with task goals and scoring systems. It entwines two strands: characteristics of competent cognitive performance and subject matter demands of assessments that permit or constrain cognitive activity. The nature of student cognition observed in a particular assessment situation can be described in terms of four general cognitive activities (e.g., Chi, Glaser, & Farr, 1988). Competent students display (a) coherent *problem representations* based on the underlying features of the topic they are studying; (b) organized, goal-oriented *strategies* which they can apply flexibly to a given problem; (c) a variety of techniques to *monitor* progress (e.g., problem recognition, rechecking work); and (d) *explanations* that demonstrate a deep understanding of the scientific principles driving their work.

Different knowledge requirements of assessment tasks affect the types of cognitive activity that may be observed. Science tasks can vary along two dimensions: content knowledge and process skills. Content knowledge demands fall on a continuum from rich to lean; process skills range from constrained to open. Together, the content and process dimensions form a 4-quadrant "content-process space" (Figure 1) that can describe the cognitive complexity of science performance assessments (Baxter & Glaser, 1998). For example, tasks that are content rich and process open (Quadrant 1) require problem representations and explanations that reflect a deep understanding of the concepts being studied, the generation of goal-directed strategies toward problem solution, and frequent, flexible monitoring. In contrast, tasks that are content lean and process constrained (Quadrant 3) involve fewer opportunities to observe the cognitive activities of problem solving. Because the content knowledge and procedures for task solution are given, students can proceed by reading and following directions and monitoring task completion. The content and process demands of a task thus establish expectations for the types of

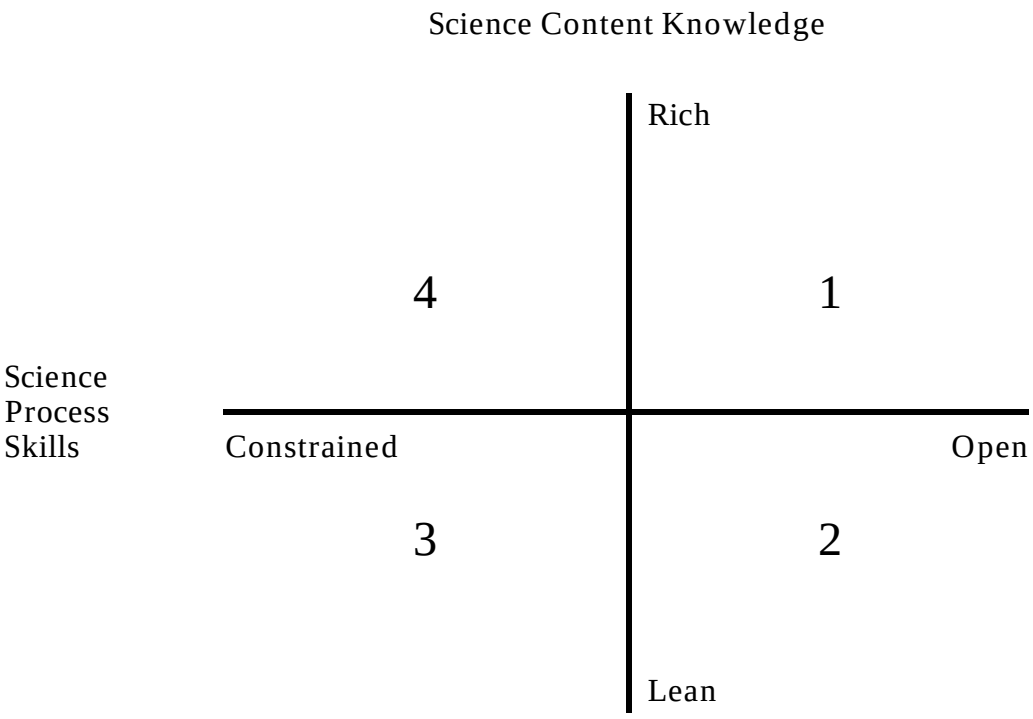

Figure 1. Content-process space.

cognitive activity likely to be observed. Analyses of cognition observed during a task can be used to confirm these expectations, expose particular features of items that facilitate or impede students' expression of their knowledge, and contribute to the foundation of a set of design heuristics for the construction of science assessments.

Consider the application of a content-process framework to the analysis of NAEP items. NAEP has been monitoring the academic abilities of 4th-, 8th- and 12th-grade students since 1969. Recently, the content of the biennial subject matter assessments has come into question as NAEP has been increasingly challenged to measure students' capacity to reason and solve problems in subject matter domains (National Academy of Education, 1997). Efforts to assess such competence in science have led to the use of hands-on performance tasks in which students conduct aspects of scientific investigations. It is assumed that performance assessments can provide students with opportunities to engage in cognitively complex activities such as representing problems, generating strategies, monitoring work, and reasoning

about information. A closer examination of these tasks, however, reveals that they do not always elicit the cognitive activities they were intended to measure. In particular, cognitive analyses of some large-scale performance assessments have identified mismatches between test developers' intentions and observations of student cognition (Baxter & Glaser, 1998).

Goals for assessing thinking and reasoning on science performance tasks in large-scale settings (such as NAEP) may be impeded on two fronts. The first concerns the demands of the assessment situation. NAEP's hands-on tasks, like those of many large-scale surveys, have been characterized as highly structured with a moderate to heavy reading load (National Research Council, 1999; Yepes-Baraya, 1997). Their design may preclude the open-ended nature of science problem solving and restrict opportunities for students to display subject matter understanding. A second problem involves the nature of the scoring system. Variations in student scores should reflect qualitative differences in cognitive performance. Conversely, if the task evokes uniform cognitive activity (a risk with NAEP's constrained assessments), all students should receive similar scores. However, when scoring criteria are based on superficial features of performance rather than evidence of thinking and reasoning, scores may overestimate the quality of cognition (Baxter & Glaser, 1998). Therefore, analyses of NAEP's science performance assessments require attention to the alignment between task goals, scores, and the quality of student cognitive activity. Toward that goal, frameworks such as those attending to content and process demands may provide an initial understanding of the effect of item characteristics on elicited performance and inform the development of future assessment situations.

This study examines the use of the content-process space for characterizing the cognitive demands of fourth- and eighth-grade NAEP hands-on science tasks and articulating heuristics for assessment design. Because questions are secure, only released items are considered. The two tasks discussed here involve estimating the concentration of an unknown salt solution. We first used think-aloud protocol techniques to observe the quality of cognitive activity that the tasks elicited. We then conducted quantitative and qualitative analyses to compare students' observed cognition to their task score. Results were used to ascertain whether the performance assessment and scoring system adequately differentiated students on the task's stated objectives (i.e., to measure students' ability to conduct an investigation and reason about data) and how the structure of items influenced observed performance.

Methods

Fourth-Grade Task

The fourth-grade Floating Pencil task is intended to measure students' ability to collect data (i.e., measure length and volume), make inferences, and apply their understanding to new situations. In the task, students are told that they can determine the difference between fresh water and salt water by doing a test. First, students are instructed to measure the length of a pencil weighted with a thumbtack (which serves as a hydrometer) floating above the surface of the water, in both fresh water and salt water. The pencil is marked with equally spaced letters from A (top of pencil) through J (bottom of pencil), and students are asked to observe where the water line comes to on the pencil and then place a mark on a picture of the pencil. Students are then directed to measure the length of the pencil that was above the water using a to-scale picture of a ruler. They repeat the Floating Pencil test to identify a "mystery water": They measure the length of pencil floating above the water in the mystery water and compare this finding with results from the previous tests. Throughout the task, students also are asked (a) whether the amount of water in the cylinder changes when the pencil is added; (b) how the way the pencil floats in salt water compares with how it floats in fresh water; (c) how dissolving more salt in the salt water would change the way the pencil floats; (d) how they can tell what the mystery water is; and (e) whether, when people are swimming, it is easier for them to stay afloat in the ocean or in a freshwater lake.

Eighth-Grade Task

The eighth-grade Salt Solutions Task is designed to assess students' ability to conduct a scientific investigation (e.g., making and applying simple observations, measuring length using a metric ruler), draw inferences from observations (e.g., identifying the concentration of an unknown salt solution), and explain the principles or goals directing their actions (e.g., why does the pencil float differently in salt water than in fresh). In the task, students read an introductory paragraph and follow step-by-step instructions to complete a scientific investigation. In the introduction, students are told that they are going to identify the concentration of an unknown salt solution by floating a pencil weighted with a thumbtack (which serves as a hydrometer) under three conditions: distilled water, a 25% salt solution, and the unknown salt solution. Students are instructed to measure the length of the pencil floating above the water in each solution two times. Next, students are directed to

plot on a graph the average amount the pencil floated in the distilled water and in the 25% salt solution and draw a line between the points. A graph is provided for this purpose, with numbered X and Y axes labeled “concentration” and “average length of the pencil above the water,” respectively.

Finally, students are asked, “Based on the graph that you plotted, what is the salt concentration of the unknown salt solution?” As they conduct the investigation, students follow instructions and respond to questions about data collection and graphing (e.g., “Use the ruler to measure the length of the pencil that was above the water. Record the length in Table 1 under Measurement 1.”), data interpretation (e.g., “Based on the graph that you plotted, what is the concentration of the unknown solution? Explain how you determined your answer.”), and the scientific principles underlying the experiment (e.g., “Explain why the pencil floats when it is placed in the water.”).

Content-Process Classification

Both tasks are classified as relatively content lean–process constrained. For example, students are given the content knowledge they need to draw conclusions (e.g., fresh water has very little salt) as well as a set of procedures for conducting the investigation. Students need some procedural knowledge to complete the investigation, such as how to measure with a ruler and (in the eighth grade) how to plot points on a graph.

Context and Students

The tasks were administered individually to 30 fourth-grade students and 27 eighth-grade students in an urban school district in southwestern Pennsylvania. The district is heterogeneous in terms of socioeconomic status and ethnicity. The sample contained a greater number of males than females and more Caucasians than African Americans (Table 1). Students were chosen by their teachers with the expressed purpose of ensuring a range of scientific abilities; interviewers were unaware of the teachers’ rankings.

Data Collection Procedure

Students were interviewed and audiotaped individually while they conducted the task. During the task, students were asked questions to simulate a think-aloud protocol. Interviewers asked students to elaborate on information they had just read or had provided on task questions they had just completed. Interviewers also noted

Table 1

Gender and Ethnicity of Participants

	Grade		
	4 (n = 30)	8 (n = 27)	Total (N = 57)
Gender			
Male	17	16	33
Female	13	11	24
Ethnicity			
Caucasian	18	17	35
African American	12	10	22

errors in strategies or procedures (e.g., incorrect reading of ruler) and student efforts to monitor progress or check understanding (e.g., re-reading directions).

Scoring

Students' task performance was assessed in two ways. First, students' written work was evaluated with the NAEP scoring standards. Next, students' verbal protocols, along with their written work, were scored using a cognitive framework designed for the purposes of this study.

NAEP Rubric Scoring

Trained NAEP personnel scored all student booklets. Based on the distribution of the total scores (which was bimodal for the fourth grade and normal for the eighth grade), students were then divided into two groups (high and low) at the fourth-grade level and three groups (high, middle, and low) at the eighth-grade level. Brief descriptions of the NAEP scoring criteria follow.

Fourth grade. NAEP scores the Floating Pencil task on the quality of student responses to two multiple-choice questions (1 point each) and on the quality of five short constructed responses (four of which are worth 3 points each, the fifth worth 5 points), and one extended constructed response (4 points). The maximum possible score is 23 points. Nineteen of the available points are given for the way the scientific investigation is conducted (i.e., the quality of the student's data collection and interpretation). The remaining 4 points are awarded for an understanding of the effect of salt on an object's ability to float.

Eighth grade. NAEP scores the Salt Solutions task on the basis of student responses to three multiple-choice questions (1 point each) and on the basis of four short constructed responses (3 points each), and three extended constructed responses (4 points each), for a maximum possible score of 27 points. Eighteen of the available points are given for the collection, graphing, and interpretation of data (i.e., scientific investigation skills). The remaining 9 points are assigned for an understanding of floating as the difference in the relative density of an object and that of the medium in which it floats.

Interview Scoring

Transcribed interviews were coded for evidence of four cognitive activities: (a) problem representation, (b) strategies, (c) monitoring, and (d) explanations. Monitoring behaviors were classified into several subcategories (e.g., problem recognition, checking data) and coded for the total number of behaviors and types. All other cognitive activities were rated as three- or four-level ordinal categories (e.g., complete, partial, or inadequate data collection strategies based on the precision and consistency of students' measurement and recording; see Appendixes A and B). More detailed analyses can be found in Bass (1999) and Magone (1999).

Data Analysis

Kruskal's gamma, a measure of association (Everitt, 1992) was calculated to describe the relationship between score level (high and low for the fourth grade; high, middle, and low for the eighth grade) and quality of problem representations, strategies, and explanations (e.g., complete, partial, and inadequate strategies). T tests and analyses of variance were used to compare the frequency of monitoring behaviors by score level. Qualitative patterns or trends in students' cognition at each score level (e.g., students identify methods but not the goal of the task; Miles & Huberman, 1994) were identified as a way to further characterize student performances.

Results and Discussion

This section describes our quantitative and qualitative findings. First, the associations between students' task score and quality of their observed cognitive activity (i.e., problem recognition, strategies, monitoring, and explanations) are described for both tasks. Then, representative case studies are provided to illustrate

additional data trends. The section concludes with a discussion of the effect of task characteristics on students' observed performance.

NAEP Rubric Scores

On the fourth-grade Floating Pencil task, students received an average of 15.3 points out of a possible 23 (*SD* = 2.7). Scores displayed a bimodal distribution and were divided into two groups (see Table 2): high and low. On the eighth-grade Salt Solutions task, students received an average of 16.6 points out of a possible 27 (*SD* = 3.5). Scores were divided into three evenly sized groups (see Table 2): high, middle, and low.

Relationship Between Task Goals, Scores and Observed Cognitive Activity

Recall that the Floating Pencil and Salt Solutions tasks were designed to measure students' ability to collect and draw inferences from data and (in the fourth grade only) apply their findings to an everyday situation. Observations of student performance generally confirmed these claims. Results indicated that the strongest associations between score level and quality of cognitive activity were for strategies (i.e., ability to collect and record data) and the explanations of the identification of the unknown salt solution. Findings are summarized in Table 3 and discussed in the subsequent text.

Table 2

Range of Fourth- and Eighth-Grade Total NAEP Scores

Group	Range of scores	Number of students	Percentage of students	Mean	*SD*
		Grade 4			
High	16-21	16	53.0	17.4	1.3
Low	9-15	14	47.0	12.9	1.6
Total	9-21	30	100.0	15.3	2.7
		Grade 8			
High	19-23	9	33.3	20.6	1.2
Middle	15-18	9	33.3	16.8	1.1
Low	11-14	9	33.3	12.6	1.0
Total	11-23	27	100.0	16.6	3.5

Note. Maximum fourth-grade score = 23; maximum eighth-grade score = 27.

Table 3

Relationships Between Score Level and Quality of Cognitive Activity by Grade

Cognitive activity	Fourth grade	Eighth grade
Problem representation	Trend: high-scoring students gave a plan for identifying the mystery water, $\chi^2(1) = 3.04$, $p = 0.08$	$\gamma = .41$
Strategies	$\gamma = .99$	$\gamma = .72$
Monitoring		
Frequency (total number of behaviors)	No variation by score level: $t(28) = .37$, $p > .05$	No variation by score level: $F(2, 24) = .41$, $p > .05$
Flexibility (total types of behaviors)	No variation by score level: $t(28) = -.93$, $p > .05$	No variation by score level: $F(2, 24) = .48$, $p > .05$
Explanations		
Identification of unknown	More high-scoring students than low identified the water correctly, $\chi^2(1) = 6.10$, $p < .05$	$\gamma = .83$
Why objects float	No difference in types of reasons by score level	$\gamma = .20$
Conservation of water volume	$\gamma = .13$	—
Easier to float in an ocean or lake	$\gamma = .43$	—

Problem Representation

On both tasks, the highest scoring students' representations changed from shallow (inadequate) to principled (complete) as students progressed through the assessment. Low-scoring students expressed less complete understanding of what they were supposed to do and how they were going to do it. For example, early in the fourth-grade assessment, nearly all students' goals focused on carrying out procedures, on making or seeing the pencil float, and on distinguishing between the two waters (e.g., determining which water "floats best"). Chi-squares (with Yates corrections because there was only one degree of freedom) were performed on individual response categories because students could give more than one answer. There were no differences between the representations of high-scoring and low-scoring groups; procedures, $\chi^2(1) = .45$, $p > .05$; make / see pencil float, $\chi^2(1) = .00$, $p > .05$; distinguish between waters, $\chi^2(1) = .00$, $p > .05$. Few high- or low-scoring students mentioned goals concerning the mystery water because at that point they had been given little indication of the role of that water in the task.

Later in the task, students' representations became less superficial, with the high scorers generating the most complete understanding of the problem. After students read that they were to identify a mystery water, most students understood what they were supposed to do (0.9 and 0.8 of the high- and low-scoring students, respectively), $\chi^2(1) = .08$, $p > .05$, but the high-scoring students had a better representation of how they were going to do it. Specifically, a trend was found in which more of the high-scoring group (0.5) than the low-scoring group (0.1) said they would be comparing and contrasting results from the mystery water with results from previous tests, $\chi^2(1) = 3.04$, $p = 0.08$.

Similarly, the quality of eighth-grade students' problem representations was weakly related to their total score ($\gamma = .41$). On beginning the task, students expressed a vague understanding of the goal of the task and could not articulate the procedures they would use to carry it out. Compared to the lower scoring students, a higher proportion of high-scoring students correctly mentioned that the goal of the task was to identify the concentration of the unknown salt solution (67% of the high-scoring students versus 33% of the middle- and low-scoring students); this statement was provided in the task instructions. A common misconception among the low-scoring students was that the activity's purpose was to determine which solution "floats the pencil the best." Higher scoring students' descriptions became more complete and coherent over time as they collected data under the experimental conditions. In contrast, low-scoring students' conceptions of the problem remained unchanged by their experiences in the assessment. A later section of the paper will elaborate on these differences.

Strategies

There was a strong relationship between quality of strategy (i.e., precision of data collection and / or graphing) and score level for both tasks ($\gamma_{4th} = .99$, $\gamma_{8th} = .72$). The highest scoring students followed the written instructions closely. Any errors made were minor and included slight numerical inaccuracies in recording data. In contrast, the lower scoring groups made one or more significant errors that invalidated the data they had collected. For example, students would test the wrong water or use inconsistent units of measurement (e.g., centimeters for the first measurement and inches for the second). Further, low-scoring eighth-grade students made a bar graph instead of a line graph when prompted to plot their findings. This became a problem when students later were asked to identify the concentration of

the unknown salt solution based on the graph they had plotted but did not have a viable graph to interpret.

Monitoring

Students' monitoring behaviors focused on completion of instructions (e.g., rereading text, recognizing problems with procedures) and were more commonly observed among eighth graders than among fourth graders. Fourth graders exhibited five different types of monitoring behaviors and eighth graders displayed six types (see Appendixes A and B). Fourth graders monitored their work an average of 1.2 times (SD = 1.1), using 0.9 types of behaviors (SD = 0.7). Eighth graders monitored their performance an average of 12.8 times (SD = 6.0), using 3.9 different types of behaviors (SD = 1.3). The frequency (total number of monitoring behaviors regardless of type) and flexibility (number of types of behaviors) of monitoring did not vary by score group for either grade level; fourth grade: $t_{\text{frequency}}$ (28) = .37, $p > .05$, $t_{\text{flexibility}}$ (28) = -.93, $p > .05$; eighth grade: $F_{\text{frequency}}$ (2, 24) = .41, $p > .05$, $F_{\text{flexibility}}$ (2, 24) = .48, $p > .05$.

Frequency analyses of individual monitoring behaviors revealed one difference between high- and low-scoring fourth-grade students: The low-scoring group showed a trend of more evidence of problem recognition behaviors (e.g., statements such as "I think I did something wrong" or corrections / adjustments) than the high-scoring group, Yates-corrected $\chi^2(1) = 3.23$, $p = .07$. This was probably because of the low scorers' sensitivity to their performance and the problems they had in carrying out the steps in the assessment.

Explanations

Explanation questions at both the fourth- and eighth-grade levels explored students' ability to identify the unknown salt solution (one question at each level) and their understanding of floating (three questions in fourth grade, five in eighth grade; see Appendixes A and B). Additionally, fourth-grade students were asked about the conservation of water volume (i.e., "Does the amount of water change when you add the pencil?") and the application of understanding to an everyday situation ("When people are swimming, is it easier for them to stay afloat in the ocean or in a freshwater lake? Explain your answer.").

At both grade levels, scores reflected students' ability to make sense of data. In the fourth grade, high-scoring students were more likely to correctly identify the mystery water as being fresh water and give a clear and complete justification for

their answer based on data from their study (0.7 of the high-scoring students versus 0.4 of the low-scoring students), Yates-corrected $\chi^2(1) = 6.10$, $p < .05$. Further, a trend was found in which that high-scoring students gave a clear and complete justification for their answer basing their justification on data from the study ($\gamma = .48$).

There was an even stronger association between water identification and task score for the eighth-grade students ($\gamma = .83$). The highest scoring students used their graph to logically estimate the unknown concentration. Middle- and low-scoring students were more likely to return to their original data table to identify the unknown concentration rather than use the graph for this purpose. They used faulty proportional reasoning to generate their answer; that is, students considered how much less the pencil floated in the unknown solution compared to the 25% salt solution only. They did this instead of comparing the unknown to both the distilled water and the 25% salt solution, which would have provided a more precise answer.

Though scores were associated with students' ability to interpret data, they were unrelated to a general understanding of why objects float. For example, fourth-grade students most commonly believed that the pencil floated because of the pencil's weight or other physical features (e.g., thumbtack, made of wood). Chi-squares (with Yates corrections because there was only one degree of freedom) were performed on individual response categories because students could give more than one answer. Analyses indicated that there were no differences between high-scoring and low-scoring groups' reasons for why the pencil floats; pencil weight, $\chi^2(1) = 1.66$, $p > .05$; physical features of the pencil, $\chi^2(1) = .34$, $p > .05$; water pushing up, $\chi^2(1) = .00$, $p > .05$; physical features of the water, $\chi^2(1) = .00$, $p > .05$; relative weight of the water, $\chi^2(1) = 2.34$, $p > .05$; gravity, $\chi^2(1) = .00$, $p > .05$; other, $\chi^2(1) = .00$, $p > .05$.

Further, there was no difference between student groups on their responses to two multiple-choice questions about how the pencil floats in the salt water compared to the fresh water (Question 5), and how the pencil would float if more salt were added (Question 8). Question wording proved difficult for students to understand, especially option 8c, "The pencil would float *lower* than it did before the salt was added." A substantial proportion of the students (0.3) chose this incorrect option even though in their explanations they correctly described the effects being assessed by the item (e.g., "*If you take the more extra salt you take out—the more salt you take out, then your pencil will go down more.*").

The quality of eighth-grade students' explanations of floating also was unrelated to their total score ($\gamma = .20$), a relationship that belies some minor differences. When asked why the pencil floats when it is placed in the water, the majority of high- and middle-scoring students generated internally consistent but scientifically inaccurate theories from their data (e.g., *"The pencil floats more in salt water because there are more molecules to push it up."*). In contrast, low-scoring students offered vague, atheoretical explanations of floating (e.g., *"Salt just makes things float more."*).

Fourth-grade students also were asked questions about the conservation of water volume and the application of their findings to an everyday situation. The question about the conservation of water volume (i.e., "Does the amount of water change when you add the pencil?") confused all students equally. Quality of explanation was unrelated to score level ($\gamma = .13$). There was a stronger relationship between explanation and score level for the application item. Students were asked whether it was easier to stay afloat in a freshwater lake or in the ocean and to explain their answers. As with the question about the identification of the mystery water, here, too, a trend was found in which high-scoring students gave more complete and correct answers than did low-scoring students. Explanations given by the high-scoring rubric group were generally more clear and complete than the explanations of the low-scoring groups ($\gamma = .43$). Low-scoring students were somewhat more likely to answer incorrectly, sometimes basing their conclusions on earlier flawed procedures, sometimes bringing in information that was irrelevant or extraneous to the assessment. As will be discussed later, ambiguity in the question wording may have compromised the effectiveness of the item to elicit what students know.

Summary

Overall, the fourth-grade Floating Pencil and eighth-grade Salt Solutions tasks most strongly differentiated students on their strategies (i.e., data collection and/or graphing) and explanations or inferences made about their data. This is in keeping with the tasks' goals. However, cognitive activities such as problem representation and monitoring proved to be less directly related to score, a fact that may be attributable to the tasks' content lean–process constrained nature. Given that the task guided students through a step-by-step series of directions, opportunities to represent problems and monitor performance were minimal.

Case Studies of Student Performance

Quantitative analyses of performances on the Floating Pencil and Salt Solutions tasks highlight their emphasis on the science process skills of data collection and interpretation. Additional information about overall performance and problem solving can be gathered from case studies. The tasks' structures created situations in which performance was enhanced or inhibited by interpretation of written directions. For both the fourth- and eighth-grade tasks, scores reflected proficiency in integrating aspects of the assessment directions and procedures into a cohesive problem representation. Further, on the fourth-grade assessment, the wording of questions impeded students' ability to express what they actually knew. These trends are illustrated below with case studies of typical student performances. For fourth and eighth grade, each section begins with a general description of performance followed by examples of specific students at the various score levels.

Fourth Grade

The highest scoring fourth graders displayed detailed problem representations, meticulous procedures, and occasional monitoring and resolution of problematic questions. In contrast, low-scoring students frequently used "bungled" procedures and maintained fragmented understandings of the task, more often bringing outside information, such as chlorine or gravity, into their explanations. The wording of some questions seemed to confuse or mislead both high-scoring and low-scoring students, although problems were more pronounced for the low scorers. Consequently, some of the items may not have elicited the relevant information students knew. Three case studies illustrate these points. Julie,[2] a high-scoring student, will be described first, followed by Zeke and David, two low-scoring students whose case studies illustrate various problems many low-scoring students had with carrying out procedures and understanding task wording.

Julie, representing high-scoring students: Coherent problem representation and explanations and accurate procedures. Julie's performance on the Floating Pencil task is typical of high-scoring students. Her rubric score of 21 was the highest score for this sample of fourth-grade students.

Julie's problem representation was well aligned with the task. At the start of the experiment, she read the instructions carefully and determined that her goal was

[2] All students' names are pseudonyms to preserve anonymity.

"to see if [the pencil] *floats in the mystery water, the salt water, the fresh water."* Later in the experiment, Julie read that she was supposed to perform the "Floating Pencil test" using the mystery water to find out if the water was fresh or salt. In line with the additional information she was given, Julie expanded her problem representation to articulate the goal of the task and a plan for reaching that goal:

> Well, what you're going to do is to find out whether this water is salt or fresh. And how you're gonna do that is by seeing how much it floats. If it floats like how the salt water—this bottle labeled salt water does—then it's gonna be salt water. If it floats like the way the fresh water did, it's gonna be fresh water.

Julie executed procedures carefully and had no difficulty following instructions. Other than the general care with which she executed procedures, Julie showed few signs of self-monitoring, perhaps because the experiment was fairly easy for her to execute. As will be described later, Julie did re-read and re-think a question she did not initially understand.

Julie's performance also was typical of the high-scoring students with regard to the explanations given for the various items. On Question 2, which asked how much water was in the cylinder after the pencil was added, students were given three choices: "(A) More water than before the pencil was added; (B) The same amount of water as before the pencil was added; (C) Less water than before the pencil was added." The task also prompted students to explain why they chose their answer. Julie chose option A, and for her explanation wrote "*I think so because the pencil makes the line go up by its capacity.*" Although Julie indicated that there was more water after the pencil was added, an incorrect answer, she seemed to be attributing this increase to the volume of the pencil (the pencil's capacity), rather than to a misunderstanding of the conservation of the water volume. Nearly all students, high scoring and low scoring, chose the wrong multiple-choice option for this question.

In contrast, Julie answered questions about the floating pencil correctly. Question 5 asked how the pencil floats in salt water compared with how it floats in fresh water and gave students four choices: "(A) In the salt water, the entire pencil sinks below the water surface; (B) In the salt water, more of the pencil is below the water than before; (C) In the salt water, more of the pencil is above the water than before; (D) In the salt water, the same amount of the pencil is above the water as in the fresh water." Julie chose option C, and her explanation was based on accurate observations. Recall that the pencil was marked with equally spaced letters from A

(top of pencil) to J (bottom) and that the higher the pencil floated, the further in the alphabet the pencil touched the surface of the water. Julie referred to those letters when justifying her answer to Question 5: *"Before I noticed that A was where it was in fresh water and now it's like maybe C—about—so that's a pretty big difference."*

Question 8 asked "If you dissolved more salt in the salt water, how would it change the way the pencil floats?" with responses "(A) The pencil would float *higher* than it did before the extra salt was added; (B) The pencil would float at the *same level* as it did before the extra salt was added; (C) The pencil would float *lower* than it did before the extra salt was added." Julie said, *"I don't like the way they're worded,"* but chose the correct answer, A. After selecting the correct answer, Julie re-read the question, trying to re-think the item's wording. She said, *"The pencil floats here—so this means that more water should be like more letters—'A' means more letters would be above."* When asked what was confusing, she said, *"I think I just read a little too quickly and it didn't make sense . . . I thought it was saying that for 'A' it was going to be probably lower."*

Both high- and low-scoring students were confused by the wording of Question 8, and answered it incorrectly. Some of the students seemed to be reading the item as though it said "The pencil would float higher before the extra salt was added" rather than—as the item stated—"The pencil would float higher *than it did* before the extra salt was added" (italics added by authors); that is, that the pencil would float higher after salt was added. Julie may have been experiencing the same confusion. She was, however, able to articulate and think through her confusion to arrive at the correct answer.

When asked to identify the mystery water in Question 10, Julie quickly wrote "fresh water." For her explanation she wrote, *"by the pencil leanining trowards the side, unlike salt water, but the main reason was that less of the pencil was sticking out"* (sic). The experimenter did not notice that the leaning of the pencil had any relationship to the salt content of the water; however, Julie seemed to observe so carefully that it seems very possible that her observation about the pencil's leaning had some validity.

Finally, in Question 11, when asked whether it was easier to stay afloat in the ocean or in the freshwater lake, Julie wrote "ocean" because *"like the pencil the salt helps them stay aflot more than in fresh water"* (sic). She said that she did not have much experience swimming in the ocean, but based her answer on the experiment with the pencil.

In sum, Julie's representations and explanations were thoughtful, lucid, and well aligned with the task. Her procedures were accurate and supported the overall coherence of her thinking.

Zeke, representing low-scoring students: Inaccurate, bungled procedures and fragmented explanations. Zeke's performance was characterized by inaccurate procedures and fragmented understanding of the concepts being studied. Zeke (rubric score = 14) developed an accurate representation of the task (e.g., *"put the mystery water in to see if it is fresh water or salt"*) corresponding to the information he was presented. He articulated a plan, *"When we put the pencil in —if it goes to B or C."* That is, he said he would observe whether the surface of the mystery water touched the letter B on the pencil, as it did for fresh water, or touched the C on the pencil, as it did for salt water. However, he erred in his execution of procedures. When Zeke had to measure the length of the pencil floating above the surface of the fresh water, he held the pencil up backwards to the ruler, thus recording the length of the pencil that was *below* the water, rather than above.

Zeke also had difficulty pouring and measuring the water. For example, Zeke began the experiment by measuring fresh water and then discarding it into a plastic dish. Next, he was to pour salt water into the cylinder up to a red line. Zeke poured too much salt water into the cylinder and then poured some of the excess into the plastic dish. At this point, the water level was too low, so he added some water from the dish of discarded water, thus mixing his salt water with fresh water and contaminating his sample. He still ended up with too much water in the cylinder.

Zeke's explanations were fragmented and inconsistent. The Floating Pencil task had students perform an experiment, but gave them little explanation for the effects they observed. Zeke seemed to try to invent his own model of floating, relating the phenomena observed to gravity. He first theorized that gravity affected floating when asked Question 8, "If you dissolved more salt in the salt water, how would this change the way that the pencil floats?" Zeke correctly chose option A (that the pencil would float higher than it did before the extra salt was added), which he simply attributed to the "gravity" of the salt water. He later contradicted himself when he mentioned that people float better in a freshwater lake than in the ocean. His justification was *"because the salt water puts more water on the object and fresh water has little gravity of the water,"* and that heavy objects like people and boats can float because *"Sometimes the gravity pulls them up on the surface of the water."* Gravity thus became an all-inclusive theory for opposing predictions and observations of floating.

Thus Zeke's bungled procedures seemed to lead to inaccurate data as well as to fragmented and inconsistent explanations. The incorporation of outside information (gravity) may have been an attempt to make sense of his ambiguous findings.

David, representing low-scoring students: Procedural errors and problems with question interpretation. David (rubric score = 13) highlights the problems some students had with task questions and procedures. First, David had trouble understanding the purpose of the activity. At the start of the task, when David was questioned about his problem representation, he said he did not know what he was going to be doing. When questioned later in the experiment, David said he would pour the water up to the red line on the cylinder. The task representations that David articulated seemed to be at a procedural level.

Inaccuracy in data recording also affected David's performance on the task. Recall that the pencil was marked with letters from A to J and that students were first told to mark the height of the pencil above the water on a picture of the pencil. Next they marked a second picture where a ruler was lined up against the pencil to show the height of the pencil that floated above the water. Early in the experiment, the pencil in fresh water floated to the letter A. David had mistakenly marked the first picture of the pencil at B rather than at A. He marked the second picture correctly, although the first picture is what he remembered. David's measurements and recordings for the salt water were accurate and indicated that the pencil floated to C. David identified the mystery water as being both fresh and salt because the pencil floated to A rather than to B (as he mistakenly remembered the fresh water) or to C (as he correctly remembered the salt water). David explained, *"I'm thinking it is like fresh water and salt water … because it is not a 'B' or 'C' when it was fresh water or salt water. … It's still a mystery water."*

David had mixed success with other questions eliciting explanations. Question 2 was a multiple-choice item that asked "How much water is in the cylinder now that you have put the pencil in it?" David indicated that there was "more water than before the pencil was added" because *"The water went over the red mark that I made."* It is not clear from this answer that David thought there was greater water volume instead of just a higher level of water. Although the item seems to have been intended to tap students' understanding of conservation of water volume, it may not have accessed what students knew about the topic.

Later explanations illustrated David's difficulties with item wording. On Question 8 ("If you dissolved more salt in water, how would it change the way the pencil floats?"), David incorrectly chose option B, "The pencil would float at the same level that it did before the salt was added." David said, *"If you was to take the salt, some salt out of it. It'll stay the same as it did before the extra salt was added."* Considered independently from the question stem, David's explanation related to option B is correct: The pencil will float the same before and after the combined action of adding salt and taking it out. This option is not, however, the correct answer. David may have had difficulty considering the options in conjunction with the stem, and picked an option that he could explain regardless of its relation to the question.

On Question 11 ("Is it easier to float in the lake or the ocean?"), David said that it is easier for people to stay afloat in fresh water *"because if salt gets in your lungs bad things will happen."* Later in the interview, however, when David was asked if he learned anything, he said that *"salt water can make you float higher than if you, if you would, if you was going swimming than if you would in salt water, I mean a fresh water*." David may have known the information to answer Question 11 correctly but may not have used it because of the way this item was worded.

Question 5 (how the pencil floats when salt is added) seemed to be easier for students to understand. David correctly answered this question, choosing option C, "In the salt water more of the pencil is above the water than before." After answering the question, his explanation for the occurrence was *"Because the salt water is like making it rise up with the fresh water . . . the chlorine is making it come down."* David was generating theories to explain what he had observed, even going beyond what was explicitly presented in the task.

In sum, David seemed to learn from the study that an object floats higher in salt water than in fresh water. However, the error he made in recording his results as well as his difficulty interpreting questions led to a low score on the task.

Eighth Grade

As noted in the quantitative section, eighth graders' performance depended mostly on the precision of their data collection and graphing, a fact reflecting the relative emphasis in the scoring system on "scientific investigation" skills. The case studies make it clear that the highest scoring students also possessed other skills that may have contributed to their competence at understanding the instructions and

performing well on the task. In particular, score levels reflected students' varying ability to coordinate individual steps of the investigation into a more complete representation of the tasks' goals and methods. Three case studies illustrate this point: (a) Lynn, a high-scoring student, (b) Tim, a middle scorer, and (c) Pete, one of the lowest scoring students.

Lynn, representing high-scoring students: Progressively principled problem representation. The highest scoring students gradually incorporated information from the task into their original problem representation. Lynn's performance is typical of the high-scoring students. At the start of the activity, Lynn (rubric score = 21) read the task instructions and correctly determined that her goal was to identify the concentration of the unknown salt solution. However, she could not describe exactly how she would do this:

> I'm going to see like how much the pencil floats in the water and then in the salt solution and this one has like 25% salt solution and this one has like I guess hardly any so I'm gonna guess how much is in there [points to unknown] by observations.

Once Lynn was familiar with the floating pencil procedure, she realized how it would help her accomplish the task's goal. After she was instructed to observe the pencil floating in the 25% salt solution, she was again asked what she was going to do. She responded with a general mathematical strategy for estimating the unknown salt concentration from the known concentrations and measurements.

> See the difference between these [distilled water, 25% salt solution] so you can guess that [unknown salt solution] . . . 'cause like if it's greater than half of a centimeter then it has more salt than the distilled water. And if it's greater than whatever this one [25% salt] will be, it'll have more of the 25% salt water. But if it's in-between, it has between none and 25%.

Nonetheless, Lynn did not later recognize that graphing her findings involved mathematical comparisons similar to the ones she had already described. By reading the instructions, Lynn understood what she was supposed to do (plot the amount the pencil floated above the water surface for the distilled water and 25% salt solution and draw a line between the two points) and what the graph would represent (the relationship between concentration and floating). However, she only acknowledged the general advantage of graphing—"*So you can like see the differences better*"—and not its specific usefulness for identifying the unknown salt concentration.

Tim, representing middle-scoring students: Errors in goal representation. In contrast to the high scorers, middle-scoring students had some problems comprehending text. Students had particular difficulty understanding the goal of the assessment. Consider Tim, a middle-scoring student (rubric score = 18). From start to finish, Tim had difficulty identifying the assessment's goal (despite reading a paragraph that stated this explicitly) and then relating that goal to the procedures for collecting, graphing, and interpreting data. When asked initially what he was going to do, he alluded to the surface features of the problem: "*Mainly putting in a cylinder, putting in water with salt in it to see, like in the Dead Sea, if it'll float like if there's a lot or less it'll float or sink.*" Later, he believed he was collecting data to see which solution "*would be best to keep* [the pencil] *above.*" Tim could always describe the procedures he was doing (e.g., floating the pencil, making a graph), yet he consistently misrepresented the overall purpose of the investigation. His overall problem representation was that of a collection of procedures with no consistent goal uniting them.

Pete, representing low-scoring students: Fragmented, inaccurate problem representation. The lowest scoring students did not substantially improve their problem representations from the beginning to the end of the assessment. Pete (rubric score = 14) represents a typical low-scoring student. After reading the introductory task description, Pete misunderstood what he would be doing and why he would be doing so: "*testing how much salt and maybe how much salt prob—, let's see, how much salt will be in fresh water and how much salt will be in ocean water. Then try to find out how much salt does it take for the pencil to float and how much for it to just stay where it is.*" He did not understand that he was supposed to identify the concentration of an unknown salt solution, nor did he apparently realize that the salt content of the fresh and salt water already had been given to him.

Further, while the middle-scoring students were able to identify the procedures they were doing, the low-scoring students misrepresented their methods. For example, when Pete graphed his findings, he said he was plotting three lines to see "*how the pencil floats in each kind of water.*" His answer was inaccurate (he was only supposed to plot two values), and his reason for making the graph had nothing to do with the goal he originally had articulated. His understanding of the task was faulty and mirrored the low-scoring students' inconsistent, mistake-ridden data collection and graphing described in the quantitative findings.

Effects of Task Characteristics on Performance

The quantitative analyses and case studies make clear that the content lean–process constrained nature of the tasks restricted opportunities to observe particular cognitive activities of problem solving. Because students were given the information and procedures needed to complete the task, emphasis was placed on reading and following directions and monitoring completion of instructions. In this context, we suggest that the presentation of the items (e.g., wording) had a major influence on students' performance (cf. Goldberg & Kapinus, 1993).

Results show that fourth graders were confused by items eliciting explanations, and consequently, the questions underestimated their knowledge. For example, Question 2 (the water volume after the pencil was added) did not seem to elicit students' understanding of conservation of water volume as it was intended to do, and nearly all students answered this item incorrectly. Students may not have understood that the change in the amount of water when the pencil was added was referring to volume of the water rather than level of water in the cylinder.

In other cases, the grammatical structure of the questions caused problems. Consider Question 8, which asked "If you dissolved more salt in the salt water, how would this change the way the pencil floats?" The correct option, "A. The pencil would float *higher* than it did before the extra salt was added," was often difficult for students to interpret, a point that was made earlier in the case studies.

Even a single word could influence student responses. Question 11 (which asked whether it is easier to float in fresh water or salt water) prompted students to use information that was not given to them in the task. For example, students answered that it was easier to stay afloat in the lake because the ocean has waves or sharks, or because salt can get into your eyes, ears, and lungs. In the case studies, students like David said that objects float higher in salt water than fresh water but did not apply that understanding to the ocean/freshwater lake question. Although students had the knowledge needed to answer the question, they were distracted by the term "easier" and therefore did not make the connection between the salt water in their experiment and the salt water in the ocean. Given students' interpretations of the word "easier," the item might have been improved by asking students if they would float *higher* in an ocean than in a lake. Such wording is more closely related to the Floating Pencil experiment and could presumably cue students to think about their data to answer the question.

Unlike the fourth graders, eighth-grade students were not confused by the wording of individual questions. Rather, their performance was affected by the interpretation of directions and subsequent attention to detail. Recall that in the assessment, students first read a general overview of what they would do (e.g., float a pencil in three different kinds of water to estimate the concentration of an unknown) and were gradually given details (e.g., procedures for collecting data and graphing results) as the task progressed. The highest scoring students understood the directions so well that they were able to conduct the experiment precisely and create principled models of the problem. Middle and low scorers, on the other hand, had difficulty following directions and generating coherent task representations.

We speculate that one or more aspects of task presentation may have affected the eighth graders' performance. First, the reading load may have influenced comprehension and execution of procedures. The low-scoring students in particular had difficulty understanding what they were to do and translating what they had read into effective data collection methods. Further, the step-by-step order of the directions may have contributed to differences in observed performance. The highest scoring students clearly were working with more information than they had taken from each individual step by virtue of their ability to synthesize those steps into a larger whole. In contrast, the middle- and low-scoring students were less proficient at reading and interpreting the instructions and had no cues within the text to give them a broader picture of what they were doing and why. Efforts are needed to provide students with this information while at the same time reducing reading demands as much as possible. At both grade levels, alternate ways of delivering information to students (e.g., through comprehensive verbal instructions as well as written instructions) and simplified linguistic features of questions (cf., Abedi & Lord, 2001) should be considered in order to diminish the confounds of reading comprehension and scientific investigation skills.

In sum, task presentation appears to be a highly influential design characteristic for the Floating Pencil and Salt Solutions tasks. When students were given precise information about what they would do, how they would do it, and how they would interpret and apply their findings, the nuances of language and order of information (i.e., step-by-step versus all-at-once) proved critical to student performance.

Conclusions

The development of performance assessments in science can benefit from frameworks that set expectations for cognitive activity and specify item design. This study considered the content and process knowledge requirements of tasks as the basis for an exploration of the cognitive demands of two NAEP science performance assessments. The tasks were classified as content lean–process constrained and consequently limited opportunities to observe problem solving. Nevertheless, in accordance with test developers' intentions, the assessments mainly differentiated between students on the basis of simple data collection and interpretation skills.

In the course of the analysis, other task characteristics that influence performance became apparent. Specifically, item presentation (e.g., wording, reading load, step-by-step instructions) appears to have affected the nature of student performance within the content lean–process constrained quadrant. Fourth graders had some difficulty understanding the questions requiring explanations and wrote answers that conflicted with their verbal explanations. Moreover, on both tasks, high-scoring students tended to coordinate the individual steps of the investigation. This suggests that high-scoring students were working with more information than were their low-scoring peers and that the reading load and step-by-step order of instructions may have circumvented efforts to provide students with the knowledge needed to conduct the investigation. This effect is particularly relevant in considering the design of highly constrained, text-heavy assessments such as those examined here.

These findings imply that frameworks that contribute to item development and analysis can be empirically studied for improving test administration. The future of effective assessment design will especially require the exploration of task characteristics salient to the other three quadrants of the content-process space. An important concurrent endeavor is empirical work that demonstrates the impact of changing design features, such as task presentation, on the quality of performance observed in assessment situations. The outcome of these efforts will improve the foundation for assessment design that is grounded in cognitive theory and its applications in instructional and large-scale settings.

Finally, this work speaks to the challenges inherent in designing assessments to elicit thinking and reasoning. While there are numerous psychometric techniques for analyzing assessments once they are implemented, there are considerably fewer

procedures available to guide task development. Assessment design is often an intuitive, time-consuming, trial-and-error process in which the most minor of details (such as task wording) can have a significant impact on the quality of information gathered about student learning. Unfortunately, the growing urgency of school accountability and pressure to monitor achievement and show improvements in student learning over a short time span can lead to expedited task development which may jeopardize the quality of assessment design. Policymakers must be sensitive to the complex nature of assessment development and allocate adequate time and resources to these efforts.

Overall, the design of innovative measures of thinking and reasoning must be grounded in the research linking cognitive theory to educational assessment (Pellegrino, Baxter, & Glaser, 1999). Efforts to specify rules or heuristics for assessing cognitively complex aspects of performance should increasingly inform the creation of assessment situations. The decision to use performance assessments in various subject matters must arise from frameworks advocating the measurement of active knowledge and be supported by extensive field testing confirming assessments' cognitive claims. Early, well-studied development efforts can maximize the potential of science performance assessments for facilitating discovery and reasoning, and thus bring into reality the educational progress that NAEP and others are trying to assess.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219-234.

Bass, K. M. (1999*). Quantitative analysis of the NAEP Salt Solutions task.* Unpublished manuscript, University of Michigan.

Baxter, G. P., & 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.

Everitt, B. S. (1992). *The analysis of contingency tables.* New York: Chapman and Hall.

Glaser, R., & Silver, E. (1994). Assessment, testing and instruction: Retrospect and prospect. *Review of Research in Education, 20*, 393-419.

Goldberg, G. L., & Kapinus, B. (1993). Problematic responses to reading performance assessment tasks: Sources and implications. *Applied Measurement in Education, 6,* 281-305.

Magone, M. E. (1999). *Quantitative analysis of the NAEP Floating Pencil task.* Unpublished manuscript, Learning Research and Development Center, University of Pittsburgh.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis.* Thousand Oaks, CA: Sage.

National Academy of Education. (1997). *Assessment in transition: Monitoring the nation's educational progress.* Stanford, CA: Author.

National Research Council. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24,* 307-353.

Yepes-Baraya, M. (1997, April). *Lessons learned from the coding of attributes for the 1996 National Assessment of Educational Progress (NAEP) science assessment: Grade 4 results.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Appendix A

Fourth-Grade Codes

I. Problem Representation

- Procedures. Low-level or general procedures, such as "pour water into the bowl," "follow directions," and "answer questions," as well as other responses, such as "don't know."
- Make / See pencil float. Responses focus on the floating pencil (e.g., "make the pencil float" and "see if the pencil floats").
- Distinguish between two waters. Responses such as determining which water "floats best" or trying to distinguish between the two waters.
- Compare / Contrast. Responses involve comparing tests of the mystery water with tests of the fresh and salt water.
- Identify. Responses such as "see if the water is fresh water or salt water."

II. Strategies

- Complete / Correct. No errors in data collection.
- Nearly Complete / Correct. Small errors (e.g., minor numerical inaccuracies in recording data or didn't mark an observation on the drawing of the ruler).
- Partial. A viable strategy, but with more major and frequent errors and omissions.
- Inadequate. Multiple or major errors (e.g., testing the wrong water or holding the floating pencil up to the drawing of the ruler backwards). The student would not be able to accurately identify the concentration of the mystery water from the data.

III. Monitoring Behaviors

- Problem recognition. Student identifies problem in comprehension or the performance of the task. This might be demonstrated by (a) statements (e.g., "I think I did something wrong"), (b) corrections / adjustments, or (c) questions to interviewers.

- Physical behaviors. Student uses physical actions to monitor work (e.g., student holds his thumb on the place on the pencil where it rose above the water).
- Rereading directions. Student rereads directions before proceeding.
- Planning. Student talks about what he/she is going to do next or what he/she needs to do to carry out the experiment.
- Consulting data. Student refers back to data in order to interpret findings.

IV. Explanations

A. Identification of unknown: Question 10, "Is the mystery water fresh water or salt water? How can you tell what the mystery water is?"

- Complete. Student identifies the mystery water and correctly and completely justifies his/her answer using a comparison with other data from the task.
- Partial. Student identifies the mystery water in a way that was consistent with other data from the task, but the explanation is missing, unclear, or incomplete.
- Inadequate. Student does not apply the data in a relevant way to answer the question; for example, the student bases his/her answer on irrelevant features such as color of the water, "icing" on the container, or bubbles in the water.

B. Understanding of floating

1. "Why does the pencil float?" (asked in interview only)

- Pencil weight. Responses such as "the pencil is light."
- Physical features of the pencil. Includes responses such as the pencil is "skinny," "hollow," has a thumbtack, eraser, no lead, has air in it, made of wood/wood floats.
- Water pushing up. Student explains that water molecules are forcing the pencil up.
- Physical features of the water. Mention of salt, "stuff," air, or bubbles in the water.

- Relative weights: pencil and water. Student mentions that the pencil floated because the water was heavier than the pencil, or because the pencil was lighter than the water.
- Gravity. References to gravity as a reason for why the water floated, even though student didn't seem to understand how it worked.
- Other. Don't know / other forces (e.g., magnetism).

2. Question 5, "How does the way the pencil floats in the salt water compare with how it floated in the fresh water?

A. In the salt water, the entire pencil sinks below the surface.

B. In the salt water, more of the pencil is below the water than before.

C. In the salt water, more of the pencil is above the water than before.

D. In the salt water, the same amount of the pencil is above the water as in the fresh water."

and Question 8, "If you dissolved more salt in the salt water, how would it change the way the pencil floats?

A. The pencil would float higher than it did before the extra salt was added.

B. The pencil would float at the same level as it did before the extra salt was added.

C. The pencil would float lower than it did before the extra salt was added."

- Complete. Student answers both Questions 5 and 8 completely and correctly. On Question 5, the student answers "C" and gives a correct explanation. On Question 8, the student answers "A" and gives a correct explanation.
- Partial.
  - Partial A. Student chooses the incorrect multiple-choice option on one of the items, although verbal explanations are correct.
  - Partial B. Student gives a correct answer and explanation for either Question 5 or Question 8.

- Partial C. Student answers “A” but has an explanation that refers to pressure, pushing, or pencil weight making the water level rise.

- Inadequate. Student gives no correct answers and shows no understanding of the question.

C. Conservation of water volume: Question 2, “Does the amount of water change when you add the pencil?

A. More water than before the pencil was added.

B. The same amount of water as before the pencil was added.

C. Less water than before the pencil was added.”

- Complete. Student answers “B” and gives a correct explanation.

- Partial.

  - Partial A. Student answers “B,” but provides no explanation or an inadequate explanation.

  - Partial B. Student answers “A,” but has an explanation that refers to the pencil taking up space to make the water rise.

  - Partial C. Student answers “A,” but has an explanation that refers to pressure, pushing, or pencil weight making the water level rise.

- Inadequate. Student answers “A” or “C” and shows no understanding of the problem.

  - Inadequate A. Responses such as “the pencil made the water level go up.”

  - Inadequate B. Responses such as “the water level went up.”

  - Inadequate C. Other responses.

D. Practical reasoning: Question 11, “When people are swimming, is it easier for them to stay afloat in the ocean or in a freshwater lake? Explain your answer.“

- Complete. Student answers “Ocean” and gives a clear and complete explanation.

  - Complete A. Explanations such as “Things float better (higher) in salt water than in fresh water.”

    - Complete B. Explanations such as “The ocean has more salt and things float better in salt water. “
    - Complete C. Explanations indicating that things would float better in salt water (ocean), giving specific reference to the tests performed in this study.
- Partial. Student says “ocean,” but the explanation is missing, unclear, or incomplete.
    - Partial A. Responses such as “the ocean has more salt” or “the ocean is heavier.”
    - Partial B. Explanations such as “salt keeps things up higher.”
    - Partial C. Incomplete references to the test, for example, “the ocean because pencils are higher.”
- Inadequate. Student does not apply the data in a relevant way to answer the question.
    - Inadequate A. Student bases the explanation on irrelevant features such as waves, sharks, or salt getting in one’s nose and eyes.
    - Inadequate B. Student gives wrong explanations due to procedural errors in data collection, (e.g., the student answers salt water because he/she had tested the other waters in the wrong order).

Appendix B

Eighth-Grade Codes

I. Problem Representation

- Complete. Student understands process (comparison of pencil floating in different concentrations of salt solution), outcome (identify an unknown salt solution) and how each step contributes to the problem solution.
- Partial: Goal. Student correctly identifies the goal of the task at least once, but doesn't consistently relate processes to outcome. The student understands that he/she is supposed to identify the concentration of the unknown salt solution, but doesn't articulate how each step will contribute to the problem solution.
- Partial: Process. Student correctly identifies the processes involved in the task, but misidentifies the outcome. The student understands that he/she is supposed to compare how much the pencil floats in three salt solutions, but doesn't explain that the goal is to identify the unknown concentration.
- Inadequate. Student identifies neither process nor outcome, or student initially identifies the correct process but subsequent representations are vague and/or fragmented.

II. Strategies

- Complete. Data and graph are complete and accurate. The student has good, appropriately transformed data from which to draw conclusions about the concentration of the unknown salt solution.
- Partial. Data and/or graph contains small errors which affect the student's ability to draw completely accurate conclusions from the data.
- Inadequate. Data collection and/or graph is inaccurate. The student would not be able to accurately identify the concentration of the unknown salt solution from the data and/or its transformation.

III. Monitoring Behaviors

- Re-reading directions. Student re-reads directions before proceeding.
- Problem recognition. Student identifies problems in comprehension or performance of the task. This might be demonstrated by (a) statements (e.g., *"I think I did something wrong"*), (b) corrections / adjustments, or (c) questions to interviewers (e.g., *"When you say distilled water, that means it's been sitting for a while, right?"*).
- Consulting data. Student refers back to data in order to graph or interpret findings.
- Physical behaviors. Student uses physical actions to monitor work (e.g., student holds his / her thumb on the place on the pencil where it rose above the water).
- Tracking progress. Student verbally monitors what he / she has done and / or how well he / she has done it.
- Planning. Student talks about what he / she is going to do next or what he / she needs to do to carry out the experiment.

IV. Explanations

A. Identification of unknown: Question 14, "Based on the graph that you plotted, what is the salt concentration of the unknown solution? Explain how you determined your answer."

- Complete. Student uses the graph or table to correctly identify the unknown solution (with "correct" being defined as based on accurately computed averages in the data).
- Partial. Student uses an appropriate strategy (i.e., graph or proportional reasoning) but makes errors because of an incorrect graph or averages.
- Inadequate. Student incompletely or inaccurately describes how to identify the unknown solution. An explanation may be inadequate if it

  (a) describes a strategy other than using the graph or proportional reasoning,

  (b) includes a vague or large range for the concentration of the unknown,

(c) describes a proportional reasoning strategy that compares the unknown to one solution but not the other, or

(d) describes a proportional reasoning strategy without quantifying the height differences between the pencil in each solution.

B. Nature of floating: A composite of responses to five questions

Question 1, "Explain why the pencil floats when it is put in the water."

Question 7, "Now take the pencil and put it in the 25% salt solution, eraser-end down. How does the pencil float in the solution compared to how it floated in the distilled water?

A. In the salt solution, more of the pencil is above the surface.

B. In the salt solution, more of the pencil is below the surface."

Question 9, "Why does the pencil float at a different level in the salt solution that in the distilled water?"

Question 10, "If you added more salt to the 25% salt solution and stirred the solution until the salt was dissolved, how would this change the way that the pencil floats?

A. Less of the pencil would be above the surface.

B. More of the pencil would be above the surface.

C. There would be no difference in the amount of pencil above the surface."

Question 13, "Based on the graph that you plotted, how does the length of the pencil that is above the surface of the water change when the salt concentration changes?

A. It increases as the salt concentration increases.

B. It decreases as the salt concentration increases.

C. It remains constant as the salt concentration increases."

- Complete. Explanation focuses on relative density and its effect on floating. The student explains that the pencil floats because it is less dense and that as salt concentration increases, so does density and the "floating force" placed on an object.

- Partial: Force. Student discusses the force applied to the pencil without discussing relative density (e.g., salt causes objects to float more because there are more molecules in the water to push against the pencil).
- Partial: Mass. Student thinks that molecules in the water take up space to hold up objects. The more molecules (e.g., salt) in the water, the more room they take up and the less room there is for the pencil. This explanation considers the mass of the object as a factor in floating but doesn't consider density or force.
- Inadequate. Student mentions features other than force, mass, or density to explain floating, or mentions a scientific term associated with floating (e.g., buoyancy) without further elaboration.

Designing Assessment Tasks. According to the constructive alignment theory by Biggs and Tang (2007), assessment tasks (AT) and teaching-learning activities (TLA) are designed to ensure that students achieve the intended learning outcomes (ILO) and develop cognitive skills at a range of levels. The learning outcomes for a topic/unit are the criteria against which instructors make judgments about student learning.Â 4.1 Analysis of elements 4.2 Analysis of relationships 4.3 Analysis of organisational principles. 5. Synthesis (structuring elements to form a pattern not previously apparent).Â Determine students prior knowledge Using a simple quiz (LumiNUS Assessment) with multiple-choice questions, faculty can quickly gauge their studentsâ€™ knowledge level. Empirical evaluation remains the most used approach for the algorithm assessment, although ML algorithms can be eval-uated through empirical assessment or theory or both, e.g., derived generalized bounds and empirical results (Marchand & Shawe-Taylor, 2002). Evaluation techniques based on multiple experiments are considered in Dietterich (1998), one of the most cited work on empirical evaluation of ML algorithms.Â The criterion for the performance of a classier is its performance on relevant documents, a well-dened unimodal positive class, independently of performance on the irrelevant documents. Precision and Recall do not depend on tn, but only on the correct labelling of positive examples Ã°tpÃž and the incorrect labelling of examples (fp and fn). Well-established assessments, such as PISA, TIMSS, and PIAAC, all make effective use of unfocused block designs of two or more subjects administered to the same student. In refining its analytic procedures, NAEP will benefit from decades of experience accumulated by the suite of international assessments. Implications for NAEP.Â Students completed writing tasks on laptop computers provided by NAEP using software similar to common word processing programs. In 2012, NCES conducted a study of computer-based writing at grade 4. Lessons learned from this study provided insights into some of the challenges encountered and solutions applied for the administration. Learn More. Science Interactive Computer Tasks (ICTs).