DOCUMENT RESUME

ED 411 314 TM 027 375

AUTHOR Solano-Flores, Guillermo; Shavelson, Richard J.; Ruiz-Primo, Maria Araceli; Schults, Susan Elise; Wiley, Edward W.; Brown, Janet H.

TITLE On the Development and Scoring of Classification and Observation Science Performance Assessments.

PUB DATE 1997-03-00

NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Classification; *Elementary School Students; Grade 5; Intermediate Grades; Interrater Reliability; *Observation; *Performance Based Assessment; Problem Solving; Science Tests; *Scoring; *Test Construction

ABSTRACT

In a project on the development of performance assessments in science, researchers have developed four types of assessment: comparative, component identification, classification, and observation. They have noted that assessments belonging to the same type of task can be scored for the same performance properties. This paper reports on the relationship between the characteristics of the task and the characteristics of the scoring system for assessments involving classification and observation. The development of two such tasks, a classification task called "Sink and Float" and an observation task called "Daytime Astronomy," is described. Scoring systems are also described. Sink and Float asks students to identify weight and size as dimensions critical to sinking and floating, while Daytime Astronomy asks students to project shadows in a box with a flashlight and solve location problems through the "sun" shadows. Sink and Float was tested with 2 5th-grade classes (32 students) who were rated by 2 raters. Daytime Astronomy was piloted with 3 5th-grade classes (58 students). There were reasonably high interrater reliabilities for both assessments, and the problems presented in each assessment distinguished different types of knowledge with the domain addressed by the assessment. Results suggest that observation assessments in particular address a very elusive kind of knowledge. Additional study is needed before these assessments are ready for general use. (Contains 6 tables and 18 references.) (SLD)

On the Development and Scoring of Classification and Observation Science Performance Assessments*

Guillermo Solano-Flores
WestEd

Richard J. Shavelson
Maria Araceli Ruiz-Primo
Susan Elise Schultz
Edward W. Wiley
Stanford University

Janet H. Brown
University of California, Santa Barbara

Paper presented at the Annual Meeting of the American Educational Research Association Chicago, IL, April 24-28, 1997

DRAFT

In recent years, we have taken steps to formalize a process of science performance assessment development. First, we identified three components that define a performance assessment: a task that poses a problem whose solution requires the use of materials that react to the students' actions; a response format that captures the students' actions, findings, and explanations; and a scoring system that records and evaluates performance numerically based on the scientific defensibility of the procedures used and the results obtained by students (Ruiz-Primo & Shavelson, 1996; Shavelson, 1995; Shavelson & Baxter, 1992). Second, we have acknowledged that these components are intimately related and must be developed together--changes in one component imply changes in the other two components (Solano-Flores & Shavelson, 1997). Third, we have postulated that the tasks in science performance assessments consist of investigations that recreate to

* This investigation would not have been possible without the collaboration of Marylin Bachman and Heather Lange and the S.E.E.D. office of the Pasadena School District, who allowed us to come into their classrooms.

some extent the conditions under which scientists work and elicit the kind of thinking and reasoning used by scientists when they solve problems.

The assessments we have developed can be characterized according to four task types that give substance to the claim that there is a knowledge domain associated with what has been lumped together as "science process skills." These task types are: (1) Comparative--conduct an experiment to compare two or more objects on some attribute; (2) Component Identification--test objects to determine their component parts, or how those parts are organized; (3) Classification--classify objects according to critical attributes to serve a practical or conceptual purpose; and (4) Observation--perform observations and/or model a process that cannot be manipulated (Ruiz-Primo & Shavelson, 1996; Shavelson, 1995).

We have observed that all assessments belonging to the same type of task can be scored for the same performance properties. For example, in comparative investigations, scoring focuses on the scientific soundness of the procedures used by students to manipulate, control, and measure variables (e.g., Baxter, Shavelson, Goldman, & Pine, 1992; Solano-Flores, 1994; Solano-Flores, Jovanovic, Shavelson, & Bachman, 1994); in component identification investigations, scoring focuses on the appropriateness of the evidence used by students to determine the presence or absence of component parts (e.g., Shavelson, Baxter, & Pine, 1991; Baxter & Elder, 1994; Druker, Solano-Flores, Brown, & Shavelson, 1996).

Research on characteristics of the tasks and scoring systems for classification and observation assessments have not yet been reported. The purpose of this paper, then, is to discuss the relationship between the characteristics of the task and the characteristics of the scoring system for these two types of assessments. We describe the process of development a classification assessment and an observation assessment and present some preliminary findings. We also discuss the challenges we have encountered in developing and using these assessments that might be used to improve the development of future classification and observation assessments.

Assessment Development

In this section we describe the process of development of a classification and an observation assessment. For each assessment, we present first a simple conceptual framework for the task type involved; then we describe the characteristics of the task, response format, and scoring system used. Our discussion emphasizes how the characteristics of a task type influence the characteristics of the scoring system.

A Classification Assessment: Sink and Float

Conceptual Framework. Classification is a fundamental science activity. Classificatory systems are used in many disciplines and are a necessary tool for the development of theories (e.g., Aldenderfer & Blashfield, 1984; Bailey, 1994; Sokal & Sneath, 1963). Classification is often conceived as just the ordering or grouping of cases based on their similarity on critical attributes. However, classification is much more than organizing objects or events. It usually involves a purpose, either conceptual or practical (see Bailey, 1994; Sokal & Sneath, 1963). Besides the process of classifying, classification is also the "end result" of that process, or the use of that "end result." Fundamental activities in any scientific discipline, such as describing, making predictions, or identifying dimensions (i.e., attributes, properties) characteristic of a phenomenon, are all instances of classification (see Aldenderfer & Blashfield, 1984; Bailey, 1994).

A classification task, then, encompasses a process (e.g., identify categories to which objects belong, identify which dimensions are needed to construct a goal-oriented classification scheme), an end result (e.g., a classification scheme based on critical dimensions, a description of how those dimensions are related), and an application (e.g., use a classification scheme to make inferences or predictions on certain objects).

Since classification is common to many scientific disciplines, there is no reason why a classification assessment should be limited to the overused, well-

known tasks of classifying leaves or rocks (e.g., the "Leaves" assessment developed by the California Assessment Program, 1992). To develop our classification assessment we selected, then, a content area other than botany or mineralogy. We selected flotation, a physics topic covered by many hands-on science curricula (e.g., Full Option Science System, Science for Early Educational Development, and National Science Resource Center). Using this content domain, we devised Sink and Float, a classification assessment for fifth and sixth graders intended to assess knowledge on flotation.

The core concept of flotation is density (d): the relation between weight (w) and volume (v).[1] Therefore, the problems included in a classification assessment on flotation should involve identifying weight and volume as critical dimensions to floating and sinking, creating and using a classification scheme based on those dimensions, and defining how those dimensions are related.

Task. In Sink and Float, students are given a tub filled with water, and 12 plastic bottles of different sizes and weights that they can place in the water. The bottles are "specimens" to can be classified by weight, volume, and whether they are "floaters or sinkers". Students are posed with four problems (see Response Format section) whose most efficient solutions involve: (1) treating weight and volume as inseparable dimensions critical to floating and sinking, and (2) identifying how the relation between these dimensions (density = weight / volume) determines whether an object is a floater or a sinker.

Response Format. The response format for Sink and Float consists of a notebook that poses classification problems and provides directions for using the equipment. The notebook is also intended to capture the students' responses--both their solutions to the problems and reasoning and strategies they used to arrive to those solutions. The notebook includes four problems: (1) Find out what makes

[1] Strictly speaking, we should use the word, "mass." However, we found that most of students are not familiar with it, so we decided to use "weight." For the same reason, in the assessment we used the word "size" instead of "volume".

bottles float or sink--identify the bottles as floaters or sinkers and determine the dimensions that are critical to floating-sinking; (2) Sort your bottles--classify the bottles according to size, weight, and whether they are floaters or sinkers; (3) Explain how size and weight make bottles float or sink--when provided with an accurate classification scheme, determine how the dimensions of weight and volume are related identify an object as a floater or a sinker; and (4) Tell floaters from sinkers without using water--based on the information about weight and size for a new set of bottles, but without actually having the bottles, classify bottles as floaters or sinkers.

Since problem 3 provides an accurate classification scheme, whereas problem 2 asks students to construct a classification scheme, the notebook is divided in two parts. When students complete Part 1 (problems 1 and 2), they return their notebooks and get Part 2 (problems 3 and 4). This reduces the possibility of carrying forward mistakes made in solving problem 2 to problems 3 and 4; also, it also prevents students from seeing an accurate classification scheme (problem 3) when solving problems 1 and 2. Figure 1 shows Problems 2 and 3. In all the problems, the students are allowed to provide answers with words, drawings, or both.

Scoring System. The scientific defensibility of a classification system depends on how well some formal criteria (e.g., exhaustiveness and mutual exclusiveness) are met as well as how relevant the dimensions used in the classification system are to the conceptual or practical purposes intended. Therefore, the scoring system for classification tasks is dimension-based--it focuses on the relevance and accuracy of the dimensions used by a student to construct or use classification schemes with specific conceptual or practical purposes.

Problem 2:
Sort your bottles.

In the space below make a chart or a drawing to sort your bottles by size and weight. Refer to the bottles with their letters. Show which bottles are floaters by circling their letters.

Problem 3:
Explain how size and weight make bottles float or sink.

In the chart below your bottles are sorted by size and weight.

White boxes show floaters. Shaded boxes show sinkers.

	small	medium	large
1 ounce	J		
2 ounces	N and G	T	V and P
3 ounces	D and K	B	R and C
4 ounces			H

Figure 1. Two problems from the Sink and Float assessment.

In Sink and Float the quality of performance is based on how effectively students: identify weight and size as dimensions that are critical to floating-sinking (Problem 1); classify bottles by those dimensions (Problem 2); explain how those dimensions interact to determine floating-sinking (Problem 3); and use information on those dimensions to predict whether objects are floaters or sinkers (Problem 4).

Figure 2 presents portions of the Sink and Float scoring form. To score a student's response, the rater must check the boxes that best describe the characteristics of the response. The small numbers in those boxes are weights assigned to those characteristics and are intended to reflect performance quality or

complexity. For problems 1 and 3, the scoring form consists of a list of attributes that characterize the precision of the descriptions and explanations given by the students.

Problem 2:
Constructing a classification scheme.

Examine and determine classification scheme used, then count the number of bottles classified correctly. Check only one box.

By Volume or Weight					*By Volume and Weight separately*					*By Volume and Weight in combination*				
1	2-5	6-9	10-12		1	2-5	6-9	10-12		1	2-5	6-9	10-12	
2	3	4	5	or	3	4	5	6	or	4	5	6	7	

Count the number of floaters identified correctly. Check one box.

Floaters			
1	2-5	4-5	6
1	2	3	4

Problem 3
Using a classification scheme to explain floating/sinking.
Check all boxes that apply.

Describes correct relationship of Volume and floating/sinking	1
Describes correct relationship of Weight and floating/sinking	1
Treats Volume and Weight as inseparable	1
States correct relationship as a principle/enumerates all levels of variable(s)/describes extreme cases	1

Add the scores for the boxes checked: ☐

Figure 2. Portions of the scoring form for the Sink and Float assessment.

For problems 2 and 4, the scoring form consists of a series of mutually-exclusive cells that describe both the number of bottles correctly classified or identified and the completeness of the strategy used. For example, in Problem 2, the student can classify all the bottles correctly, but that is not enough to obtain the

maximum score because the bottles can be classified with three different classification schemes of varied effectiveness: by either volume or weight, by volume and weight separately (e.g., two charts, one for volume, one for weight), or by volume and weight in combination (e.g., on a single chart in which volume and weight are treated as inseparable).

An Observation Assessment: Daytime Astronomy

Conceptual Framework. Although observation is inherent to any science activity, it becomes a type of investigation in its own right when the phenomena under study are not directly accessible to the senses, are typical of phenomena that take place over long periods of time, occurred a long time ago, or are beyond manipulation or control. Formally speaking, an observation investigation is actually an indirect observation investigation--it depends on the evidence of the phenomena studied, rather than the observation of the actual phenomena. Observation does not occur by itself; it implies the development of models that make sense of the data gathered and represent the phenomena studied (e.g., Bunge, 1967; Hesse, 1963).

What scientists look for when they perform observations is, of course, influenced by their prior knowledge or models of the phenomenon under study. Those models influence how the results of their observations are interpreted (Carin, 1993). An observation task, then, encompasses performing observations on a phenomena, using a model that determines how those data are gathered, and describing the results obtained.

We selected astronomy observation, an earth science topic covered by many hands-on science curricula (e.g., Full Option Science System, Science for Early Educational Development, and National Science Resource Center). Through the study of the unit, students observe and record, for several days, the shape of the moon and the motion of shadows projected by the sun; develop and discuss models on the position of the earth, the sun and the moon; and explain the results of their

observations. The key skill in astronomy observation is the use of models to interpret observations. The problems included in an observation assessment on daytime astronomy, then, should involve the use of models to collect and interpret data. Using this criterion, we devised Daytime Astronomy, an observation assessment for fifth and sixth graders intended to assess the use of models on the motion and position of the earth and the sun.

A major challenge to developing an observation investigation assessment is the fact that the activities included in observation-type hands-on instructional units are completed over several days. Yet, to insure standardization, the assessment must be administered in a single session of roughly 45 minutes. We used an earth globe and a flashlight to simulate the motion and position of the earth and the sun (see below), which not only served the purpose of involving the use of models (a characteristic of observation investigations), but also helped to simplify the administration of the assessment.

Task. In Daytime Astronomy students are given an earth globe inside a carton box (the box is large enough for the globe to spin and insures enough darkness so the shadow projected with the flashlight can be readily seen), a pocket flashlight, and a set of "sticky towers" (see Figure 3). Students are asked to use the pocket flashlight as if it were the sun, to project sun shadows with the towers, and to solve six location problems (see Response Format below) by observing the sun shadows projected by the towers. The correct solutions to these problems involve: (1) pointing the flashlight onto the equator and (2) modeling the earth rotation from West to East.

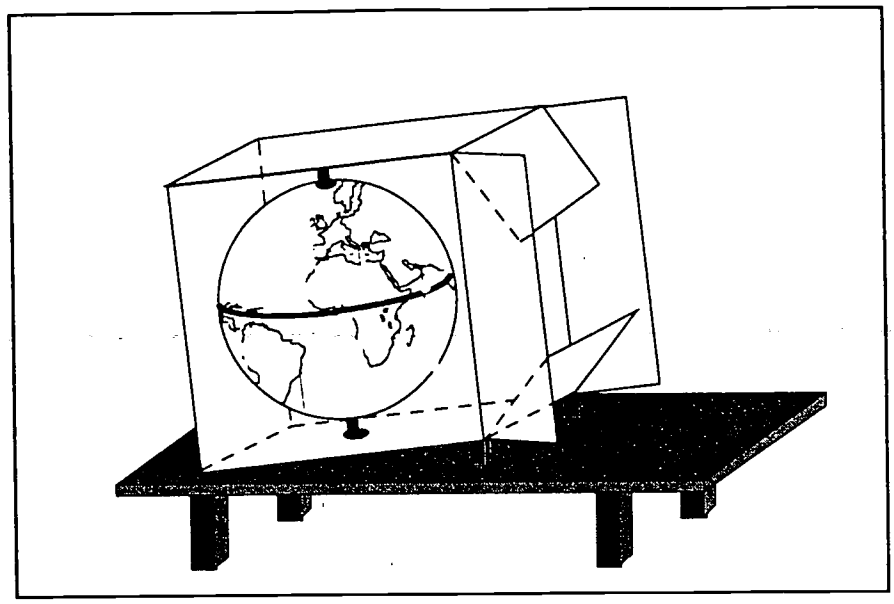

Figure 3. The Daytime Astronomy assessment.

Response Format. The response format for Daytime Astronomy consists of a notebook that poses observation problems and provides directions on the use of equipment. The notebook also captures the students' responses--both their solutions to the problems and the reasoning and strategies they used to solve those problems. The notebook has six problems: (1) Where in the US is tower C (given the location, length, angle, and orientation of the shadows for towers A and B, and a the length, angle, and orientation of the shadow for Tower C)?; (2) What does Tower A look like at 10 AM and 3 PM?--model what a tower's shadow looks like at a specific location in the Northern Hemisphere at 10 AM and 3 PM; (3) What time is it in Seattle when it's noon for Towers A and B (given the location, length, angle, and orientation of the shadows for towers A and B)?; (4) What do Sun shadows and the Earth's motion have to do with time?; (5) What does Tower D look like at 10 AM, noon, and 3 PM?--model what a tower's shadow looks like at a specific location in the Southern Hemisphere at 10 AM, noon, and 3 PM; and (6) Do Sun shadows move and change the same way in the Northern and Southern hemispheres?--

hemispheres?--describe similarities and differences of sun shadows in the Northern and Southern hemispheres.

Since Problems 2 and 5 are parallel, to insure their mutual independence the notebook is divided in two parts. When students complete Part 1 (problems 1 to 4), they return their notebooks and get Part 2 (problems 5 and 6).

Figure 4 presents Problem 1. In some cases students are asked explicitly to provide their answers with drawings. In others, when they have to justify their actions or provide explanations, they are allowed to use words, drawings, or both.

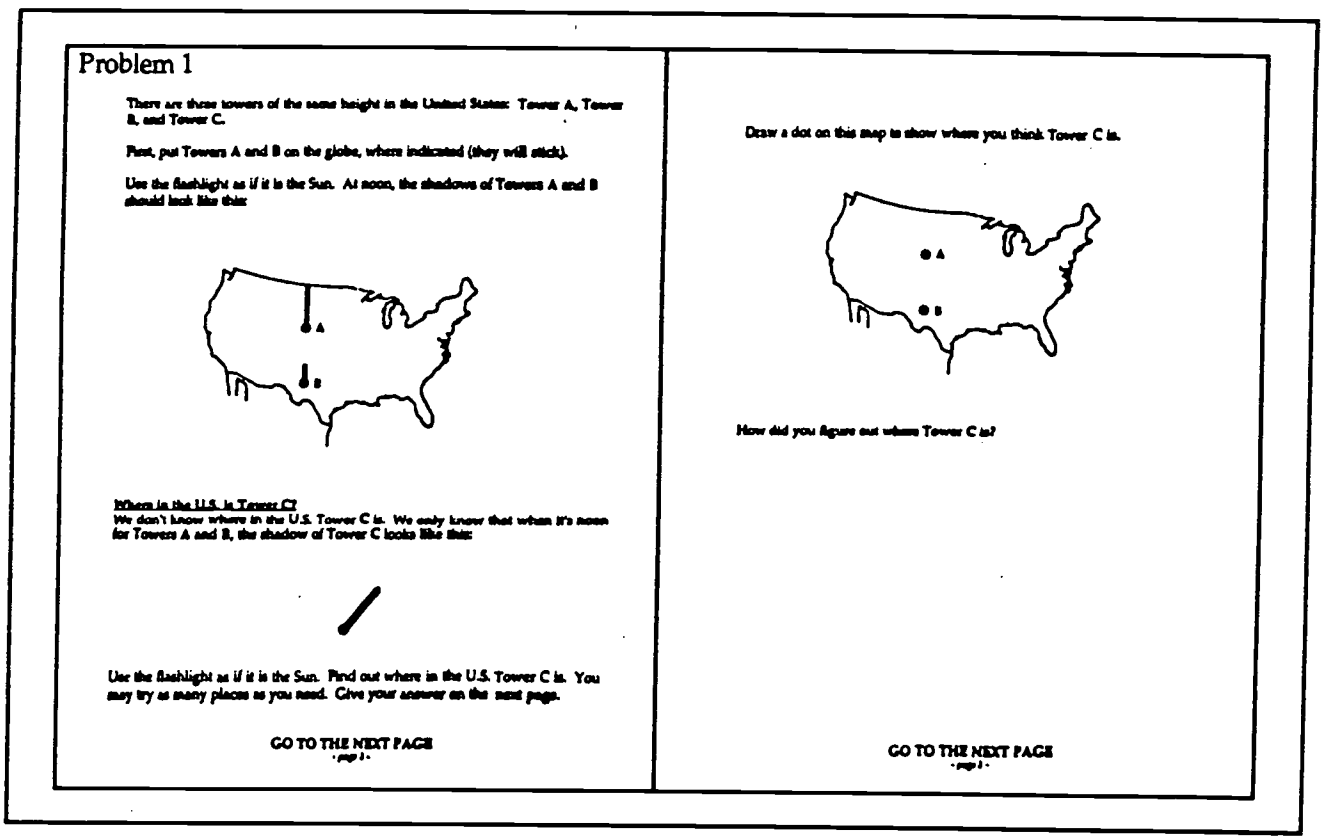

Figure 4. Portion of the response format of the Daytime Astronomy assessment.

Scoring System. The scientific defensibility of an observation investigation is determined by how well the student's model represents the application of relevant knowledge about the phenomena being studied, the quality of observations carried out, and the quality of the descriptions/explanations that make sense of those

observations. Therefore, the scoring system for an observation task is accuracy-based--it focuses on the accuracy of the model used, the accuracy of the results, and the accuracy of the description and interpretation of the results.

In Daytime Astronomy the quality of performance is determined based on the accuracy of three performance components: Observations/Results (Problems 1-3 and 5), Data Gathering/Modeling (Problems 1-3 and 5) and Description/Explanation (Problems 4 and 6). These three components can be thought of as: "What results did you get?," "How did you obtain those results?," and "How do you account for those results?"

Observations/Results performance is assessed based on physical evidence provided by students, like the place on the US map where the student draws a dot to indicate where a tower is located, or the characteristics of the shadow (angle, length, and orientation) projected by a tower. Data Gathering/Modeling performance is assessed based on the students' descriptions of the procedures they used to obtain or model data (e.g., what they did to model the shadows projected by the towers at a certain location and time of day). Description/Explanation performance is assessed based on the students' descriptions and interpretations of results (e.g., what they observed and how they explained the results obtained).

The process of developing the scoring form for Description/Explanation was very complicated; it took many iterations. Since the scoring of this component relies heavily on language use, the major challenge was not to penalize or privilege students for their writing. In addition, students' actions and explanations are frequently confounded in their descriptions, or they provide information that does not allow one to distinguish whether they are reporting something they did or something they thought.

We attempted to make up for the ambiguity of language by designing a scoring form that specifies at least most of the possible approaches students can use in solving the problems. The price for this level of detail is the length of the scoring form, a page for each problem (See Figure 5) that functions as an inventory of the

possible performance characteristics (variables). Variables are grouped in sections (separated by heavy lines). To score a student's response, the rater must check the boxes for the variables (separated by thin lines) that best describe the characteristics observed in that response. The small numbers in the boxes for the variables are weights that reflect performance quality. A component score is computed by adding the weights for the boxes checked.

Where in the US is tower C?

Observations/Results

Tower C is in Eastern US		1
Tower C is in North Eastern US		1
Tower C is somewhere between Pennsylvania and Maine		1

Data Gathering/Modeling

Flashlight Position	Points flashlight at Equator	2
Flashlight motion	Moves flashlight from E to W	2
Globe Rotation	Rotates globe	1
	Rotates globe from W to E	2
Towers	Moves tower C around on the map/globe until shadow is matched	1
	Moves tower C around on the map/globe only in the E/NE region until shadow is matched	2
Shadows	Uses shadows of towers A and B as reference	1

Description/Explanation

Sun Position	Mentions that sun rays hit the earth around the equator	2
Shadow Orientation	Mentions shadow orientation	1
	Shadows point to the left if Tower C is placed on the W	2
	Shadows point to the right if Tower C is placed on the E	2
	Shadows point to the right when tower is on the E and to the left when tower is on the W	3
	Shadow orientation varies according to where tower C is located with respect to the sun rays	3
Shadow Length	Mentions shadow length	1
	The higher the tower is placed on the map/globe, the longer its shadow gets	2
	The lower the tower is placed on the map/globe, the shorter its shadow gets	2
	The higher the tower is placed on the map/globe, the longer its shadow gets, and the lower the tower is placed, the shorter its shadow gets	3
	Shadow length increases with latitude	3
	Shadows are longer at places far from Equator and shorter at places close to Equator	3
Shadow Angle	Mentions shadow angle	1
	Shadow angle varies according to where tower is placed	2
	Shadow angle is wider as tower is placed farther right	2
	Shadow angle is wider as tower is placed farther left	2
	Shadows angle is wider as towers is placed farther right or farther left	3
	Shadow angle varies with meridians	3

Figure 5. Portion of the scoring system for the Daytime Astronomy assessment.

For Observations/Results, the sections (each containing one variable) describe the physical characteristics of the results (e.g., the section describes the location of the dot drawn by the student to represent the location of the Tower C in problem 1). The score reflects how accurate the student's response is about the tower location. All the sections are weighted 1. The maximum score is the sum of the variables scored '1.'

For the Data Gathering/Modeling and Description/Explanation components, the sections describe a variety of approaches to solve the problem. Students are not expected to use all the approaches; the variables for many sections may not be selected. The variables within a section are mutually exclusive; only one variable must be checked.

Piloting the Assessments

The evaluation of the assessments has focused on interrater reliability and two aspects of validity: (a) knowledge domain specification--the ability of the four Sink and Float problems or the Daytime Astronomy performance components to distinguish different kinds of knowledge; and (b) sensitivity to differences due to instruction.

Sink and Float

We administered Sink and Float to two classes of fifth-grade students from a middle-to-high SES school in California with a curriculum that emphasizes hands-on science. Class 1 (n = 16) used the hands-on instructional unit, "Sink and Float," developed by the National Science Resource Center (NSRC, 1994). Class 2 (n = 16) did not study the unit. Both classes were tested at the same time on two occasions, before and after Class 1 studied the unit.

Two raters were trained to use the scoring form for Sink and Float with a sample of responses selected randomly. The raters scored the notebooks independently, discussed the differences they found, agreed upon the ways in which

the scoring forms should be interpreted, and, when necessary, modified the scoring forms to make them more explicit. This was repeated with another sample of responses until an interrater reliability of at least .90 was reached. Once no further modifications were needed and raters achieved a reliability .90, the raters independently scored the students' notebooks. Both raters scored all student responses.

Interrater Reliability. Interrater reliability coefficients for pre-test and post-test total scores were reasonably high (.87 and .83, respectively; Table 1). The coefficients obtained for problem scores are moderate to high, except for Problem 3.

In reviewing raters disagreements, we found that one of the raters had difficulty in identifying whether students treated the critical dimensions as inseparable. This rater tended to provide the highest score when students just mentioned the two dimensions, even though they did not relate the dimensions in any way. The same rater had similar difficulties with Problems 2 and 4 identifying the approach students used to create a classification scheme and make predictions. Recalibration, then, may be especially important in this kind of assessment.

Table 1
Interrater Reliability Coefficients for the Sink and Float Assessment by Problem and the Composite Total Score Across Groups.

Problem	Pre-Test	Post-test
Problem 1	.79	.95
Problem 2	.86	.75
Problem 3	.77	.64
Problem 4	.75	.83
Total Score	.87	.83

Knowledge Domain Specification. Table 2 shows the estimated variance components obtained with a series of student x rater x problem Generalizability (G) studies. Averaging across classes and occasions, the problem facet accounts for 46.59 percent of the score variability, indicating substantial differences in difficulty across problems. The considerable score variability due to the student x person interaction (which, averaged across classes and occasions accounts for 30.56 of the total score variability) indicates that a given problem was not equally difficult for all students. Thus, the four problems seem to distinguish different kinds of knowledge.

Table 2

Estimated Variance Components and Generalizability Coefficients for a Student x Rater x Problem Design in the Sink and Float Assessment.

	Pre-Test		Post-Test	
Source of Variation	Estimated Variance Components	Percent of Total Variability	Estimated Variance Components	Percent of Total Variability
Class 1-Instruction				
Student (s)	.00306	4.31	.00958	9.66
Rater (r)	.00004	0.05	.00002	0.02
Problem (p)	.03593	50.65	.05108	51.53
s x r	0*	0.00	.00123	1.24
s x p	.02185	30.80	.02739	27.63
r x p	.00013	0.18	0*	0.00
srp,e	.01006	14.18	.00983	9.92
$\hat{\rho}^2$ $(n_r = 2, n_p = 4)$	.31		.52	
$\hat{\phi}$	.16		.31	
Class 2-No instruction				
Student (s)	.00945	16.58	.00134	1.42
Rater (r)	.00020	0.46	.00074	0.79
Problem (p)	.03970	37.90	.04362	46.30
s x r	0*	0.00	0*	0.00
s x p	.01785	24.48	.03754	39.85
r x p	0*	0.00	.00016	0.17
srp,e	.00572	7.84	.01097	11.64
$\hat{\rho}^2$ $(n_r = 2, n_p = 4)$	.65		.11	
$\hat{\phi}$	.38		.05	

* Negative variance components set to zero; in no case was the variance component more than -0.00124.

Relative coefficients were higher ($\hat{\rho}^2$ = .40 averaging across groups and occasions) in magnitude than absolute coefficients ($\hat{\phi}$ = .23 averaging across groups and occasions), reflecting especially the difference in problem difficulty.

Sensitivity to Instruction. Table 3 presents mean scores and their standard deviations by class across occasions for each problem. The most striking result is that mean scores are, in general, lower on occasion 2 in both classes, even for the class that had instruction(!). A series of split-plot ANOVAs performed for both problems and total scores revealed no significant differences ($p > .05$) between classes (C), across occasions (O), or their interaction (CxO) for Problems 1, 2, 3 and Total Score (Problem 1: $F_C = .21$, $F_O = .98$, $F_{CxO} = .11$; Problem 2: $F_C = .15$, $F_O = ..01$, $F_{CxO} = .06$; Problem 3: $F_C = 1.39$, $F_O = .35$, $F_{CxO} = .78$; Total Score: $F_C = 2.57$, $F_O = .001$, $F_{CxO} = .19$). In Problem 4 we found a significant difference between classes ($F_C = 9.78$, $p < .05$), but not between occasions or for their interaction ($F_O = .06$, $F_{CxO} = 1.47$; $p > .05$).[2] Averaged across occasions, Class 2 performed better than Class 1 in predicting which bottles would sink or float.

Possible interpretations for these findings, taken together, are: (1) students either had some naive knowledge of what makes things sink or float (receiving on average, 17 out of 24 possible points), or could attain this score through trail-and-error. (2) Whatever the conceptual difficulties that led to less than perfect performance, these difficulties were not ameliorated by instruction. Hence, essentially no gain from pre- to post-test and no between classroom mean differences were observed. In the end, the assessment may not sufficiently overlap the instruction students received to show changes. Indeed, the teacher found the instructional unit difficult to teach and spread across many important ideas.

[2] To perform the analysis for total scores, we transformed the score on each problem into a proportional score--the score on that problem divided by the maximum score on that problem--and added the proportional scores

Table 3

Mean and Standard Deviations for Each Problem by Occasions and Problems in the Sink and Float Assessment.

	Class 1-Instruction				Class 2-No Instruction			
	Pre-Test		Post-Test		Pre-Test		Post-Test	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	SD
Problem 1 (Max. = 4)	2.59	.71	2.43	1.30	2.78	.75	2.47	.91
Problem 2 (Max. = 11)	8.19	.54	8.31	.48	8.00	.73	7.69	2.44
Problem 3 (Max. = 4)	2.16	1.30	1.69	1.08	2.25	1.35	2.34	1.34
Problem 4 (Max. = 5)	4.47	.67	4.34	.39	4.72	.41	4.91	.26
Total Score (Max. = 24)	17.41	1.87	16.78	2.54	17.75	2.46	17.41	3.10

Daytime Astronomy

Method. We administered the Daytime Astronomy assessment to three classes of fifth-grade students of middle-to-high SES school in Southern California with a curriculum that emphasized hands-on science. Class 1 (n = 20) and Class 2 (n = 19) were taught the hands-on unit on Daytime Astronomy developed by Science for Early Educational Development, SEED (Hamilton, 1994). Class 3 (n = 19) did not receive any instruction. The assessment was administered to the three classes at the end of the instruction of Classes 1 and 2.

Three raters were trained to use the scoring form with a sample of responses selected randomly. The raters scored the notebooks independently, then discussed the differences found, agreed upon the ways in which the scoring forms should be interpreted, and, if necessary, modified the scoring forms to make them more explicit. This was repeated with another sample of students until an interrater reliability of .90 was reached. When no further modifications were needed, the raters independently scored the student

notebooks. The three raters scored all the student responses. Because of the complexity of the scoring form, raters took, on average, almost ten minutes to score each student response.

Interrater Reliability. A series of student x rater G studies were carried out to estimate the magnitudes of measurement error due to raters separate from residual error due to other sources. The G studies were performed with the total scores across problems for each class. The patterns of score variability due to student, rater, and the residual were similar across the three classes. Averaging across classes, they accounted, respectively, for 88.93 %, 1.08 %, and 9.98 % of the total score variation. Also averaging across classes, the $\hat{\rho}^2$ and $\hat{\phi}$ coefficients were .96. Interrater reliability was not a problem. These results were confirmed with other G studies carried out (see below). Percent of variability due to raters was always below 1 percent.

Knowledge Domain Specification. We performed a series of G studies to determine whether the three performance components (i.e., Observation/Results, Data Gathering/Modeling, Description/Explanation) addressed different kinds of knowledge.[3] Because Problems 4 and 6 do not involve performing observations or modeling, two assessment composites were created and treated separately, one with Problems 1-3 and 5 (Composite A), the other with Problems 4 and 6 (Composite B).

For Composite A, a series of student x rater x problem x component G studies was performed, one per class. Table 4 presents the results only for Class 1--Instruction. Results revealed that the facet Component was the major source of score variability. It accounted for 43.10 percent of the total variability. The same pattern was observed on the other two classes (Class 2 = 53.79 percent and Class 3 = 51.48 percent). Thus, the three components address different kinds of knowledge. Relative and absolute coefficients were low, .24 and .08 respectively. Students

[3] To perform this analysis, we transformed the scores of the three components--Accuracy of Observations/Results, Data Gathering/Modeling, and Description/Explanation--into proportions: the score on each component divided by the highest possible score on that component--and added those proportional scores

performed differently across components and problems. In one problem they did better in one component, but in the other problem they did better in another component.

Table 4

Estimated Variance Components and Generalizability Coefficients for a Student x Rater x Problem x Component G Study Design by Class in the Daytime Astronomy Assessment. Composite A, Class 1.

	Observation, Modeling & Explanation	
Source of Variation	Estimated Variance Components	Percent of Total Variability
Class 1-Instruction		
Student (s)	.00293	2.00
Rater (r)	.00005	0.03
Problem (p)	0*	0.00
Component (c)	.06301	43.10
s x r	0*	0.00
s x p	.01522	10.41
s x c	.00184	1.2
r x p	0*	0.00
r x c	0*	0.00
p x c	.00156	1.07
s x r x p	0*	0.00
s x r x c	.00134	0.92
s x p x c	.05100	34.88
r x p x c	.00032	0.22
pra,e	.00894	6.11
$\hat{\rho}^2$ $(n_r = 3, n_p = 4; n_c = 3)$	.24	
$\hat{\phi}$	.08	

* Negative variance components set to zero; in no case the variance component was more than -.00313.

Based on these results we treated performance components as a fixed facet and carried out a series of student x rater x problem G studies for each of the components. We combined the three classes for these G studies (Table 5). Across the three aspects of performance the pattern of variability is the same: the largest variance component was for the error component, student by problem interaction.

Not surprisingly, students' relative standing varied from one problem to the next. This result is not new, task interaction has consistently been found a major source of unreliability (e.g., Shavelson, Baxter, Gao, 1993). Raters did not introduce error variability into the scores (percent of variability is insignificant). Notice that the percent of variability among students is higher for the Data Gathering/Modeling score than for the Observations/Results and Description/Explanation scores (see Table 5). This indicates that modeling scores reflect better the differences in students' performance than the other two types of performance components. The highest relative and absolute "reliability" coefficients were for this type of score ($\hat{\rho}^2$ = .50 and $\hat{\phi}$ = .47). Restriction of score range and difficulty seem to be the reasons for low coefficients in the other two types of scores.

Table 5

Estimated Variance Components and Generalizability Coefficients for a Student x Rater x Problem Design For Each Performance Component in the Daytime Astronomy Assessment The Three Groups Combined.

	Observation		Modeling		Explanation	
Source of Variation	$\hat{\sigma}^2$	%	$\hat{\sigma}^2$	%	$\hat{\sigma}^2$	%
Class 1--Instruction						
Students (s)	.00163	4.28	.00863	15.23	.00156	3.07
Rater (r)	0*	0.00	0*	0.00	.00025	0.09
Problem (p)	.00754	2.16	.00486	8.58	.00005	0.49
s x r	0*	0.25	.00097	1.71	.00196	3.86
s x p	.09635	90.08	.02917	51.49	.03308	65.06
r x p	.00011	0.00	0*	0.00	.00001	.01
srp,e	.00560	3.22	.01302	22.98	.01393	27.40
$\hat{\rho}^2$	.06		.50		.13	
$\hat{\phi}$	.05		.47		.13	

Results for Composite B (i.e., Problem 4 and 6: Only Description and Explanation Score) indicated that the major source of measurement error was

problem (47.32 percent, averaged across the three classes). We interpret this to mean that the two problems considered in this composite are tapping different aspects of the students' performance. Indeed, mean scores were higher for Problem 4 (i.e., explaining the relation between sun shadows and earth rotation; $\overline{X}$ = .48 averaged across classes), than Problem 3 (i.e., explaining similarities and differences between shadow on the Northern and Southern Hemispheres; $\overline{X}$ = .15 averaged across classes). Not surprisingly, the next largest variance component was for the interaction student x problem (35.68 percent, averaged across the three classes).

We will use these results to revise the assessment. Among other things, we need to decide whether the performance component, Description/Explanation, should be eliminated or if we need to modify both the questions and the scoring form.

Sensitivity to Instruction. Table 6 presents the problem and total mean scores and standard deviations for the three classes. The direction of the difference between scores--which are low even for the two groups that received instruction, reflects the differences in instruction. We conducted a One-Way ANOVA to determine the statistical significance of the difference between the total scores. The ANOVA revealed statistically significant differences (F= 5.88; p <.005); a Turkey's HSD test indicated that the significant difference was only between Classes 1 and 3. No significant difference was observed between Classes 2 and 3, despite the fact that Class 2 received instruction. Daytime Astronomy was not consistently sensitive to differences due to instruction.

Table 6

Mean and Standard Deviations for Problem and Total Scores by Class in the Daytime Astronomy Assessment.

	Class 1-Instruction		Class 2-Instruction		Class 3-No Instruction	
	Mean	S.D.	Mean	SD	Mean	SD
Problem 1 (Max = 3)	1.16	.66	1.09	.50	.69	.30
Problem 2 (Max = 3)	1.23	.47	.98	.38	.89	.52
Problem 3 (Max = 3)	1.24	.62	1.15	.47	1.25	.54
Problem 4 (Max = 1)	.57	.27	.49	.26	.38	.25
Problem 5 (Max = 3)	.99	.52	.89	.34	.81	.35
Problem 6 (Max = 1)	.21	.12	.12	.14	.12	.09
Total Score (Max = 14)	5.40	1.44	4.73	.98	4.14	.94

Summary and Conclusions

We have described a conceptual framework for conceiving science performance assessments that modestly recreate the conditions in which scientists work and that are intended to elicit from students the activities and thinking of scientists when they solve problems. According to this framework, science performance assessments can be classified by types of science investigations and performance on those assessments can be scored based on the scientific defensibility of the approaches used by students.

The defensibility of classification activities (which may involve a process, an end result, and an application) lies in the relevance of the dimensions involved in the classification; therefore, performance on a classification investigation assessment can be scored based on the dimensions used by students to classify. The defensibility of observation activities lies in the accuracy of the data collected and the models used to collect and interpret those data; therefore, performance on an

observation investigation assessment can be scored based on the accuracy of the data obtained, the models used, and the explanations provided by students.

We constructed two assessments to explore the nature of the classification and observation tasks in our framework. Sink and Float, a classification assessment, consisted of four problems, each intended to address a different type of knowledge. Daytime Astronomy, an observation investigation assessment, consists of six problems. The components across those problems are intended to address different aspects of observation skills.

We presented some findings obtained from administering these assessments to fifth-grade students. We found reasonably high interrater reliabilities for both assessments. Based on score variability, we found that the problems presented in Sink and Float and the components of Daytime Astronomy distinguished different aspects of knowledge within the domain addressed by each assessment. Finally, we found no evidence that Sink and Float and Daytime Astronomy were sensitive to differences in instruction.

Regarding sensitivity to instruction, the results suggest that the classification problems of Sink and Float were too easy for the students. We are currently analyzing data obtained with fourth-grade students and fifth-grade students with different curricular experiences. If we find no significant score differences due to instruction among fourth-graders, but significant score differences between them and the fifth-graders used in this investigation, we might conclude that the classification problems are not relevant to the content of flotation. For Daytime Astronomy, the lack of sensitivity to differences due to instruction is unclear to us. We believe that more studies are need before any definite conclusion can be given.

Based on the experience developing the scoring form for Daytime Astronomy, it seems that the observation investigation assessments address a very elusive kind of knowledge. Although we obtained reasonably high interrater

reliabilities, from a practical standpoint the scoring form for this assessment still has to be improved to make it simpler and quicker to use.

Needless to say, additional development work and research is needed before classification and observation performance assessments are ready for "prime time."

References

Aldenderfer, M. S. & Blashfield, R. K. Cluster analysis. Quantitative Applications in the Social Sciences, Vol. 44. Thousand Oaks: Sage University Paper.

Bailey, K. D. (1994). Typologies and taxonomies. An introduction to classification techniques. Quantitative Applications in the Social Sciences, Vol. 102. Thousand Oaks: Sage University Paper.

Baxter, G. P. & Elder, (1994). On the use of embedded assessments to support learning in elementary science classrooms. Unpublished manuscript. University of Michigan.

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

Bunge, M. (1967). Scientific Research II. The Search for truth. New York, NY: Springer-Verlag New York Inc.

California Assessment Program (1992). Science performance field test, grade 5. Sacramento, CA: California Department of Education.

Carin, A. A. (1993). Teaching Science Through Discovery, Seventh Edition. New York: Macmillan Publishing Company.

Druker, S., Solano-Flores, G., Brown, J. H., & Shavelson, R. J., (1996, April). A comparison of two approaches to score science performance assessments. Paper presented at the AERA annual meeting. New York, NY.

Hamilton, E. (1994). Daytime Astronomy: Teacher's Guide. Project SEED Office. Pasadena, CA.

Hesse, M. B. (1963). Models and analogies in science. London: Sheed and Ward.

National Science Resource Center (1994). "Floating and Sinking": Instructional kit for fifth grade. North Carolina: Biological Supply Company.

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessment: An update. Journal of Research in Science Teaching, 33(10), 1045-1063.

Shavelson, R. J. (1995). On the development of science performance assessments technology. Unpublished manuscript. Stanford University. Stanford, CA.

Shavelson, R. J, & Solano-Flores, G. (1997). Toward a science performance assessment technology. Manuscript submitted for publication. Educational Researcher.

Shavelson R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-232.

Shavelson, R. J., Baxter, G. P., Pine, J. , Yuré, J., Goldman, S. R., & Smith, B. (1991). Alternative technologies for large scale science assessment: Instrument of education reform. School Effectiveness and School Improvement, 2(2), 97-114.

Sokal, R. R. & Sneath, P. H. A. (1963). Principles of numerical taxonomy. San Francisco: W. H. Freeman and Company.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., and Bachman, M. (1994, April). Development of an item shell for the generation of performance assessments in physics. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

AREA 1997

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE
(Specific Document)

ERIC

TM027375

I. DOCUMENT IDENTIFICATION:

Title: ON THE DEVELOPMENT AND SCORING OF CLASSIFICATION AND OBSERVATION SCIENCE PERFORMANCE ASSESSMENTS	
Author(s): GUILLERMO SOLANO-FLORES, R.J. SHAVELSON, M.A. RUIZ-PRIMO, S.E. SCHULTZ, E.W. WILEY & J.H. BROWN	
Corporate Source: PAPER PRESENTED AT THE ANNUAL MEETING OF THE AMERICAN EDUCATIONAL RESEARCH ASSOCIATION CHICAGO, IL	Publication Date: APRIL 24-28, 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here
Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature:	Position: SENIOR RESEARCH ASSOCIATE
Printed Name: GUILLERMO SOLANO-FLORES	Organization: WESTED
Address: WESTED 12345 EL MONTE ROAD LOS ALTOS HILLS CA 94022-4594	Telephone Number: (415) 949-8466
	Date: 6/2/97

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

February 21, 1997

Dear AERA Presenter,

Congratulations on being a presenter at AERA[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a printed copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the AERA Conference. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality. You can track our processing of your paper at http://ericae2.educ.cua.edu.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (523)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the AERA web page (http://aera.net). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

---

[1]If you are an AERA chair or discussant, please save this form for future use.