APPLIED MEASUREMENT IN EDUCATION, 4(4), 347-362

Performance Assessment in Science

Richard J. Shavelson and Gail P. Baxter
University of California, Santa Barbara

Jerome Pine
California Institute of Technology

The call for alternative assessments of science achievement grows out of the current constructivist reform in science curriculum and cognitive research. This article presents and applies guidelines for developing performance assessments aligned with this research and reform. We sample classroom activities or tasks from a domain of activities and construct performance assessments with them. Using this approach, three hands-on science investigations were constructed so that each could be scored by observers in real time. These investigations were considered benchmarks for performance assessments. Because these investigations are costly to develop and administer, surrogates were developed: student notebooks in lieu of observers, computer simulations of the investigations, free response questions paralleling parts of the investigation, and multiple-choice items with alternatives keyed to student hands-on performance. Data have been collected from over 300 fifth- and sixth-grade students using these assessments. We found that hands-on assessments can be developed through an extensive, iterative, development process; hands-on assessments are very delicate instruments. Moreover, they can be scored reliably, even in real time. However, with both benchmarks and surrogates, task heterogeneity—variations in an individual student's performance among tasks—limits the generalizability of performance to the larger domain of interest. Similarly, method heterogeneity—variations in an individual student's performance depending on whether the hands-on investigation, computer simulation or pencil-and-paper exercises was used—limits the exchangeability of the surrogates for the benchmarks.

Three forces in the United States have converged to create the impetus for alternative assessments of science achievement. One force has been recent advances in research on cognition and instruction (e.g., Glaser, 1984;

Requests for reprints should be sent to Richard J. Shavelson, University of California-Santa Barbara, Graduate School of Education, Santa Barbara, CA 93106.

Resnick, 1987). This research has changed our notions about learning and how instruction might be designed to facilitate learning. Rather than arranging instruction in a series of small steps that move students from basic skills and facts to concepts, and from concepts to problem solving, a more holistic approach is taken. Students are viewed as active agents in the teaching–learning process, constructing personal and shared meaning in a subject matter. The subject matter is well contextualized in a culture of learning and problem solving, one that encourages group as well as individual work. Hands-on activities and long-term projects are the rule rather than the exception.

A second force is the reform of science curricula. Curricula now stress active learning in which students solve concrete problems, hands-on, in small groups. Consistent with cognitive research, these curricula focus on doing rather than hearing about science. Moreover, the curricula integrate disciplines, as is the case in doing science. Mathematics and science go hand in hand, and writing about scientific ideas and keeping lab notebooks is routine.

The third force is public and professional disenchantment with the current testing technology. Based on a constructivist perspective inherent in both cognitive research and curricular reform, multiple-choice technology is now recognized as too limiting a measure of science achievement (e.g., Shavelson, Carey, & Webb, 1990). The goal is to measure students' understanding of important concepts in a subject matter, not their recall of facts (Murnane & Raizen, 1989; Raizen, Baron, Champagne, Haertel, Mullis, & Oakes, 1989; Resnick & Resnick, in press; Shavelson, et al., 1990).

SOME GUIDELINES FOR CONSTRUCTING ALTERNATIVE ASSESSMENTS

To be consistent with the new developments in cognitive research on science learning and curricular reform, we propose several guidelines for performance assessments:

1. Alternative technologies for assessing achievement need to go beyond factual recall and selecting a single correct response from among alternatives. They need to capture students' scientific understanding, reasoning, and problem solving, as well as permit novel, creative responses.
2. The assessments need to involve students responding actively with manipulatives or experimental apparatus. Some hands-on assessments need to be objective and standardized; others need to be longer-term projects that cannot be carried out in a single testing session.
3. Although desirable, long-term projects and hands-on investigations are expensive and time consuming to administer. Alternative technologies, then, need to build on advances in computer technology.

4. Alternative technologies need to reflect developments in cognitive research, notably the work on "mental models," and assess student knowledge structures that reflect understandings as well as misunderstandings in science.

5. Alternative technologies need to be aligned with curricular reform. One reason for this is to encourage teachers to orchestrate the curriculum in a manner consistent with reform. A second reason is that the interpretations of test scores are content referenced—the scores have meaning within the subject matter. Otherwise, the present science curricular reform, like that of the 1960s in the United States, will be frustrated by a mismatch between curriculum and testing. (Shavelson, et al., 1991).

A SAMPLING APPROACH TO ASSESSMENT DEVELOPMENT

A central issue in creating assessments is their content representativeness. To what extent do the assessments represent important concepts within a subject matter domain? To what extent do they fit with local or state curricula? Simply to assert that the assessments seem reasonable to a group of subject matter experts, such as teachers or university professors, is inadequate. Rather, some basis is needed for arguing that the assessments produce scores that are directly interpretable within a science domain (cf., Guion, 1979; Wigdor & Green, 1986). To this end, we sample science assessment activities from a large universe of possible activities.

Goals for the activities had to be established. Instructions had to be crafted to ensure that students understood what was expected. Materials had to be built that formed an integral part of the activity, permitting active exploration by students. A system for scoring students' performance had to be developed, one that captured a diversity of performance. Finally an iterative process was carried out to fine tune the tasks and scoring; a process of development, test (with students talking aloud as they performed), revise, and retest.

Two features of this procedure for developing performance assessments are noteworthy. First, we have taken a sampling approach to assessment construction. We view the assessments we create as exchangeable for an indefinitely large number of assessments that could be developed. Our intent is to generalize the findings from a sample of assessments to a large domain of science process performance.[1]

[1]From this perspective, we can examine the generalizability of these performance measurements using the formal statistical apparatus of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989).

Second, we view assessment as the flip side of the instruction coin and as symmetric with instruction (Shavelson & Baxter, 1990). Good instructional activities can be translated into assessments; good assessments can be used as instructional activities. A possible criterion for determining content representativeness is to ask, "Would this assessment make a good teaching activity?"

HANDS-ON PERFORMANCE ASSESSMENTS AND SOME SURROGATES

Our work on science assessments over the past 3 years involved a team of researchers and science teachers at the University of California, Santa Barbara and the California Institute of Technology who have been developing and evaluating hands-on performance measures and surrogates to them (Figure 1). This research compares these alternative technologies with traditional multiple-choice science tests such as the Comprehensive Test of Basic Skills (CTBS).

These alternatives are based on students' performance of concrete, meaningful investigations. Moreover, they are scored so as to preserve the procedures used in carrying out the investigation in addition to providing a common metric on which to score a wide variety of creative performances.

We developed and collected data with three hands-on investigations (a) Paper Towels—determine which of three different paper towels soaks up the most/least water, (b) Electric Mysteries—determine the contents of six mystery boxes by connecting circuits to them, and (c) Bugs determine sow bugs' preferences for various environments (e.g., dark or light, dry or wet). The performance of over 300 fifth- and sixth-grade students in a model hands-on science curriculum and students who received little science instruction was observed and scored by science educators.

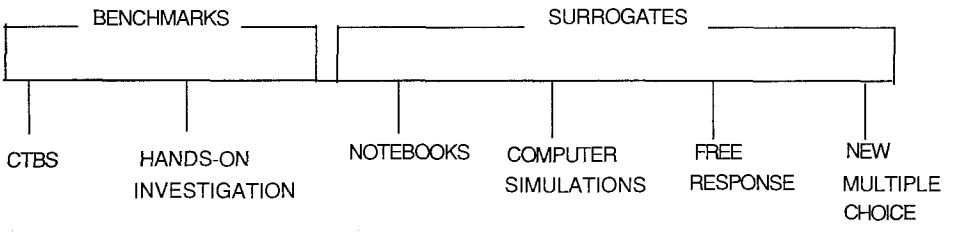


FIGURE 1 Hands-on performance assessments and their surrogates.

In conceiving the project, we recognized that, for large-scale assessment,[2] hands-on performance assessments scored by expert observers are impractical. Consequently, we examined less costly and time-consuming surrogates of the real thing (see Figure 1). In order of decreasing verisimilitude, the alternatives are notebooks based on the hands-on investigation, computer simulations, open-ended paper-and-pencil exercises, and new forms of multiple-choice tests based on mental models research.

Hands-on Investigations

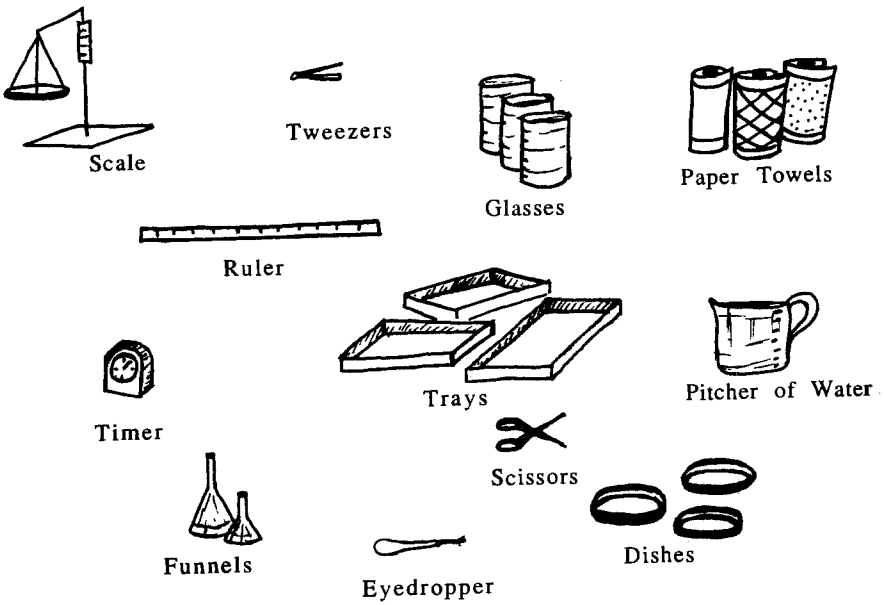
Paper towels. Students were given a laboratory set up to conduct an investigation to determine which of three paper towels held the most and least water (Figure 2). Students were told that they could use all or some of the equipment, whatever they needed. A scoring scheme was developed to capture both the diversity of procedures used to carry out the experiment and to score this diversity of performance on a common scale (Figure 3). An outstanding experiment 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 all this was done carefully. For example, a student might have saturated the towel in the pitcher of water and weighed 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 lead to less than outstanding scores. Moreover, the scoring scheme captured the procedure used and could thereby characterize performance in terms of both processes and outcomes.

Bugs. Students were provided laboratory apparatus and asked to conduct a series of experiments to determine the preferences of sow bugs for light and dark, and damp and dry environments. The scoring scheme used in the towels investigation was readily adapted to the bugs investigation.

Electric mysteries. This investigation was a bit different. Students were asked to use batteries, bulbs, and wires in a circuit to determine the contents of a set of mystery boxes (see Figure 4). Their performance was scored on the basis of (a) their determination of the contents of each box and (b) the sequence of tests they conducted on the box to determine the contents.

[2]Our work focuses on large-scale assessment. Nevertheless, the assessments created might be embedded in a hands-on science curriculum or used to gain diagnostic information. However, we have not validated the assessments for these proposed uses.

You have three different kinds of paper towels in front of you and some equipment for doing scientific experiments.



Problems:

1. Find our which paper towel can hold, soak up or absorb the most water.
2. Find out which paper towel can hold, soak up or absorb the least water.

FIGURE 2 Hands-on Paper Towels investigation.

Notebook Surrogates

Students were asked to keep notebooks enumerating the procedures they used in their investigations. They were asked to describe their investigation so that a friend could repeat it exactly. By using notebooks instead of expert

Hands-On Paper Towels Score Form

Student _____ Observer_____ Score____

1. Method for Getting Towel Wet

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/put towel in

D. No Method

2. Saturation **A.** Yes **B.** No **C.** Controlled (same amount of water- all towels)

3. Determine Result

A. Weigh towel

B. Squeeze towel/measure water (weight or volume)

C. Measure water in/out

D. Count # drops until saturated

E. Irrelevant measurement (ie. time to soak up water, see how far drops spread out, feel thickness)

F. Other_____

4. Care in saturation and/or measuring Yes No A little sloppy (+/-)

5. Correct result Most Least

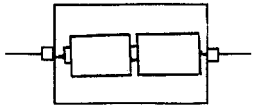
Grade	Method	Saturate	Determine Result	Care in Measuring	Correct Answers
A	Yes	Yes	Yes	Yes	Both
B	Yes	Yes	Yes	No	One or Both
C	Yes	Controlled	Yes	Yes/No	One or Both
D	Yes	No **or**	Inconsistent	Yes/No	One or Both
F	Inconsistent **or**	No **and**	Irrelevant	Yes/No	One or Both

FIGURE 3 Scoring form for hands-on Paper Towels investigation.

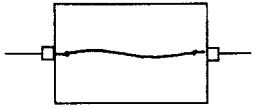
observers, large numbers of students could be tested with hands-on investigations. Moreover, notebooks provide an opportunity for students to express themselves in writing, an important skill in doing science and a way of integrating curricular areas. The notebooks were scored in a very brief

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.

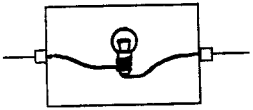
Two batteries:



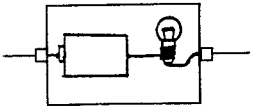
A wire:



A bulb:



A battery and a bulb:



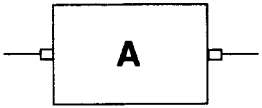
Nothing at all:

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 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 ?**

_____

_____

Box B: Has _____ inside.

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

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

_____

_____

FIGURE 4 Hands-on Electric Mysteries investigation.

amount of time, on the order of one to two minutes per student. Notebooks, then, preserved a great deal of the hands-on investigation while reducing time and cost of expert observers (see Figure 5). Moreover, they captured the rather inventive nature of the investigations and ways of reporting on them.

Computer Simulations

Computer simulations were developed for the electric mysteries and bugs investigations. The simulations were developed so as to replicate, as nearly as possible, the hands-on investigations. For the electric circuits investigation, students used a Macintosh computer with a mouse to connect circuits with the mystery boxes to determine their contents (Figure 6). The intensity of the luminosity of the bulb in a real external circuit was accurately simulated. Students connected a multitude of circuits if they so desired. Alternatively, they could leave one completed circuit on the screen for comparative purposes. Instructions on how to record their answers, erase wires, save their work, or look at a previous page of their work on the screen were given in a teacher-directed tutorial format prior to the test. The computer recorded every move the student made.

The bug simulation was constructed similarly. Figure 7 shows an experimental set up to determine whether sow bugs choose light or dark environments.

Computer simulations have a number of desirable properties for assessment. They are less costly and time consuming to administer than hands-on assessments although development costs are considerable. Students can be tested in groups by a parent or other volunteer who has been briefed on how the simulations work. Student performance can be scored quickly and easily. 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 with other types of assessments.

Pencil-and-Paper Surrogates

Free response and multiple-choice items were developed to parallel the three hands-on investigations. Examples of these items for the electric mysteries investigation are presented in Figure 8 (free response) and Figure 9 (multiple choice). The alternatives in the multiple-choice items were based on students' misconceptions inferred from observations of the hands-on investigation.

We have found a fundamental difference between these and other surrogates. The paper-and-pencil surrogates do not provide immediate

A. Steps in Experiment: Please number each step in order.

Step	What you did
1.	First of all I tore the paper towels from their rolls. Then I made a list of all three types of towels by color (white, orange, blue) one at a time they were soaked in water. When they were brought out of the water the excess water was dripped off, after that the towel was weighed. One by one the same procedure was followed, recording the weight each time. It was done twice for each towel. The different weights of each towel, two trials were then added and averaged. After all that the averages were compared find the ones with the most & least weights.
	The same procedure mentioned before.

B. Here are some questions about your experiment. Answer each of the questions "yes" or "no".

1. Were all the paper towels the same size? no
2. Were all the paper towels completely wet? yes
3. Did you use the same amount of water to get each paper towel wet? no
4. Did you let each towel soak in the water for the same amount of time? yes

C. How did you know from the experiment which paper towel holds, soaks up or absorbs the most water and which paper towel holds, soaks up or absorbs the least water?

Most: first the were soaked in water, weighed, and recorded twice, then they were averaged

Least: same as above

D. Francisco thinks all of the paper towels must be completely wet before you can decide which paper towel holds the most water and which holds the least. Sally does not think the paper towels have to be completely wet. What do you think?

must be completely wet

FIGURE 5 Example of paper towels notebook.

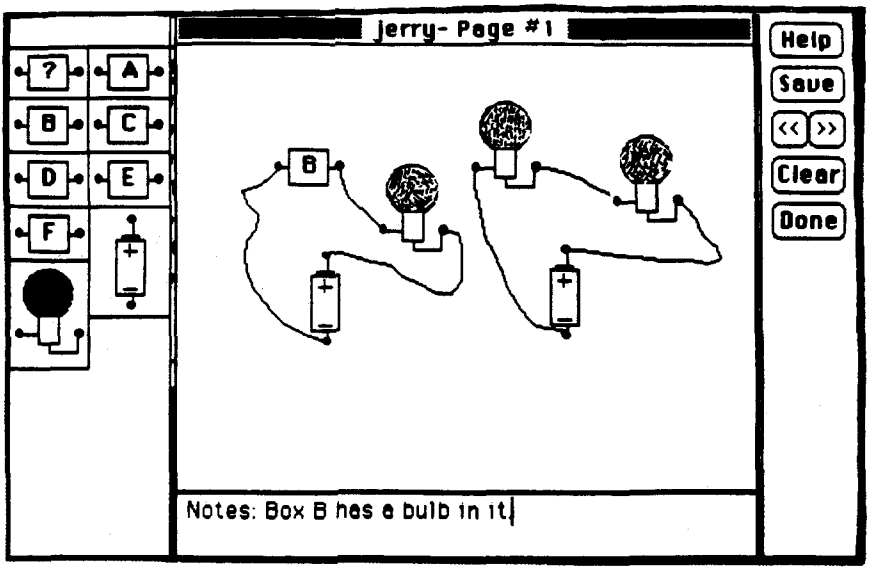


FIGURE 6 Computer screen for Electric Mysteries simulation.

responses to the actions taken by the students. Even if some type of test that provided written feedback were developed, we doubt it would have the same impact as the real-life (hands-on) or life-like (computer) responses of the other assessment methods. We may not be able to develop paper-and-pencil surrogates that overcome this limitation.

PROMISES AND PERILS OF ALTERNATIVE ASSESSMENT TECHNOLOGIES

The findings of our research are informative, especially in the current environment in which politicians are pushing implementation of alternative assessments way ahead of the development research and technology. First, the good news. We seem to be measuring something different about science process performance than what was measured by traditional multiple-choice tests, and we can do so reliably, at least we can if we take each hands-on investigation individually. Now, the bad news. There are considerable limitations to performance assessments that still need attention before we are in a position to use them as alternative technologies to measure science achievement.

Hands-on Performance Measures

Hands-on performance assessments can be developed for large-scale assessment purposes, and they make good teaching activities as well. But they are

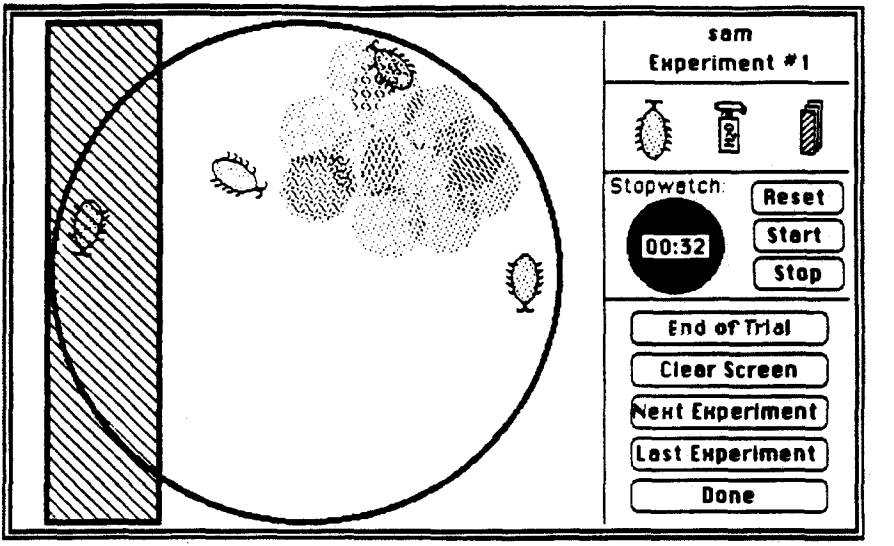


FIGURE 7 Computer scree for Bugs simulation.

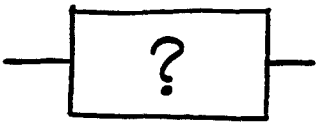
time consuming to develop and administer and are delicate instruments requiring fine tuning.

Scoring systems can be developed to capture the diversity of hands-on performance, and raters can be trained to code reliably scientific procedures and score performance on a common scale (Baxter, Shavelson, Goldman, & Pine, in press). But performance tasks are heterogeneous. They vary on a number of factors, especially their knowledge-domain specificity and requirements for students to monitor their own performance as they proceed with a task. Some are inherently more difficult than others. More importantly, some students perform well on one task and others perform well on another task. Consequently, a substantial number of assessment tasks are needed to generalize, with any degree of confidence, from students' observed performances to the science domain of interest.

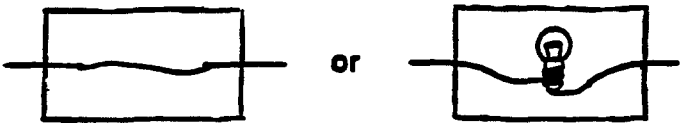
Hands-on performance assessments distinguish students experienced in hands-on science from students who have received a more traditional text-book approach, especially with knowledge-domain specific investigations (e.g., Electric Mysteries). Each draws less on traditional cognitive abilities than do multiple-choice achievement tests, and they measure different aspects of science achievement.

Surrogates

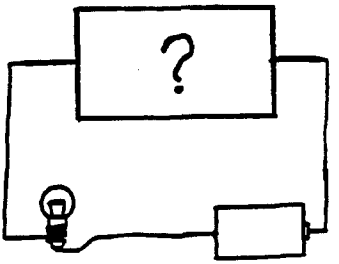
Some surrogates can be developed to reflect the complexity of hands-on investigations. But their exchangeability with hands-on assessments varies



The box with the question mark has either a wire or a bulb in it.



Susan hooked up this circuit.



How could she tell what was in the box without looking inside?

FIGURE 8 Example of free response paper-and-pencil item.

considerably with notebooks and computer simulations providing a closer match than the less expensive paper-and-pencil measures. Moreover, the level of a student's performance depends, in part, on the tasks sampled and the assessment method used. For example, some students scored high on Bugs and low on Electric Mysteries. Finally, some students who scored low

This box  contains either a wire or nothing.

Look at the circuits below. Circle the letter of the circuit you would use to find out what is in the box.

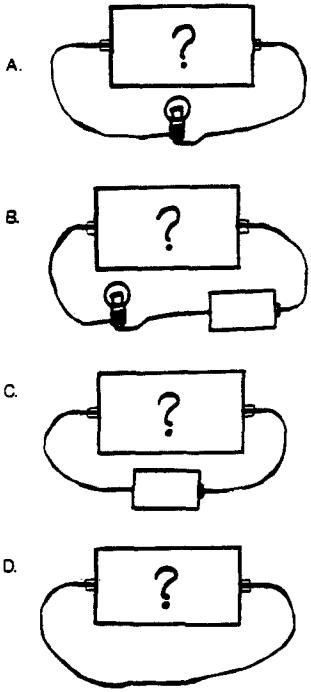


FIGURE 9 Example of multiple-choice paper-and-pencil item.

with the hands-on version of Electric Mysteries (e.g., scores of 1 or 2) scored high with the computer simulation (e.g., scores of 5 or 6), and vice versa. Large samples, even of paper-and-pencil measures, are needed due to task and method heterogeneity. As is the case in hands-on performance measurement, content representativeness is at issue, especially because different methods tap somewhat different aspects of the content.

The same scoring system developed for the hands-on performance

measures applies to high fidelity surrogates; the hands-on scoring rubrics can be used with notebooks and computer simulations; not so with the paper-and-pencil measures. The surrogates can be reliably scored; rater agreement is not a problem. But unreliability is introduced by task heterogeneity across all surrogates.

Impact

If educational systems react to the alternative assessment technologies in the way they have to traditional multiple-choice tests, teachers will teach to the test. Ideally, teachers would focus on content and building knowledge and skills in doing hands-on science instead of teaching children strategies for selecting among multiple-choice alternatives. If this happens, the curriculum as experienced by students may reflect, at least to some degree, the curriculum as envisioned by reformers. Teachers' plans for instruction and the classroom implications of these plans will, of necessity, change. Manipulatives, experiments, and student group work will become integral aspects of instruction. Indeed, teachers will have to change their everyday routines for teaching science; these changes may lead to restructuring schools (Shavelson & Baxter, in press-a).

ACKNOWLEDGMENTS

This research was supported by a grant (No. SPA–8751511) from the National Science Foundation (NSF).

The ideas presented reflect those of the authors and not necessarily those of the NSF.

REFERENCES

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (in press). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement.*

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39,* 93–104.

Guion, R. M. (1979). *Principles of work sample testing: III. Construction and evaluation of work sample tests* (TR–79–A10). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Murnane, R. J., & Raizen, S. A. (Eds.). (1989). *Improving indicators of the quality of science and mathematics education.* Washington, DC: National Academy Press.

Raizen, S. A., Baron, J. B., Champagne, A. B., Haertel, E., Mullis, I., & Oakes, J. (1989). *Assessment in elementary school science education.* Washington, DC: National Center for Improving Science Education.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Resnick, L. B., & Resnick, D. P. (in press). Tests as standards of achievement in schools. In B. R. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement and instruction.* Boston: Kluwer Academic Publishers.

Shavelson, R. J., & Baxter, G. P. (1990, September). *The symmetry of teaching and testing: Implications for teachers' instructional decisions.* Paper presented at the International Symposium on Research on Effective and Responsible Teaching, University of Fribourg, Fribourg, Switzerland.

Shavelson, R. J., & Baxter, G. P. (in press). Linking assessment with effective and responsible teaching: Implications for instructional decisions. In F. Oser, J. L. Patrie, & A. Dick (Eds.), *Effective and responsible teaching: A new synthesis.* San Francisco: Jossey-Bass.

Shavelson, R. J., Baxter, G. P., Pine, J., Yurè, 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.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71,* 692–697.

Shavelson, R. J., Webb, N. M., & Rowley, G. (1989). Generalizability theory. *American Psychologist, 44,* 922–932.

Wigdor, A. K., & Green, B. F. (1986). *Assessing the performance of enlisted personnel: Evaluation of a joint-service research project.* Washington, DC: National Academy Press.