What We've Learned About Assessing Hands-On Science

Assessing scientific inquiry is more complex than political rhetoric pushing performance tasks indicates, this team of scientists found. And, unless carefully crafted and blended into science instruction, assessments alone are unlikely to boost achievement.

RICHARD J. SHAVELSON AND GAIL P. BAXTER

Over the past three years, our team of researchers, scientists, and science teachers at the University of California, Santa Barbara, the California Institute of Technology, and the Pasadena Unified School District has sought to create assessments that support good science teaching at the elementary level (Baxter et al. in press; Shavelson et al. 1991). In particular, our goal has been to develop activities that permit students to pursue an experimental inquiry focusing on process skills (such as observing and inferring) and construction of new knowledge (such as understanding the effects of insulators on electric current). In our definition, assessments consistent with good teaching invite students to perform concrete, meaningful tasks such as a laboratory experiment to determine, for example, how certain kinds of insects respond to changes in environment. Scoring of performance focuses on the reasonableness of the procedure used to carry out the investigation, not just on the "right answer."

In our work, we assumed that the ideal assessment would be direct observation of a student pursuing a scientific inquiry with laboratory equipment and materials. This observation would be made by scientists and science teachers trained to score performance in real time. That is, the ideal was predicated on the assumption of the symmetry of teaching and testing: an ideal assessment would be a good teaching activity and, indeed, might even serve as a teaching activity when not used for assessment.

However, observations of individual student performance are costly, time consuming, and difficult to obtain. With the ideal performance assessment as a benchmark, we developed and evaluated alternatives (or surrogates) to the benchmark. They were, in order of decreasing verisimilitude, (1) lab notebooks in which students recorded their procedures and conclusions; (2) computer simulations of the hands-on investigations; (3) short-answer paper-and-pencil problems in planning, analyzing, and/or interpreting experiments; and (4) multiple-choice items developed from observations of students conducting hands-on investigations. Finally, we compared the benchmark and surrogate assessments to a traditional multiple-choice science achievement test, the Comprehensive Test of Basic Skills (CTBS).

We developed and collected data using three hands-on investigations:

- Paper Towels — Given laboratory equipment, conduct an experiment to determine which of three different paper towels soaks up the most water.
- Electric Mysteries — Given six "mystery boxes," determine their contents by connecting electric circuits to them.
- Bugs — Determine sow bugs' preferences for various environments (for example, dark or light, dry or wet).

The performance of more than 300 students, some experienced in hands-on science and some who had received minimal science instruction from a textbook on health, was observed and scored in real time by science educators. In addition, all students completed corresponding notebooks, computer simulations, paper-and-pencil measures, and the CTBS.

Four questions guided our research: (1) Could reliable measures of hands-on performance and of surrogate assessments be developed? We wanted these measures to permit a wide variety of student responses found when doing science. We also wanted to develop a method to score performance that captured the diversity of procedures and put them on a common scale. (2) Could the performance of students with different instructional experiences (hands-on vs. textbook) be distinguished? We expected students experienced with hands-on science to perform better on the

benchmark and closest alternatives than students in a textbook curriculum program. (3) Do the performance assessments provide information about student achievement not available from traditional multiple-choice science tests? If not, perhaps nothing new had been developed. (4) Do the surrogate assessments capture the information gained from the benchmark? If so, then dollars and time can be saved in administration and scoring.

Hands-On Investigations

Students conducted the three investigations in approximately 1 1/2 hours while being observed by a scientist or science educator trained to score student performance.

Paper Towels. Students used a laboratory setup to determine which of three paper towels held the most and least water. Students were told that they could use all or some of the equipment, whatever they needed. A scheme was developed to score the diversity of procedures used to carry out the investigation on a common scale (see fig. 1). An outstanding investigation completely saturated each towel, determined the amount of water each held by a method that was consistent with the way the towel was wetted, and the entire procedure was done carefully. For example, a student might saturate the towel in the pitcher of water and weigh it in the scale, carefully removing the excess water in the scale after weighing each towel. Carelessness, inconsistencies in the method of wetting the towel and measuring the results, incomplete saturation, and irrelevant methods resulted in less than outstanding scores. The scoring scheme identified the procedure used and could thereby characterize performance in terms of both processes and outcomes. Moreover, several different performances

High-quality assessments throughout a course are vital if teachers are to accurately appraise performance.

FIGURE 1

PAPER TOWELS INVESTIGATION—HANDS-ON SCORE FORM

Student_____ Observer_____ Score _____ Script_____

1. **Method** A. Container: Pour water in/put towel in; Put towel in/pour water in; 1 pitcher or 3 beakers/glasses. B. Drops. C. Tray (surface): towel on tray/pour water on; pour water on tray/wipe up

2. **Saturation** A. Yes B. No C. Controlled

3. **Determine Result**
   - A. Weigh towel
   - B. Squeeze towel/measure water (weight or volume)
   - C. Measure water in/out
   - D. Time to soak up water
   - E. No measurement
   - F. Count # drops until saturated
   - G. See how far drops spread out
   - H. Other _____

4. **Care in Measuring** Yes No

5. **Correct Result** Yes No

Grade	Method	Saturate	Determine Result	Care in Measuring	Correct Answer
A	Yes	Yes	Yes	Yes	Yes
B	Yes	Yes	Yes	No	Yes/No
C	Yes	Yes/Controlled	Error		Yes/No
D	Yes	No	Missing		Yes/No
F	----------	----------	No Attempt	----------	----------

could result in the same letter grade.

Bugs. Students used laboratory apparatus to determine the preferences of sow bugs for environments that were light or dark, damp or dry, or some combination of the four. The scoring scheme resembled the one used in the towels investigation.

Electric Mysteries. This investigation was a bit different from the others. Students were given batteries, bulbs, and wires and asked to determine the circuit components hidden in a set of mystery boxes (see fig. 2). Performance was scored on the basis of (1) their determination of the contents of each box and (2) the sequence of tests they conducted to determine the contents.

Notebook Surrogates

For each investigation students were asked to keep notebooks describing their investigation so that a friend could repeat it exactly. Notebooks have several advantages for large-scale assessment. Through the use of notebooks instead of expert observers, many students can be tested with hands-on investigations. Notebooks provide an opportunity for students to express themselves in writing, an important skill in science and a way of integrating curricular areas. And the notebooks can be scored quickly — in about one to two minutes per student. Notebooks, then, preserve much of the hands-on investigation while reducing time and cost of expert observers. They also capture the rather inventive nature of the investigations and ways of reporting on them.

Computer Simulation Surrogates

We developed our own computer simulation for the Electric Mysteries and Bugs investigations to replicate, as nearly as possible, the hands-on investigations. (The Paper Towels investigation could not be simulated adequately.) For the electric circuits investigation, students used a Macintosh computer with a mouse to connect circuits to the mystery boxes to determine their contents. The intensity of the luminosity of the bulb in a real external circuit was accurately simulated. Students could connect a multitude of circuits if they so desired. Alternatively, they could leave a completed circuit on the screen for comparison. A teacher-directed tutorial prior to the test provided students with instructions on how to record their answers, erase wires, save their work, or look at a previous page of their work on the screen. The computer recorded every move the student made. The bug simulation was constructed similarly, though it was not possible to record every move the student (and the bugs) made.

Computer simulations have a number of desirable properties for assessment. They are less costly and time consuming to administer than hands-on assessments, though development costs are considerable. Students can be tested in groups by a parent or volunteer who has been briefed on how the simulations work. Student performance can be scored quickly and easily using the scoring system developed for the hands-on investigations. In addition, a computer simulation maintains a full record of performance, so that teachers and/or students can review problem-solving processes. Finally, students experiment with the technology, discovering solutions to problems that they might not have found with other types of assessments. In other words, the computer assessment provides a good instructional tool.

Pencil-and-Paper Surrogates

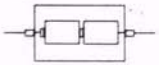
Short-answer and multiple-choice questions were chosen to parallel, in content, the three hands-on investiga-

FIGURE 2

HANDS-ON ELECTRIC MYSTERIES INVESTIGATION

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.

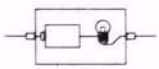
Two batteries:



A wire:

A bulb:

A battery and a bulb:



Nothing at all:

You can use your bulbs, batteries, and wires any way you like. Connect them in a circuit to help you figure out what is inside.

When you find out what is in a box, fill in the spaces on the following pages.

Box A: Has _____ inside.

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

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

_____

_____

_____

Do the same for Boxes B, C, D, E, and F.

_____

_____

tions. For the Electric Mysteries short-answer questions, students received a pictorial representation of a problem similar to one encountered during the hands-on investigation. For example, students might be asked how they would determine the contents of a particular mystery box without looking inside. For the Paper Towels and Bugs questions, students were given descriptions of portions of the investigations and questioned about the control of variables, the setup of the experiment, or the best method to use for measuring results.

Multiple-choice questions began much like the short-answer questions. Rather than formulate a response, students chose among four alternatives, all of which were based on observed performance. For example, an Electric Mysteries question presented alternative circuits connected to a box and asked students to indicate which circuit would tell them what was inside the box.

The paper-and-pencil surrogate assessments differ fundamentally from the other surrogates; they do not respond to the actions taken by the students. Even if a paper-and-pencil test provided immediate written feedback on a decision made by a student (Shulman and Elstein 1975), we doubt it would have the same impact as the feedback from the real-life (hands-on) or lifelike (computer) assessments. We may not be able to develop paper-and-pencil surrogates that overcome this limitation.

Findings

We found the process of creating performance assessments symmetric with good teaching activities to be time consuming, requiring considerable scientific and technological know-how. Development of quality performance assessments requires multiple iterations through a sequence of development, tryouts with students (getting their thoughts and comments), and revision. Short-circuiting this process leads to ill-conceived and poorly constructed assessments. Such assessments are as likely to lead to poor teaching — if teachers teach to the test, and they do (Smith 1991) — as are ill-conceived and poorly executed classroom activities.

If teachers teach to poorly constructed assessments, these assessments are likely to distort good hands-on science teaching.

Once the performance assessments were constructed, our research posed four questions about them: (1) Can they provide reliable measurements? (2) Are they sensitive to students' instructional experiences? (3) Do they provide achievement information that differs from traditional measures? (4) Are they interchangeable? Our findings (some good news, some bad news) balance the political rhetoric pushing implementation of performance assessments with a cold reality.

Reliability. Raters can reliably evaluate students' hands-on performance on complex tasks in real time. Reliabilities are high enough (above 0.80) that a single rater can provide a reliable score. But task-sampling variability is considerable. Some students perform well on one investigation while others perform well on a different investigation: general "expertise" is more in the mind of the beholder than in performance. To get an accurate picture of individual student science achievement, the student must perform a substantial number of investigations — perhaps between 10 and 20.

Instructional history. Performance assessments can distinguish students with different instructional histories. Assessments that are closely linked to a specific domain of knowledge (for example, electric circuits/electricity) are more sensitive to performance differences than more general process assessments (for example, the Paper Towel investigation). But to be sensitive to instructional history, performance assessments must be carefully crafted to measure more than science process. They need to measure the application of both concepts and procedures.

Relation to multiple-choice tests. The good news is that performance assessments do *not* duplicate information about student achievement provided by traditional tests (average correlation is about 0.45). They tap somewhat different aspects of achievement. But we are not sure what aspects of achievement multiple-choice tests or performance assessments do and do not tap. Indeed, a combination of indicators (multiple-choice, performance assessments, and others) may be needed to provide a complete picture of achievement.

Interchangeability of surrogates. Certain surrogate assessments appear to be interchangeable with their corresponding benchmark. This is especially true of notebooks for student-level assessment. Computer simulations are interchangeable with their corresponding benchmarks if the intent is to estimate *classroom-level* mean performance. But for individual students, measures of science achievement are highly sensitive not only to the investigation used (for example, Bugs vs. Electric Mysteries), but also to the method used to measure performance (for example, observation vs. simulation). Some students' scores depend on the particular investigation

(Electric Mysteries, Paper Towels, Bugs) *and* on the particular method used to assess performance (observation, notebook, computer simulation, paper-and-pencil). Indeed, each combination of investigation and method provides different insight into what students know and can do.

Conclusions. A fundamental assumption made by policymakers (for example, Bush 1991) and other education reformers (for example, Wiggins 1989) is that, by changing the nature of the achievement test, teachers who teach to the test will have to change and improve their teaching. Our experience with performance assessments suggests that this assumption is, at best, half true. Teachers will indeed change the way they teach if held accountable by performance assessments. But, without high quality assessments and staff development in quality instruction, they very well may not improve their teaching. Moreover, one-shot, end-of-year tests cannot provide adequate information on individual-level student achievement. Continuous assessment throughout the course of instruction is needed to accurately reflect student science achievement.

Assessment Development

Performance assessments are very delicate instruments. They need to be carefully crafted, each requiring a specially developed or adapted scoring method. Shortcuts taken in developing these assessments will likely produce poor measuring devices. If these instruments are used to judge the quality of education in classrooms — and they will be used for that purpose — then teachers will teach to the test. If teachers teach to poorly constructed assessments, these assessments are likely to distort good hands-on science teaching.

This conclusion was brought home to us in our observations of someone we consider to be an outstanding hands-on science teacher. In her class, a unit on electricity is taught in a series of lessons. Small groups of students conduct hands-on investigations, not unlike the ones we have developed. Students keep notebooks and draw conclusions based on the outcomes of their experiments and their discussions with other groups of "scientists" in the class.

Excited about hands-on science teaching, our teacher volunteered her class for pilot testing California's new hands-on science assessments. These assessments were constructed under severe time and cost constraints, and consequently involved minimal trials with students. To meet testing time and space constraints, and in recognition of differing curriculums in the state's elementary schools, the assessments were accompanied by detailed directions to students. Students read the instructions, followed explicit procedures, and reported what they found in the spaces provided. The assessments were more like recipes (first do this step, then do this step) than like scientific investigations.

Based on her experience with these assessments, our exemplary teacher began to modify her teaching to correspond to the state's pilot assessments. Rather than open-ended, group problem solving, emphasis was placed on reading instructions and carefully following format, carrying out a set of required procedures, and recording findings in a prespecified format in notebooks. (For example, students were admonished to "be sure to write complete sentences or the state will grade you down.") The essence of *doing science* was becoming one of following procedures. The story ends happily. After we pointed out what was happening, the teacher went back to her "old ways."

Staff Development

Unless provided the scientific and pedagogical knowledge required for hands-on science teaching, teachers may very well flounder in their attempts to match their teaching to the testing. Once again, an anecdote may bring home this conclusion. One of the teachers whose class was participating in our research knew of the Paper Towels investigation. Teaching to the test, she instructed her students to saturate the towels completely, using the timer to ensure that each towel was saturated for at least 10 minutes — a total of 30 minutes for saturation! In reality, the towels could be saturated in a matter of seconds. This led her students, when tested, to perform in a clearly stylized manner. She had informed the students, perhaps unintentionally, that science is a set of precise steps that must be carried out invariably, regardless of whether they make sense. Her approach was not particularly conducive to scientific exploration. Although the teacher could teach the students how to "do" the experiment, what was missing was an understanding of the essence of doing science.

Curriculum-Embedded Assessments

To obtain sufficiently large samples of student performance, assessments may need to be taken throughout the academic year. For example, students might receive the Electric Mysteries assessment embedded in the activities composing a unit on electricity. Likewise, assessments would be embedded in three or four other units as well.

The embedded assessments have several desirable characteristics. They provide almost immediate feedback to teachers on their students' perfor-

mance, and on how this performance compares with that of students in comparison schools. Moreover, the assessments reinforce hands-on teaching and learning. Nowhere is the symmetry between teaching and assessment more apparent than with embedded assessments.

Embedded assessments do not preclude an end-of-year examination. The latter provides both additional information on achievement and an external audit ensuring data credibility to the various audiences interested in educational accountability.

Will Achievement Improve?

Results of our research suggest that the political rhetoric calling for immediate reform of national, state, and local testing systems far exceeds current technological capability and ignores educational and social consequences. No doubt assessment systems will be changed in the very near future. The politicians will have their day. We suspect that the initial impact will be to change classroom activities and the nature of assessment, possibly embedding assessments in classroom activities. However, without quality instrumentation and extensive staff development, the bottom line — achievement — will probably not change.

The nation may be capable of producing the kind of assessment systems currently envisioned by the rhetoric if the politicians stick to their guns and do not retreat to the usual multiple-choice testing. Politicians need to provide resources for preparing beginning and current teachers for teaching and testing reforms, and for fine-tuning the assessments through research, social debate, and revision. With the symmetry between assessment and teaching firmly established, the bottom line — achievement — may very well improve. □

Authors' note: This research was supported by a grant from the National Science Foundation (No. SPA-8751511). The ideas presented reflect those of the authors and not necessarily the NSF.

References

Baxter, G. P., R. J. Shavelson, S. R. Goldman, and J. Pine. (In press). "Procedure-Based Scoring for Hands-On Science Assessments." *Journal of Educational Measurement.*

Bush, G. W. (1991). *America 2000: An Education Strategy.* Washington, D.C.: U.S. Department of Education.

Shavelson, R. J., G. P. Baxter, J. Pine, and J. Yure. (1991). "Alternative Technologies for Assessing Science Understanding." Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shulman, L. S., and A. S. Elstein. (1975). "Studies of Problem Solving, Judgment, and Decision Making: Implications for Educational Research." In *Review of Research in Education* 3: 3-42, edited by F. N. Kerlinger.

Smith, M. L. (1991). "Put to the Test: The Effects of External Testing on Teachers." *Educational Researcher* 20, 5: 8-11.

Wiggins, G. (1989). "A True Test: Toward More Authentic and Equitable Assessment." *Phi Delta Kappan* 70, 9: 703-713.

Richard J. Shavelson is Dean of the Graduate School of Education and Professor of Research Methods. **Gail P. Baxter** is a researcher in the Graduate School of Education and co-principal investigator on two National Science Foundation grants. They can be contacted at the University of California, Santa Barbara, CA 93106.