Rhetoric and Reality in Science Performance Assessments: An Update

Maria Araceli Ruiz-Primo and Richard J. Shavelson
School of Education, Stanford University, Stanford, California 94305-3096

Abstract

This article addresses the rhetoric of performance assessment with research on important claims about science performance assessments. We found the following: (a) Concepts and terminology used to refer to performance assessments often were not consistent within and across researchers, educators, and policymakers. (b) Performance assessments are highly sensitive not only to the tasks and the occasions sampled, but also to the method (e.g., hands-on, computer simulation) used to measure performance. (c) Performance assessments do not necessarily tap higher-order thinking, especially when they are poorly designed. (d) Performance assessments are expensive to develop and use; technology is needed for developing these assessments in an efficient way. (e) Performance assessments do not necessarily have the expected positive impact on teachers' teaching and students' understanding. (f) If teachers are to use performance assessments in their classrooms, they need professional development to help them construct the necessary knowledge and skills. This article attempts to address some of these realities by presenting a conceptual framework that might guide the development and the evaluation of performance assessments, as well as steps that might be taken to create a performance assessment technology and develop teacher inservice programs.

In the last few years, traditional multiple-choice achievement tests have been criticized on numerous grounds and authentic assessments have been proposed as important alternatives. Of the variety of authentic assessments proposed, we focus on performance assessment—an assessment that provides students with laboratory equipment, poses a problem, and allows students to use these resources to generate a solution. The virtues of performance assessments have been heard over and over by state legislators, educators, and the public in general. However, the reality of performance assessment may not have the immediate promise that was expected some years ago (Shavelson, Baxter, and Pine, 1991). This article shares the findings of our research group on science performance assessments with the goal of tempering rhetoric with research reality and encouraging our fellow researchers to address assessment challenges.

Science Performance Assessments: The Rhetoric

Traditional multiple-choice achievement tests in science have been criticized in several ways (e.g., Shavelson, Carey, & Webb, 1990). Despite their efficiency (economical to develop, administer, and score) they do not measure some aspects of knowledge that are valued in science education: for example, the ability to formulate a problem or carry out an investigation. Hence
multiple-choice tests are limited in capturing students' conceptual understanding and problem-solving skills; they are limited in their very nature by the requirement to select, not produce, a response; they do not look like the science conducted in the laboratory or the field, and consequently may provide only limited information about what students know and can do in science. Finally, they lead teachers to teach a multitude of often unrelated facts rather than conceptual and procedural understanding. Performance assessments have caught public attention in the past years as a complement to multiple-choice tests. Performance assessments are assumed to tap higher-order thinking processes and be more directly related to what students do in their classroom and what scientists actually do—observe, hypothesize, record, infer, and generalize. They are also assumed to be a useful policy instrument for science curriculum reform. If teachers teach to the test, performance assessments can signal to teachers new instructional goals and new conceptions of teaching.

The promising advantages of performance assessments led educators and policy makers to assume that the evaluation and monitoring of educational progress can be enhanced by this new testing idea. However, before performance assessments are used widely, we need to know more about them. Is there a conceptual sharing among researchers, teachers, and advocates about what a performance assessment is? Can performance assessments meet the technical standards of reliability and validity? Can performance assessments meet the utility and practicality standards that teachers demand? Can they be developed with the same efficiency as multiple-choice tests? Do they promote high-order skills? Do they have a positive impact on curriculum, teaching, and students as expected? Do teachers know how to use and select performance assessments? Research findings on these issues are discussed below.

Science Performance Assessments: The Reality

In this section we review what we have learned, largely from our research, trying to answer these questions.

Is There a Conceptual Sharing among Researchers, Teachers, and Advocates about What a Performance Assessment Is?

If performance assessments are to be promoted and used as a centerpiece in science curriculum and assessment reform, we must take the time and effort to clarify the concepts and the terminology to delineate better what a performance assessment is (e.g., Baker, O'Neil, & Linn, 1993; Worthen, 1993). In other words, researchers need to construct a framework that can help one another and other educators not only to use a common terminology, but also to guide the development and evaluation of performance assessments. To begin addressing this need, we present one possible conceptual framework for science performance assessments.

Performance Assessment Components. We conceive of a science performance assessment as a combination of (a) a task that poses a meaningful problem and whose solution requires the use of concrete materials that react to the actions taken by the student; (b) a format for the student's response; and (c) a scoring system that involves judging not only the right answer, but also the reasonableness of the procedure used to carry out the task. Without all three, a performance assessment is undefined (Figure 1).
Performance Assessment Components

- Task
  - Invites students to solve a problem or conduct and investigation
  - Requires the use of concrete materials that react to the actions of students
  - Provides feedback to students on their actions

- Response Format
  - Provides opportunities to record findings
  - Allows students to decide how to summarize findings
  - Asks students to justify answers

- Scoring System
  - Reflects the goals for which the task was selected
  - Captures "right" answer and reasonableness of procedures or evidence
  - Provides insight about the students' problem solving procedures and errors

Figure 1. Performance assessment components.

Consider the Bugs and the Electric Mysteries performance assessments (e.g., Shavelson et al., 1991) as examples to instantiate the triple. With the Bugs investigation (Figure 2), elementary students are asked to find out in which environments bugs reside by carrying out investigations with equipment provided (the task). The student is asked to draw a conclusion about the environment in which the sow bugs reside and describe the procedures used to carry out the investigation (the response). The student's performance is scored on the basis of the procedure used and the conclusion drawn from the results of the investigation (the scoring system).

In Electric Mysteries (Figure 3), fifth and sixth graders are asked to find out what is inside six mystery boxes by hooking up batteries, bulbs, and wires (the task). The box may contain two batteries, a battery and a bulb, a bulb, a wire, or nothing at all. Students are asked to conclude what is inside each box and provide the circuit that helped them find out (the response). Students' performance is scored on the basis of the evidence presented by them and the conclusions drawn from that evidence (the scoring system).

Performance Assessments Methods. We have used different measurement methods to collect information on students' performances: direct observation, notebooks, computer simulation, short-answer and multiple-choice tests (Baxter & Shavelson, 1994; Shavelson et al., 1991; Shavelson & Baxter, 1992). With direct observation, an observer records a student's performance and response (the procedure the student uses in the bugs investigation for determining choice, or the external circuit the student constructs in the mystery box investigation) as the student proceeds with the investigation. This method is considered to be the ideal measurement method or benchmark. Other alternative methods, considered to be surrogates, were developed to reduce cost and human resources. These methods are (a) notebooks, in which students recorded their procedures and conclusions; (b) computer simulations of the hands-on investiga-
Bugs Investigation

Task

You have some sow bugs and some equipment.

Your tasks are to answer these three questions:
1. When given a choice, do sow bugs choose light or dark?
2. When given a choice, do sow bugs choose damp or dry?
3. When given a choice, do sow bugs choose:
   a) damp and dark?
   b) damp and light?
   c) dry and dark?
   d) dry and light?

Response Format

Notebook for Experiment 1:
A. In the space below, draw a picture of your experiment. Make sure you show each of the following:
1. How you set up the experiment
2. the number of bugs you used
3. where you put the bugs to start the experiment

B. Steps in Experiment 1. Please number each step in order.
Step | What you did

C. How did you know from experiment 1 whether sow bugs choose light or dark?

Scoring System

Sow Bugs Score Form

Student __________________ Observer __________________ Score #1: __________

1. Method
   A. Two conditions: one dish
   B. One condition at a time

2. Control of Manipulation
   A. Equal area for each condition
   B. Number of bugs
   C. Starting location of bugs
   D. Sufficient time

3. Care in handling of bugs (i.e., by, down, or under):
   Yes
   No

4. Determine Result
   A. Count the number of bugs in each location
   B. Observe how busy the bugs are
   C. Other __________________

5. Result (logically follows from 4A)
   Yes
   No

<table>
<thead>
<tr>
<th>Grade</th>
<th>Method</th>
<th>Area</th>
<th>Control of Bugs</th>
<th>Manipulation Location</th>
<th>Time</th>
<th>Case</th>
<th>Determine</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>IA</td>
<td>Yes</td>
<td>6-5</td>
<td>Middle/Large</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>4A</td>
</tr>
<tr>
<td>B</td>
<td>IA</td>
<td>Yes</td>
<td>12</td>
<td>Middle/Large</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>4A</td>
</tr>
<tr>
<td>C</td>
<td>IA</td>
<td>Yes</td>
<td>12</td>
<td>Any</td>
<td>Yes/No</td>
<td>Yes</td>
<td>Yes</td>
<td>4A</td>
</tr>
<tr>
<td>D</td>
<td>IA</td>
<td>No</td>
<td>12</td>
<td>Any</td>
<td>Yes/No</td>
<td>Yes</td>
<td>Yes</td>
<td>4A</td>
</tr>
<tr>
<td>F</td>
<td>IB</td>
<td>Yes</td>
<td>Any</td>
<td>Any</td>
<td>Yes/No</td>
<td>Yes/No</td>
<td>Yes/No</td>
<td>4A or 4B</td>
</tr>
</tbody>
</table>

Figure 2. Bugs investigation assessment components.
Electric Mysteries Investigation

Task

Find out what is in the six mystery boxes A, B, C, D, E and F. They have five different things inside, shown below. Two of the boxes will have the same thing. All of the others will have something different inside.

- Two batteries:
- A wire:
- A bulb:
- A battery and a bulb:
- Nothing at all:

For each box, connect it in a circuit to help you figure out what is inside. You can use your bulbs, batteries and wires any way you like.

When you find out what is in a box, fill in the space on the following page.

Response Format

Name: _______________________

Box A: Has ______________________ inside.

Draw a picture of the circuit that told you what was inside Box A:

[Diagram of circuit]

How could you tell from your circuit what was inside Box A?

__________________________________________________________

Box B: Has ______________________ inside.

Draw a picture of the circuit that told you what was inside Box B:

[Diagram of circuit]

How could you tell from your circuit what was inside Box B?

__________________________________________________________

Scoring System

Electric Mysteries End of Unit Assessment Score Form

<table>
<thead>
<tr>
<th>Mystery Box</th>
<th>What's Inside</th>
<th>Circuit</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>two batteries</td>
<td>![Image of circuit]</td>
</tr>
<tr>
<td>B</td>
<td>bulb</td>
<td>![Image of circuit]</td>
</tr>
<tr>
<td>C</td>
<td>wire</td>
<td>![Image of circuit]</td>
</tr>
<tr>
<td>D</td>
<td>battery</td>
<td>![Image of circuit]</td>
</tr>
<tr>
<td>E</td>
<td>nothing</td>
<td>![Image of circuit]</td>
</tr>
<tr>
<td>F</td>
<td>wire</td>
<td>![Image of circuit]</td>
</tr>
</tbody>
</table>

Score: _______________________

Total: _________

Figure 3. Electric mysteries assessment components.
tions; and (c) paper and pencil problems in planning, designing, and/or interpreting experiments based on the hands-on investigations (i.e., short-answer and multiple-choice tests).

A Sampling Framework for Science Performance Assessments. Performance assessments can be viewed as a concrete task, with its corresponding response format and scoring system (e.g., Bugs) performed by a student on a particular occasion (e.g., third week of April) and scored by a rater (e.g., a teacher) who judges the student's performance based on the procedure used by the student and the accuracy of the response (Shavelson, Baxter, & Gao, 1993; Shavelson, 1995a). The measurement method might be hands-on, notebook, computer simulation, or paper and pencil.

The tasks represent the content in a subject-matter domain, occasions represent all possible occasions in which a decision maker would be equally willing to accept a score on the performance assessment, raters represent all possible individuals who could be trained to score performance reliably, and methods represent all permissible performance assessment measures that a decision maker would be equally willing to interpret as bearing on a student's achievement (Shavelson et al., 1993; Shavelson, 1995a).

A student's performance can thus be conceived as a sample drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. Task, occasion, and rater traditionally have been considered as sources of measurement error. Measurement method, however, traditionally has been considered as a validity issue. When performance assessment scores differ from one measurement method to another, we speak of the lack of convergent validity owing to method sampling variability (Baxter & Shavelson, 1994).

Based on this sampling framework for performance assessments, we present a short review of the evidence provided by research on reliability and validity.

Can Performance Assessments Meet the Technical Standards of Reliability and Validity?

New evidence has accumulated in the last couple of years on reliability and validity issues. Here we focus on these technical issues with some good and bad news.

Reliability. Research findings on rater sampling, or consistency of scores across raters, are positive. Raters can be trained to reliably evaluate student performance (Shavelson et al., 1991; Ruiz-Primo, Baxter, & Shavelson, 1993; Solano-Flores, Jovanovic, & Shavelson, 1994). However, high reliability coefficients may mask important disagreements among raters, and these disagreements may lead to erroneous decisions about a student's performance. Sometimes there are raters who, even after training, do not perform well. Consequently, when a single rater is used, as is the case in some statewide assessments, unreliability due to raters may still be problematic. To address this issue, training raters is not enough. Their performance must be continually monitored and calibrated; those who do not perform well should not participate in scoring (cf. Wigdor & Green, 1991).

Findings on task sampling are also consistent but not quite as optimistic. Task sampling variability is a major source of measurement error in large-scale assessment at both individual and school levels (Shavelson et al., 1993; Gao, Shavelson, & Baxter, 1994). We have found a strong person by task interaction, and a moderate school by task interaction. The bottom line is that a substantial number of tasks is needed to reliably estimate student and school level performance.
At the individual level, 8 to 23 tasks would be needed to reach an approximate reliability of .80 (relative G coefficient) (Shavelson et al., 1993; Gao et al., 1994). At the school level, a tradeoff between the number of students and tasks should be considered if we want to rank order schools consistently. For example, to achieve a coefficient of .80, 50 students and 15 tasks or 100 students and 12 tasks may be needed (Shavelson et al., 1993). In a second study, Gao et al. (1994) found that as few as 7 tasks may be needed for a sample of 25 students. Still, large variation among students within a school creates uncertainty about school level performance.

The impact of task sampling on time, cost, and human resources is substantial: It may take about 2.25 hours of testing to obtain a generalizable measure of student achievement if we consider 7 tasks of 20 minutes each. Moreover, these tasks are costly and time-consuming to develop. A technology for generating tasks is needed.

Finally, occasion sampling has rarely been studied despite its importance for understanding the reliability of performance assessments. One study on this issue suggests that occasion-sampling variability may be as large as task-sampling variability (Ruiz-Primo et al., 1993). Indeed, occasion sampling may even be confounded with task sampling because assessments are usually given on one occasion (Cronbach, Linn, Brennan, & Haertel, 1995). Simply put, students' performance varied from one occasion to the next, even though no further instruction or practice was received. However, educators, policy makers, and the public will generalize a student's performance score earned on one occasion to a whole set of possible occasions.

To increase the dependability of this generalization, a substantial number of tasks or more occasions need to be sampled. Again, either strategy for increasing reliability will be costly and time-consuming. Nevertheless, the stakes in large-scale assessment seem to warrant the expense. Moreover, as this new technology is refined, cost and time should decrease (Shavelson, 1995a).

Validity. Research has shown that not all performance methods—direct observation, notebooks, computer simulation, and paper-and-pencil—are exchangeable. There is a large student by task by method interaction (Baxter & Shavelson, 1994). Students' scores depend on the particular tasks sampled and on the particular methods used to assess performance.

The paper-and-pencil method is least exchangeable with the direct observation benchmark ($r < .30$). In contrast, results suggest that notebooks are a reasonable, less expensive surrogate for direct observation ($r > .80$) (Baxter & Shavelson, 1994).

Computer simulations fall in between notebooks and paper and pencil as surrogates; they correlate moderately with observation and notebook scores ($r \sim .45$) (Baxter & Shavelson, 1994). Subsequent studies produced the same level correlation as did the original study, even though we attempted to improve the link between computer-generated icons and physical equipment (Ruiz-Primo, Solano-Flores, Brown, Druker, & Shavelson, 1994).

Whatever performance assessments are measuring about science understanding is highly sensitive not only to the task and occasion sampled, but also to the method used to assess performance. Results show that measurement methods seem to tap different aspects of science achievement. Each method may provide a different insight into what students know and can do.

These findings make evident the need to examine more carefully the technical qualities of performance assessments before their scores are reported to students, parents, and policy makers. Results also indicate that larger numbers of tasks are needed if we want to get generalizable measures of achievement.
Can Performance Assessments Meet the Utility and Practicality Standards that Teachers Demand?

If we want teachers to use performance assessments in their everyday practice, performance assessments have to compete with the efficiency with which multiple-choice tests can be administered and scored. The dilemma for teachers is clear: Preparation, administration, and scoring of multiple-choice tests, while requiring adaptation, is relatively straightforward. Performance assessments, in contrast, require more work before, during, and after the test. Scoring cannot be done from a right–wrong key; expert judgment is required to evaluate students’ performances.

Performance assessments thus should meet standards other than the usual technical ones. Two such additional standards are utility and practicality (Shavelson et al., 1994). Utility refers to the extent to which performance assessment scores provide useful information for monitoring instruction and student progress. Practicality refers to the extent to which performance assessments can be used in a classroom without excessive cost, effort, or disruption.

Our evaluation of the utility and practicality of science performance assessments is based on the administration and scoring of different investigations with hundreds of students (Shavelson et al., 1991; Ruiz-Primo et al., 1994; Solano-Flores, Ruiz-Primo, Baxter, & Shavelson, 1991; Solano-Flores et al., 1994). Information gained from this experience has taught us that for performance assessments to be useful for teaching, they need to (a) be linked directly to instructional units, and (b) have a well-designed scoring system that clearly reflects what students know and can do.

We also found that performance assessments can be managed as efficiently as hands-on science instruction, and that scoring can be easy and quick to learn and to use. How manageable performance assessments are thus depends on the instructional approach used in the classroom. We found that teachers who opt to use an activity-based curriculum in their classroom can administer performance assessments with no more difficulty than that associated with inquiry science. As with utility, the quality of the design of the scoring system will determine scoring efficiency. With well-designed scoring systems, we found that after a 45-minute scoring-training session, and after having practiced scoring about 10 students’ performances, scoring can be fast and easy.

In summary, the utility and practicality of performance assessments is a function, at least in part, of the quality of the performance assessment itself and of teachers’ experience with inquiry science. These findings suggest the need to find criteria for developing and selecting performance assessments.

Can Performance Assessments Be Developed Cost-Efficiently?

To date, we lack a performance assessment technology equivalent to that used to develop multiple-choice item banks. Although general guidelines have been provided for developing performance assessments (Baron, 1991; Wiggins, 1992), they are limited in their ability to generate tasks as efficiently as is done for multiple-choice tests. At present, performance assessments require an extensive iterative developmental process of review, revision, and tryout that takes months and is costly. However, our group is now attempting to develop a technology for designing this new form of assessment in a more efficient way.

Toward a Performance Assessment Technology. In research with our colleagues, we discovered regularities in the nature of science assessments built for large-scale testing (Shavelson et al., 1991; Solano-Flores et al., 1994) and for specific science curricula (Solano-Flores, 1994).
We have discovered (Shavelson, 1995a) a small number of different types of tasks that characterize a wide variety of science performance assessments (Figure 4): (a) **comparative**—compare two or more objects on some attribute while controlling other variables (e.g., Bugs); (b) **component identification**—determine the components that make up the whole (e.g., Electric Mysteries); (c) **classification**—create a classification scheme using attributes of a set of objects and a particular goal for classification (e.g., Rocks and Charts); and (d) **observation**—observe and systematically record an attribute of an object over a period of time (e.g., Daytime Astronomy). Of course, the framework includes an "other" category which represents the current state of our ignorance. Additional research may very well expand our category system or a new system may replace it. We think we may be on the right track; one reason is that each type of task can be identified with several areas of science (Shavelson, 1995a). For example, comparative tasks are frequently found in physics, component identification tasks in chemistry, and observation tasks in astronomy.

Regularities in scoring systems have also emerged in our work (Baxter, Shavelson, Goldman, & Pine, 1992). For example, the scoring system for comparative tasks is procedure-based—it focuses on the procedures used to solve the problem and the accuracy of the solutions (Figure 2). The scoring system for component identification is evidence-based—it focuses on the evidence provided to confirm or disconfirm the presence of a component and the accuracy of the response (Figure 3).

Notice that in Figure 4 we also distinguish between Analytic and Holistic scoring systems.

<table>
<thead>
<tr>
<th>Types of Scoring Systems</th>
<th>Types of Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comparative Investigation</td>
</tr>
<tr>
<td>Procedure Based</td>
<td>• Paper Towels</td>
</tr>
<tr>
<td>Evidence-Based</td>
<td>• Electric Mysteries</td>
</tr>
<tr>
<td>Dimension-Based</td>
<td>• Rocks &amp; Charts</td>
</tr>
<tr>
<td>Data Accuracy-Based</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
</tr>
<tr>
<td>Holistic</td>
<td></td>
</tr>
<tr>
<td>Rubric</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 4.* Types of investigations and scoring systems in performance assessments.
Analytic scoring systems focus on different components of a student's performance. For example, in the Bugs investigation (Figure 2), the scoring system focuses on how the student set up the experimental conditions, how she controlled different variables, how careful she was in handling the bugs, how she determined her results, and whether or not the solution was correct. Holistic scoring systems, in contrast, focus on overall judgments about the student's performance. Rubrics, for example, provide a set of statements that describe levels of student performance ranging from poor to outstanding. Sometimes these statements are accompanied with examples of actual student work at each score level. Figure 5 provides examples of the range of statements used to score performance holistically on a Leaves classification assessment.

This distinction among the types of scoring systems may help to (a) clarify conceptually the language used in performance assessment (e.g., rubric is not a synonym for a scoring system); (b) create a common terminology (e.g., analytic or holistic scoring system); and, most important, (c) modify thinking that performance can only be scored by holistic judgment. Matching the scoring system with assessment purposes (e.g., monitoring students' progress, instruction, or accountability) should be an item on the research agenda.

The importance of this classification of types of tasks and scoring systems lies in its ability to guide assessment development. Once we have decided to develop a comparative assessment, we need to know a lot about the structure of the task and the response format, and the nature of the scoring system. Using this schema, we have saved up to 6 months time in the development of performance assessments.

Of course, the development process is still long and costly if we compare it with multiple-choice testing. For this reason, we are looking for new, efficient ways to generate dependable performance assessments.

Solano-Flores et al. (1994) developed a blueprint or item shell for generating comparative tasks. The item shell is based on a specification of a knowledge domain. It prescribes the actions that test developers should perform to create an assessment in that domain. The item shell includes four science process skills (i.e., planning and designing, performing, analyzing and interpreting, and applying), and four levels of inquiry (no, low, medium, and high) from which to choose. The assessments generated with this item shell (one on friction and one on incline

<table>
<thead>
<tr>
<th>Rubric Scoring System</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Examples of three performance level statements)</td>
</tr>
</tbody>
</table>

**Outstanding...Rating = 4**
Gives complete answer with logical responses; uses complex attributes (beyond color and size); developing of a process of thinking is evident; provides both detailed and specific comparisons; rationale is clearly stated and consistent with grouping; demonstrate understanding of the concept; may include pictures in addition to written explanations; description match object; complete task.

**Competent...Rating = 3**
Gives complete answers with some logical responses; shows development of a process of thinking; sortings are clearly stated; descriptions are clear - may have only basic attributes; rationale is consistent with groupings; shows understanding of the concept; second page completed, however, only the first page is outstanding.

**Serious Flaws...Rating = 1**
Gives answers that are not complete or understandable; doesn't show understanding of the process of grouping (no grouping or illogical grouping); may group some leaves but doesn't provide explanation or rationale, questions #2 and #4 are not answered or answered inaccurately; task is not completed.

*Figure 5.* Examples of three statements that describe levels of student performance.
produced generalizable scores across raters and were sensitive to students' instructional experiences. Not surprisingly, the major source of measurement error was task sampling (some students performed better on Incline Planes and others on Friction).

Our goal with item shells is to enable educators to develop sound performance assessments. To test this hypothesis, Solano-Flores et al. (1994) asked an elementary school teacher, whose activity-based science teaching was considered exemplary, to develop a performance assessment using the comparative investigation item shell. The assessment developed by the teacher—the nutrition assessment—was successfully completed. Although the teacher was not familiar with the purpose of the research, the instructions provided by the item shell were clear enough to guide her actions. The assessment—the task, the response format, and the scoring system—was highly similar to the assessments developed by the researchers with the item shell (i.e., Friction and Incline Planes).

The results of using the item shell are promising. They suggest that item shells can contribute to the development of a performance assessment technology: If we construct item shells for each type of task linked with a corresponding type of scoring system, educators might be able to develop science performance assessments of the genre we have described more efficiently.

Do Performance Assessments Promote High-Order Skills?

The motto of performance assessment advocates has been that these assessments measure important aspects of higher-order thinking (Wiggins, 1989). Again, rhetoric may have surpassed reality. Precious little research has been conducted to support empirically the inference of higher-order thinking from observed performance.

One exception is the work of Baxter, Glaser, and Raghavan (1993). They linked performance scores with level and kind of reasoning and understanding, by analyzing verbal protocols, observing students' performances, evaluating students' notebooks, and examining scoring systems. Baxter et al. focused on the extent to which the two different assessment tasks—a Exploring the Maplecorder and Electric Mysteries investigation—provided students with the opportunity to engage in higher-order thinking and whether the scoring systems reflected differential performance of students with respect to the nature of the cognitive activity.

Their findings suggested that performance assessments may not necessarily lead to higher-order thinking. Characteristics of the assessment task hold the key. Tasks that provided procedural instructions to students may not allow them to show what they know and can do. Also, scoring systems that were inconsistent with the characteristics of the task, and did not reflect students' meaningful use of knowledge and problem-solving procedures, tended not to capture higher-order thinking. In one of the tasks analyzed, for example, the authors concluded that even though the task of exploring the maplecorder was rich in providing students opportunities to engage in higher-order thinking, the scoring system failed to capture the essence of proficient students' performances. The scoring system ignores the scientific validity of the way students manipulate the variables; it considers only the number of variables used by students.

Even though performance assessments intend to measure higher-order thinking processes, poorly designed assessments do not guarantee that such thinking processes occur. More research linking performance assessment scores with students' cognitive processes is needed before we can be sure that performance assessments measure higher-order thinking.

If performance assessment tasks are to measure higher-order thinking, the assessment should permit students to demonstrate what they know and can do, rather than direct their performances. Assessment structure (i.e., task and response format structure) then becomes an important issue in the design of performance assessments. Assessment structure may be thought
Assessment Structure

<table>
<thead>
<tr>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree of Structure</td>
<td></td>
</tr>
<tr>
<td>(Recipe Mode)</td>
<td>(Discovery Mode)</td>
</tr>
<tr>
<td>Defining the Problem</td>
<td>Provides students with a problem</td>
</tr>
<tr>
<td>Planning &amp; Conducting the Investigation</td>
<td>Provides students with a plan for solving a problem and with step-by-step procedures for conducting the investigation</td>
</tr>
<tr>
<td>Analyzing &amp; Representing Data</td>
<td>Tells students how to transform, represent, synthesize, etc. results</td>
</tr>
</tbody>
</table>

*Figure 6. Assessment task structure.*

of as a continuum (Figure 6). An assessment can be more or less structured according to what we want to know about students' performances. A highly structured assessment—one that provides students step-by-step instructions to follow—will tell us more about how good students are at following directions than about what they know and can do in science. The information students provide is thus limited by task-structure demands. In contrast, a low-structured assessment will reveal more about what students know and can do because the problem to be solved—one faced in daily scientific research—is to determine the procedures for the investigation. That is, the problem to be solved is giving structure to the investigation. Incidentally, the levels of inquiry of the item shell described previously draws heavily on this notion of assessment structure.

*Do Performance Assessments Have a Positive Impact, as Expected, on Teachers, Curriculum, or Students?*

*Impact on Teachers.* Performance assessments have been hailed as an instrument of reform (Resnick & Resnick, 1992). If performance assessments are well designed, they can provide new conceptions of science teaching for classroom teachers (e.g., activity-based science teaching) (Wiggins, 1989) and new information to monitor their teaching. If teachers teach to the test, the use of performance assessments can lead teachers to teach science in the direction of science curriculum reform.

Little research has been done to evaluate empirically the effect of performance assessments on teachers and science classrooms. Druker and Shavelson (1994) examined four teachers' responses to three of California's science education reform policy instruments: the Science Framework for California Public Schools, the California Science Implementation Network (CSIN), and the pilot phase of the California Assessment Program (CAP). From interviewing teachers, observing them in their classroom, and analyzing artifacts (e.g., homework, worksheets), they concluded that teachers' beliefs about teaching and learning determined how they received, interpreted, and used the information provided by the three policy instruments. The expectation that performance assessments will enhance exemplary teaching for teachers does
not recognize that teachers' perspectives will influence what parts of the testing situation they will choose to model.

The results obtained by Druker and Shavelson (1994) also indicated that some teachers had made changes in the direction that the science reform intended; for example, they did more activity-based teaching. However, for 3 of the 4 teachers, no fundamental changes were observed from their rather traditional, didactic teaching styles to a more constructivist approach. The bottom line is that teachers who use performance assessments and adapt their teaching in the manner that the reform intends are likely to be those who already have beliefs about teaching and learning aligned with the reform. Much more is needed to change beliefs and performance of other teachers.

Impact on Curriculum Development. Research has not evaluated how performance assessments affect the development of science curriculum. However, our experience with curriculum developers may provide some insight (Shavelson, 1995b).

A team of science assessment developers from the University of California, Santa Barbara, was invited to a Center for Science Curriculum Development to demonstrate the process of performance assessment development. The curriculum developers chose Soil, one of the units they were developing, to instantiate the assessment development process. Our team provided an activity-based workshop that helped the curriculum developer conceptualize performance assessments (e.g., types of tasks, response formats, and scoring systems) and experience the performance assessment development process (Figure 7).

Three assessment development teams were formed with the center staff. The teams went through the first four steps of the process applying it to develop an end-of-unit assessment for soil. After 2 days, and even without going through the whole process of development, the center staff recognized that the unit needed revision.

The identification of the unit's goals, a critical step in the development process, helped the teams recognize that the list of objectives at the beginning of each lesson was not enough to assure that they were pursuing the same goal, or even to identify what the goal was. Disagreements among teams on what the goal was and why they included that goal in the unit were discussed and an agreed-upon goal emerged in the process.

Through the development process the goal became clearer for the teams. Soon the advan-

![Performance Assessment Development Process](image)

**Figure 7.** Performance assessment development process.
tages of having a well-defined goal were evident. As the development process evolved, curriculum developers questioned the appropriateness of one lesson, and recognized that others needed to be fine-tuned to focus on the knowledge and skills embodied by the unit’s goal and the embedded and end-of-unit assessments. With a clearer vision of the unit’s goal, the relevance of activities to the unit was questioned, activities were modified, and even the sequence of activities was changed.

We believe that there are good reasons to think that performance assessments may have a positive impact on curriculum development. However, “whether the romance between assessment and curriculum will lead to marriage or separation remains an open question. As with other romances, initial infatuation gives way to reality” (Shavelson, 1995b, p. 73).

**Impact on Gender Differences.** When multiple-choice tests are used to measure science achievement, females tend to perform more poorly than males. Rhetoric of performance assessment asserts that this new form of assessment may eliminate gender bias in testing (Jenkins & MacDonald, 1989). The reasoning goes as follows: Because performance assessments focus more on problem solving and critical thinking than recall of discrete pieces of knowledge for selecting 1 answer from 4 alternatives, female disadvantages may disappear. Performance assessments are supposed to give females the opportunity to show what they know and can do in science. Once again, empirical evidence to support this belief is scarce.

Jovanovic, Solano-Flores, and Shavelson (1994) examined gender differences in science performance assessments in several content areas. Results show that on 5 of the 10 performance assessments used, no gender effect was found. Differences on the other 5 assessments seemed to be related to the particular science content that was assessed. Males had an advantage over females on those assessments related to electricity. Females performed better on those assessments related to classification activities (e.g., classifying leaves and rocks). Jovanovic et al. concluded that performance assessments themselves are not going to eliminate gender differences in science achievement; the issue is more complicated. Students’ prior experiences, their interaction with instruction, and teachers’ preconceptions all play an important role in testing students’ performances.

**Impact on Language Minority Students.** August, Hakuta, and Pompa (1994) and Hakuta & Valdés (1994), among others, argue that limited English-proficient (LEP) students should be assessed in their native language to find out what they know and can do in science. Thus, the development of performance assessments written in the LEP students’ native language has been recommended.

However, only a few studies have been conducted to evaluate whether these assessment modifications make a difference on students’ performances (Solano-Flores et al., 1991; Solano-Flores, Othman, & Shavelson, 1994; see also Kopriva, Lowrey, & Martois, 1994, in mathematics and writing). Solano-Flores et al. did not find significant differences in science achievement due to assessment format (i.e., English-only and English-and-Spanish version). However, in both studies restricted score range at the lower end of the scale and small samples in some cells led to a lack of statistical power.

Although findings are inconclusive, they suggest that the problem is more complex than offering the bilingual version of the assessment. After reviewing the LEP students’ notebooks, they concluded that many students were not literate in either languages. For them, bilingual versions did not make a difference in their performance. The systematically low performances also suggest that students have not acquired the science knowledge and skills assessed. The
problem thus goes beyond the type of assessment used. It reflects instructional and social problems that cannot be masked by changing or modifying the assessments.

Do Teachers Know How to Select and Use Performance Assessments?

If we want teachers to align their assessment practices with science curriculum reform, they need to have knowledge about these new assessment practices and skills in their applications. How can science teachers be helped to learn and own new assessment practices? How can they be helped to use performance assessments in their classrooms?

We believe that the success of the assessment reform depends to a large extent on a sustained program of inservice education, not just political rhetoric (Hurd, 1986; Shavelson, Copeland, Baxter, Decker, & Ruiz-Primo, 1994). With this idea in mind we developed an inservice education program which aimed to provide teachers with knowledge and skills needed to understand, select, and use performance assessments (Shavelson et al., 1994). We provided teachers with activity-based experiences that give them some insight into the nature of performance assessments (e.g., task, response format, and scoring system, performance assessments methods, types of performance assessments and scoring systems), how to judge good performance assessments (i.e., reliability, validity, utility, and practicality), and how to use them in their classroom (e.g., embedded and end-of-unit performance assessments).

After implementing the program more than 10 times, we learned that many teachers think of performance assessments as anything that asks students to manipulate materials (the hands-on part of the assessment). Not surprisingly, the characteristics of the task demands, the response formats, and the scoring system are not considered in judging the quality of the assessments (the mind-on or -off part of the assessment).

Through the inservice program we introduced, explained, and illustrated with data from research the technical and practical qualities by which performance assessments can be evaluated. Figure 8 presents an overview of criteria for teachers to use when selecting performance assessments. Notice that these criteria focus on questions that teachers find closely related to their everyday practice.

Without extensive inservice education and high-quality instrumentation, little can be changed in the quality of instruction and assessment. Therefore, we need inservice programs that help teachers to construct the knowledge and skills necessary to select and use performance assessments.

Conclusions

The purpose of this article was to counter rhetoric with research on important claims about performance assessments. Specifically, we provided information on the following questions: Is there a conceptual sharing among researcher and advocates about what a performance assessment is? Can performance assessments meet the psychometric standards of reliability and validity? Can performance assessments meet the utility and practicality standards that teachers require? Can they be developed with the same efficiency as multiple-choice tests? Do they promote high-order skills? Have they a positive impact on curriculum, teaching, and students, as expected? Do teachers know how to select and use performance assessments?

Research does not provide the same picture as rhetoric. Indeed, our review points to the need to test all assertions made about performance assessments. Our review also points to the need for a long-term research agenda to develop a performance assessment technology for use in curriculum, and for use in state and national examinations. We need to (a) examine what factors
Selection of Performance Assessments

<table>
<thead>
<tr>
<th>Reliability</th>
<th>Are the scores provided by the assessment consistent across raters or tasks? A desirable range for a reliability coefficient is .80 to 1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Do the tasks and procedures used in the assessment overlap with what I am currently teaching?</td>
</tr>
<tr>
<td>Utility</td>
<td>Do the assessment scores provide helpful information in evaluating performance of a student or class of students?</td>
</tr>
<tr>
<td></td>
<td>Do the assessment scores provide useful information for monitoring instruction?</td>
</tr>
<tr>
<td></td>
<td>Do the assessment scores provide information helpful to the students in judging the quality of their own performance?</td>
</tr>
<tr>
<td>Practicality</td>
<td>Is the assessment cost effective in terms of time and money?</td>
</tr>
<tr>
<td></td>
<td>How long does it take to set up, administer, and clean up the task?</td>
</tr>
<tr>
<td></td>
<td>How long does it take to score each student?</td>
</tr>
</tbody>
</table>

Figure 8. Some questions for teachers to ask to judge the quality of performance assessments.

may contribute to the large sampling variability and how to reduce it; (b) look for other methods of measurement, besides performance assessments, to be used as indicators of students’ science achievement (e.g., concept maps) (Ruiz-Primo & Shavelson, 1996); (c) develop a performance assessment technology; (d) study more closely the impact of performance assessments on students, teachers, curriculum, and the public; and (e) develop and implement inservice programs that help teachers to become more acquainted about performance assessment.

The good news is that research is helping to developing a performance assessment knowledge base, perhaps even a technology, that will help to overcome some of the problems we are facing now. We hope that others will extend the breadth and depth of research described here.

Research reported in this article was supported in large part by a National Science Foundation (NSF) grant (TPE 90-55443) and conducted by a team of researchers including Gail P. Baxter, Guillermo Solano-Flores, Xiaohong Gao, Jasna Jovanovic, Stephen Druker, Abdul Othman, Janet Brown, and Jerry Pine. The findings and interpretations reported are those of the authors and do not necessary reflect those of the NSF or their colleagues.

Notes

1 Performance assessments in other subject domains can be described by this triple combination, but the task, response format, and scoring system definition will vary across subject matters.
The performance assessments described here are influenced greatly by the work of Black and colleagues (Black, 1990).

Baxter et al. (1993) analyzed three different assessment tasks. However, only two of those were performance-based assessments; the third was a conceptual integration task that required students to write an essay on a science topic.

Specifically, we asked the groups to address two questions: Why are we teaching this unit? and What do you want students to know and be able to do when they have completed the unit?

References


modifications in a special study of CLAS, the California Learning Assessment System. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.


Received July 17, 1995
Revised June 21, 1996
Accepted July 8, 1996