Performance Assessments
Political Rhetoric and Measurement Reality

RICHARD J. SHAVELSON  GAIL P. BAXTER  JERRY PINE

By most yardsticks—international comparative studies of achievement, national assessment of educational progress, and some state testing programs—the United States has lost its academic competitiveness. The stakes are high. According to President Bush: "Today, education determines not just which students will succeed, but also which nations will thrive in a world united in pursuit of freedom in enterprise" (1991, p. 1). He went on to argue that:

The time for all the reports and rankings, for all the studies and the surveys about what’s wrong in our schools is past. If we want to keep America competitive in the coming century . . . we must accept responsibility for educating everyone among us . . . (p. 2)

The solution to "America’s" education woes, according to Bush, is nothing short of revolution. ("There will be no renaissance without revolution" [Bush, 1991, p. 3].) He concludes: "For the sake of the future, of our children and the nation’s, we must transform America’s schools" (p. 2, emphasis ours).

Part of the president’s strategy for the transformation of "America’s schools" lies in an accountability system that would track progress toward the nation’s education goals as well as provide the impetus for reform. Here we focus primarily on issues of accountability and student achievement.

Political Strategy for Education Reform

The president has a strategy for transforming the nation’s education system, one that includes parental choice of schools and the establishment of a private-sector corporation to fund school research-and-development teams. As a first step, however, Bush puts nationwide testing and accountability center stage:

We will develop voluntarily . . . national tests for 4th, 8th, and 12 graders [sic] in the five core subjects [English, mathematics, science, history, and geography]. These American Achievement Tests will tell parents and educators, politicians and employers, just how well our schools are doing. I’m determined to have the first of these tests for 4th graders in place by the time that school starts in September of 1993. And for high school seniors, let’s add another incentive . . . a Presidential Citation to students who excel on the 12th grade test. (1991, p. 5)

This is quite a remarkable strategy. The president, a Republican president no less, has proposed a system of national examinations. This proposal is a long way from the democratic notion of local autonomy and control of schools, once a fundamental tenet of education in the United States.

Not only does the president recommend voluntary national examinations; he set a time schedule for their implementation and added in an extra incentive: a Presidential Citation for achievement.

Perhaps equally striking is that he did not call for the usual cost-efficient multiple-choice tests that pervade achievement testing in the United States. Rather, he seeks to challenge "not only the [educational] methods and the means that we’ve used in the past, but also the yardsticks that we’ve used to measure our progress" (Bush, 1991, p. 2, emphasis ours). The tests would be tied to "World Class Standards" ("U.S. students will be first in the world in science and mathematics achievement" [p. 19]) and designed "to foster good teaching and learning as well as to monitor student progress" (p. 21).

The good teaching and learning would, presumably, result by aligning the testing program with the curriculum (National Goals Panel, 1991).

In counterpoint to the president’s assertions about voluntary national examinations using new testing technologies, we present the findings of a program of psychometric work on performance assessments in science. This research casts doubt on the extent to which a reformed accountability system can be attained in the president’s time frame. The paper concludes by cautioning that accountability systems may not be an appropriate policy instrument for achieving the nation’s goals in the year 2000.

Reality of Alternative Achievement Assessments

Over the past four years, a team of researchers at the University of California–Santa Barbara and the California Institute of Technology have developed and evaluated alternative assessment technologies in science, assessments consistent with the emerging constructivist assumptions about learning and teaching. These alternatives are based on students’ performance of concrete, meaningful tasks. Moreover, the assessments are scored so as to capture not just the "right answer," but also the reasonableness of the procedure used to carry out the task or solve the problem. Finally, these alternatives are developed with recognition of the symmetry between testing and teaching. That is, a good assessment makes a good teaching activity, and a good teaching activity makes a good assessment. The research focuses on upper elementary and middle-school children and is presented with the intent of providing concrete examples of the new vision of assessment, even if they only approximate the ideal.

The Goals

Our research sought to evaluate the reliability and validity of science performance assessments. We assumed that hands-on investigations observed and scored by scientists
and science educators approximated the ideal or benchmark assessment. We recognized that in both dollars and human resources, the benchmark was costly to use in large-scale accountability systems. Consequently, we examined the quality and interchangeability of other, less costly "surrogates" (Figure 1). These alternatives were, in order of decreasing verisimilitude: notebook reporting the investigation, computer simulations of the investigation, short-answer questions paralleling parts of the investigation, and multiple-choice tests with alternatives based on observed student procedures.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Surrogates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Investigations</td>
<td>Notebooks</td>
</tr>
</tbody>
</table>

FIGURE 1. Study design—performance assessment benchmark and surrogates.

We asked a number of questions about the reliability and validity of the performance assessments. In terms of reliability, we asked: How large a sample of observers is needed to produce a reliable measurement? How large a sample of assessment tasks is needed to produce a reliable measurement?

In terms of validity, we asked a series of questions: First, could the performance assessments distinguish students who had received hands-on science instruction from those who had received the traditional textbook curriculum? We expected students immersed in a hands-on science curriculum to outperform students in a traditional textbook curriculum. Second, do the benchmarks and multiple-choice tests measure different aspects of science achievement? If the correlation between the two types of measures is high, the additional costs of redesigning an assessment system would be difficult to justify. Third, are the correlations between aptitude and the benchmarks lower than the correlation between aptitude and the multiple-choice science achievement test? With the concrete, well-contextualized nature of performance assessments, we expected traditional aptitude to play a smaller role. Finally, in an attempt to discover cost-efficient surrogates for the benchmark, we asked: How interchangeable is each surrogate for the benchmark?

The Study

We collected data on over 300 fifth- and sixth-grade students in two school districts. One district was located in a suburban area and recognized for its hands-on science curriculum (Begley, 1990). The other district, located in an urban center, did not have a regular science curriculum. Rather, science was encountered as part of a textbook course on "health." All students in both districts took a measure of general aptitude and a standardized, multiple-choice science achievement test. Then students responded to the performance assessments in the following order (and separated by 3 weeks): paper-and-pencil measures, observed investigations with notebooks, and finally computer simulations.

We developed three observed investigations. In the first investigation, "Paper Towels," students determined which of three kinds of paper towels soaked up the most and least water (Department of Education and Science, 1994). A laboratory setup including a pitcher of water, trays, beakers, a scale, a timer, and so on was provided. In the second investigation, "Electric Mysteries," students determined the circuit components hidden in each of six mystery boxes (Figure 2) using a collection of wires, batteries, and bulbs. For the third investigation, "Bugs," students conducted three experiments with sow bugs to determine the bugs' preferences for various environments. Two experiments manipulated a single independent variable: dampness (damp or dry), and darkness (light or dark). The third experiment was a 2 x 2 factorial combining the dampness and darkness variables.

Following the Paper Towels investigation, students completed a "scientific notebook" in which they described the steps taken in their investigation and the variables controlled. For Electric Mysteries, students recorded their findings in a notebook as they proceeded. They (a) noted the contents of each mystery box, (b) drew a diagram of the circuit used to determine the contents, and (c) gave a reason for their decision (Figure 3). For each of the three Bugs experiments, students (a) drew a picture of their experimental setup, (b) described the steps used, and (c) explained the basis for arriving at their conclusions.

Both the Electric Mysteries and Bugs investigations were simulated on a Macintosh computer. (Paper Towels could not be adequately represented by the computer.) The Electric Mysteries simulation (Figure 4) replicated, as nearly as possible, the hands-on investigation. Students used the Mac mouse to connect circuits to a set of mystery boxes to deter-

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.

FIGURE 2. Electric mysteries investigation.
mine their contents. The software emulated the behavior of a real circuit. When students finished and signed off the computer, their every move was recorded and could be played back for scoring or instructional purposes. The bugs simulation implemented, as realistically as possible, the hands-on bugs experiments. Students could place up to five bugs in a dish using the mouse. The bugs moved randomly, obeying probability distributions for speed and direction. With the implementation of the experimental conditions, their behavior became systematic but still stochastic. When the student finished the assessment and logged off the computer, each experiment was captured for scoring later.

Short-answer and multiple-choice questions were chosen to parallel, in content, the three hands-on investigations. For the Electric Mysteries short-answer questions, students were given 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 box without looking inside (Figure 5). For the Paper Towels and Bugs questions, students were given descriptions of portions of the investigations and probed as to the control of variables, the setup on an experiment, or the best method to measure the result for a given experimental procedure.

Multiple-choice questions had stems similar to the short-answer questions. 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 mystery box and asked students to indicate which circuit would tell them what was inside the box (Figure 6).

Observed investigations, notebooks, and computer simulations were all scored using the same scoring system. For both Paper Towels and Bugs, because of the variety of experiments carried out by students and the necessity of linking a score to the logic and consistency of the particular pro-

---

**Box A:** Has a battery and a bulb inside.

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

![Diagram of Box A with a battery and a bulb]

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

I could tell because the bulb lights, but it was dim so I think it was a battery and it light the bulbs.

**Box B:** Has a wire inside.

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

![Diagram of Box B with a wire]

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

I think it was just a wire in side because it would not light without a battery.

**FIGURE 3.** Electric mysteries notebook.

---

**FIGURE 4.** Electric mysteries computer simulation.

---

**FIGURE 5.** Electric mysteries short-answer question.
A procedure used, a procedure-based scoring system was developed (Baxter, Shavelson, Goldman, & Pine, 1992). Letter grades A through F were determined from the experimental procedures and converted to a 14-point scale (A+ = 14; F = 1). A variety of procedures could result in the same letter grade. For Electric Mysteries, the sequence of circuits students used to test each box and the students’ decisions as to the contents of a given box were taken into consideration in evaluating performance (Figure 7). Each box was scored 1 if both the sequence of circuits and the contents of the box were correct, otherwise 0. The paper-and-pencil items were scored right or wrong.

The Findings

For the benchmark—observed hands-on investigations—we (e.g., Shavelson, Baxter, Pine, Yurê, Goldman, & Smith, 1991; Shavelson, Baxter, Pine, & Yurê, 1991) found that:

- Interrater reliability was consistently high for all investigations and varied little by students’ curricular experience. In addition, raters agreed almost perfectly on procedures students used to conduct the investigation.
- Intertask reliability (“internal consistency”) was more difficult to attain. Some students performed well on one task (e.g., mystery box or bug experiment) while other students performed well on another task.

- For all investigations, mean performance was higher for students in the hands-on than in the textbook curriculum. The magnitude of the difference varied by investigation; Electric Mysteries, requiring domain-specific knowledge, produced the greatest difference in performance.
- The correlations with the standardized multiple-choice test were only moderate in magnitude, suggesting that these tests measured different aspects of science achievement.
- The correlations between aptitude and the benchmark were lower than between aptitude and the standardized science test.

The interchangeability of the surrogates for the corresponding benchmarks varied. Notebooks provided the closest approximation in reliability and validity. The next closest surrogate was the computer simulation. Mean performance was comparable to the benchmarks as were the patterns of correlations. However, some students who scored high on the benchmark scored low on the computer simulation and vice versa (Pine, Baxter, & Shavelson, 1991).

The paper-and-pencil surrogates did not fare as well. Compared with the benchmark, the short-answer items were less reliable and correlations with the standardized achievement test and the aptitude test were lower. Moreover, mean performance of students experienced in hands-on science did not differ significantly from performance of students receiving traditional instruction (Baxter, Shavelson, Pine, & Yurê, 1991). The multiple-choice items fared even worse.
Implications for the President’s Testing Agenda

The measurement reality, judging from our research, is more sobering than the political rhetoric. Before rushing out and declaring a national examination system with new assessment technologies within the next several years, the nation should follow a more reasoned pace. Here is the good news for the president’s agenda tempered by the bad news.

Good news. Raters are able to reliably evaluate student hands-on performance on complex tasks in real time. Reliabilities are high enough that a single rater can provide a reliable score.

Bad news. Task-sampling variability is considerable. In order to estimate the student’s achievement, a substantial number of tasks may be needed.

Good news. Performance assessments can distinguish students with different instructional histories. Assessments that are closely linked to a specific domain of knowledge (e.g., electric circuits, see Figure 8b) are more sensitive than more general process assessments (e.g., paper towels, Figure 8a).

Bad news. Performance assessments must be carefully crafted to measure more than science aptitude. To be curriculum sensitive, they need to measure the application of both declarative and procedural science knowledge. Student performance, then, will depend on access to the curriculum being tested. Hard decisions will have to be made about curriculum coverage. A curriculum syllabus may have to be set for a particular academic year, stressing fewer topics and delving into them in greater depth.

Good news. Taken one at a time, the performance assessments correlate about 20 points less with aptitude than do standardized science achievement tests.

Bad news. Taken in the aggregate, a combination of the alternative assessments correlates about the same with aptitude as does the standardized science achievement test. Aptitude, then, is a major factor in generalizing performance across assessment tasks.

Good news. Certain surrogates appear to be interchangeable with their corresponding observed investigation benchmarks. This is especially true of notebooks for student-level assessment, and also true of computer simulations if the intent is to estimate a population mean.

Bad news. Measures of science achievement are highly sensitive, not only to the investigation sampled but also to the method used to measure performance. Some students’ scores depend on the particular investigation sampled (e.g., Electric Mysteries, Bugs) and on the particular method used to assess performance (e.g., observation, notebook, computer simulation). Indeed, each method provides different insight into what students know and can do. For example, some students received the maximum score, 6, on the observed Electric Mysteries task and a score of 1 on the computer simulation (see Figure 9). Decisions will need to be made about how achievement is defined in a subject-matter domain.

President Bush’s call for reform in the nation’s accountability systems is, indeed, timely. However, we believe that his and others’ expectations about what technologies currently exist and the time frame needed to develop new technologies are unrealistic. If new accountability systems are developed hastily, and this is what is currently happening at both federal and state levels, they may be just as likely to drive education in unwanted directions as in the desired direction. More debate and tryouts, informed by basic and evaluative research, are needed. The political rhetoric is simply too far ahead of the technical reality.

Our research has not addressed the social impact of new accountability systems. Untested assumptions are made about the salutary effects of performance assessments on students, teachers, and the public (Shepard, 1991; Smith, 1991). In the final analysis, we suspect that this nation may be placing far too much weight on accountability to achieve its reform agenda. Judging from past experience, those states with the strongest and most technically sound accountability systems have not achieved their desired reforms (Jaeger,
Perhaps what is needed is far less account taking and far greater consideration and resources given to teaching and learning, especially for students drawn from diverse social, economic, cultural, and language backgrounds.

An increased emphasis on and bully-pulpit use of high-stakes testing may, paradoxically, have a deleterious effect on U.S. education. Tactical, political ‘fixes’ are not what is needed. Rather, we believe that a long-term, realistic approach to assessment—one that transcends politicians’ terms of office—is.

Note

This research was supported, in large part, by a grant from the National Science Foundation (#SPA-8751511). A version of this paper was read at a conference on “Mehrdimensionale Lehr-Lern-Anordnungen: Lernen, Denken, Handeln in Komplexen Ökonomischen Situationen,” the Georg-August Universität, Göttingen, Germany.

References


Richard J. Shavelson is dean and Gail Baxter is a researcher at the Graduate School of Education, University of California—Santa Barbara, Santa Barbara, CA 93106. Dr. Shavelson specializes in psychometrics and policy analysis, and Dr. Baxter specializes in psychometrics. Jerry Pine is a professor at the Department of Physics, California Institute of Technology, Pasadena, CA 91125. He specializes in high energy physics and biophysics.

Wanted: Funny Folk

In order to add new mirth-inducers to next year’s Triennial Travesties in Atlanta, all would-be travestites are invited to submit evidence of funniness for review by independent panels of comedically oriented reviewers. Submitted evidence may take any form, for example, essay, audiotape, videotape, or major motion picture. Please send four copies of all evidence by August 14, 1992, to W. James Popham, IOX Assessment Associates, 5420 McConnell Avenue, Los Angeles, CA 90066.