Development of Performance Assessments in Science: Conceptual, Practical, and Logistical Issues

Guillermo Solano-Flores
WestEd
and
Richard J. Shavelson
Stanford University

What are practical and logistical constraints in developing science performance assessments (SPAs)? What are key components in a framework for conceptualizing the process? What are the major steps in SPA development?

Knowledge of the strengths and weaknesses of science performance assessments (SPAs) has increased significantly in the last few years. Although the psychometric challenges posed by performance assessment are far from being completely addressed and exhaustively investigated, we now have a better idea about what we can and should not expect about task sampling variability (Shavelson, Baxter, & Gao, 1993), assessment method variability (Baxter & Shavelson, 1994; Dalton, Morocco, Tivnan, & Rawson, 1994), interrater reliability (e.g., Baxter, Shavelson, Goldman, & Pine, 1992; Dunbar, Koretz, & Hoover, 1991), and stability (Ruiz-Primo, Baxter, & Shavelson, 1993). Moreover, new scoring approaches have been investigated (Druker, Solano-Flores, Brown, & Shavelson, 1996; Solano-Flores & Shavelson, 1994a) as have techniques for generating SPAs (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1994).

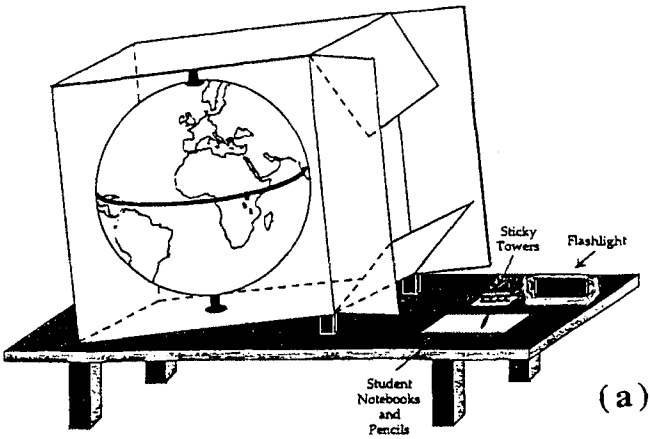
The assessments used in the body of research cited above were developed for both statewide testing and specific science curricula, Grades 4–8. We have developed some assessments to tap knowledge at the end of an instructional unit (Solano-Flores, Shavelson, Ruiz-Primo, Schultz, Wiley, & Brown, 1997), other assessments as a part of an effort to align assessment with instruction by embedding assessments within instructional units to inform teaching (e.g., Baxter & Elder, 1994; Druker, Solano-Flores, Brown, & Shavelson, 1996), and still other assessments for large-scale assessment programs (e.g., Gao, Shavelson, & Baxter, 1994).

In developing and using these assessments, we have encountered different sets of conceptual, practical, and logistical challenges. In this article, we present assessment development lessons learned with the intention of moving assessment from rhetoric to practical reality (see Shavelson & Baxter, 1992; Shavelson, Baxter, & Pine, 1992; Ruiz-Primo & Shavelson, 1996). To date, literature on the technical aspects of performance assessment either focuses on reliability and validity issues (e.g., Phillips, 1996) or provides rather general guidelines for assessment developers (e.g., Baron, 1991; Brown & Shavelson, 1996; Shavelson, Baxter, & Pine, 1991; Wiggins, 1992). By contrast, our intent here is to provide educators with conceptual tools and procedures that promote sound performance assessment practices (see Blum & Arter, 1996; Stiggins, 1994).

More specifically, we discuss and illustrate the need for SPA construction techniques and the challenges faced by researchers and teachers when they develop or use performance assessments in the classroom (e.g., materials with specific properties are difficult to obtain, classrooms are small, schools have tight schedules). The importance of these challenges should not be underestimated. They have serious implications for large-scale testing, standardization,

Guillermo Solano-Flores is a Senior Research Associate at WestEd, 12345 El Monte Rd., Los Altos Hills, CA 94022-4599. His specializations are measurement and assessment development.

Richard J. Shavelson is James Quillen Dean of the School of Education and Professor of Education and (by courtesy) Psychology at the School of Education, Stanford University, Stanford, CA 94305. His specializations are measurement and validity theory, and cognitive performance aspects of mathematics and science teaching and assessment.



(a)

Draw a dot on this map to show where you think Tower C is.

How did you figure out where Tower C is?

(b)

Observations/Results

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

Data Gathering/Modeling

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

(c)

FIGURE 1. *Performance assessment components of* Daytime Astronomy *(Solano-Flores, Shavelson, Ruiz-Primo, Schultz, Wiley, & Brown, 1997). (a)* Task: *The equipment consists of a spinning Earth globe inside a carton box, three sticky towers, and a pocket flashlight; students stick Towers A and B at two specific locations on the globe and are told what Tower C's shadow looks like when it is noon for Towers A and B; they have to find out where in the U.S. Tower C is; the solution requires modeling the sunlight by using the flashlight to project the towers' shadows onto the globe. (b)* Response Format: *Students record in notebooks their solutions, the actions they carried out, and the reasonings underlying those actions. (c)* Scoring System: *Student performance is scored based on the accuracy of the results and the accuracy of the modeling, reasoning, and observations.*

and score validity, to say the least (see Haertel & Linn, 1996). For example, an SPA might not accurately estimate some students' performance if it cannot be properly administered in both large and small classrooms; instruction may be stalled if setting up and administering SPAs take so much time that other curricular topics suffer. Whether assessment reform succeeds may ultimately depend on how properly these challenges are surmounted.

A Conceptual Framework for Science Performance Assessment

A simple conceptual framework has proven to be a good start for developing SPAs. First, three components are needed to define a performance assessment (Figure 1): (a) a *task* that poses a well-contextualized problem the solution of which requires the use of concrete materials that react to the actions taken by the student (see Wigdor & Green, 1991, for a formal, general definition), (b) a *response format* in which the student's responses are captured (e.g., record the procedures used to solve the problem, draw a graph, construct

a table, write a conclusion), and (c) a *scoring system* to score the student's responses about his or her scientific reasonableness and accuracy (Ruiz-Primo & Shavelson, 1996). Often, task and response format are so closely linked that they are not easily discernible. Yet, for conceptual clarity, we distinguish them.

Second, SPAs can be conceived as tasks that recreate the conditions in which scientists work and elicit the kind of thinking and reasoning used by scientists when they solve problems. The assessments that we have developed belong to four task types: *comparative*—conduct an experiment to compare two or more objects on some attribute, *component identification*—test objects to determine their components or how those components are organized, *classification*—classify objects according to critical attributes to serve a practical or conceptual purpose, and *observation*—perform observations and/or model a process that cannot be manipulated (Table 1).

Although there must be more task types yet to be discovered, the existence of these four and the fact that all investigations of the same type can be scored based on the same properties (Ruiz-Primo & Shavelson, 1996; Shavelson, 1995; see Table 2) give substance to the claim that there is a knowledge domain associated with what has been lumped together as "science process skills."

Table 1
Examples of Four Types of Science Tasks

Comparative investigation
- *Paper Towels:* Discover which of three kinds of paper towels holds the most water and which holds the least (Baxter, Shavelson, Goldman, & Pine, 1992).
- *Bubbles:* Discover which of three soapy solutions produces the most durable bubbles (Solano-Flores, 1994; Solano-Flores & Shavelson, 1994b).
- *Incline Planes:* Determine the relationship between the angle of inclination and the amount of force needed to move an object up a plane (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1994).

Component identification
- *Electric Mysteries:* Determine the components of the mystery box (Shavelson, Baxter, & Pine, 1991).
- *Mystery Powders:* Given a bar containing substances commonly found in the kitchen (e.g., baking soda, starch, sugar), determine which substances are in the bag (Baxter, Elder, & Glaser, 1995; Baxter & Shavelson, 1995).
- *Motor:* Given a motor, a battery, and a box containing a battery, determine the polarity of the battery that is inside the box (Druker, Solano-Flores, Brown, & Shavelson, 1996).

Classification
- *Sink & Float:* Create a classification system that allows you to predict whether an object will sink or float in tap water (Solano-Flores, Shavelson, Ruiz-Primo, Schultz, Wiley, & Brown, 1997).
- *Rocks & Charts:* Given a set of minerals, test the minerals for known attributes and create a classification system using those attributes (Druker, 1997).

Observation
- *Daytime Astronomy:* Model the path of the sun from sunrise to sunset and use direction, length, and angle of shadows to solve location problems (Solano-Flores, Shavelson, Ruiz-Primo, Schultz, Wiley, & Brown, 1997).

The framework for developing assessments can be taken a step further. We constructed a *shell* (blueprint) for comparative tasks that provides step-by-step instructions for generating comparative investigation assessments (Solano-Flores, Jovanovic, Shavelson, & Bachman, 1994). This shell addressed six types of science knowledge (planning and designing investigations, conducting a hands-on investigation, analyzing and interpreting data, and applying science knowledge), and it can be used to generate assessments at a level of inquiry that best fits assessment needs (wide open to procedural). The inquiry level is defined by the characteristics of the task and the response format (e.g., whether or not the assessment provides conceptual information, directions on how to use the equipment, aids for reporting results; Table 3). To use the shell, the assessment developer has to select a level of inquiry and follow the steps indicated. We have found that two assessments generated with the shell by a team of developers have parallel structures and similar appearances (Solano-Flores, Jovanovic, Shavelson, and Bachman, 1997). Thus, shells can potentially be used to generate parallel assessment forms for large-scale, year-to-year comparisons.

The conceptual framework, then, may provide assessment development capability to a wider audience than just those doing SPA research. The framework provides assessment developers with a scheme for thinking about and selecting task types that should be represented in large-scale comprehensive assessments. The framework also provides a way to link scoring to the task performed by students. Once a task type has been decided on, assessment developers know a great deal not only about the structure of the task but also about the characteristics of the scoring system. This knowledge saves considerable development time and cost—it may also form the basis for constructing shells for all task types.

Assessment Dimensions

We use the term *dimensions* to refer to the methodological requirements and practical and logistical constraints that must be addressed to attain an assessment's measurement goals. These dimensions can be grouped into three areas: content, equipment, and use (Table 4). The list is not intended to be exhaustive. Although many of these dimensions are *universal* (i.e., they apply to the development of *any* assessment), we exemplify them with our experience with SPAs, and we illustrate them

Table 2
Scoring Systems of Four Types of Science Tasks

Comparative

Student conducts an experiment to compare two or more objects on some property. The scoring system is **procedure-based**—it focuses on the scientific defensibility of the procedures used by the student to compare the objects. For example, in *Paper Towels,* the student conducts an experiment to find out which of three kinds of paper towels holds the most water and which holds the least water. If the student does not completely saturate one of the towels, even though he or she gets the right answer, the investigation is flawed.

Component identification

Student tests objects to determine their components or how those components are organized. The scoring system is **evidence-based**—it focuses on the quality of the evidence used to confirm or disconfirm the presence of components. For example, in *Electric Mysteries,* the student has to test 6 mystery boxes to determine their contents—two batteries, a wire, a bulb, a battery and a bulb, or nothing (two boxes have the same contents). A student who tests a mystery box first with a simple circuit containing a light bulb and, then, if the bulb doesn't light, tests the circuit with a battery and a bulb, uses a scientifically defensible way of confirming or disconfirming the presence of components.

Classification

Student classifies objects according to critical attributes to serve a practical or conceptual purpose. The scoring system is **dimension-based**—it focuses on how well the classification system constructed uses attributes that are relevant to the purposes of classification. For example, in *Sink and Float,* the student has to construct a classification scheme based on variables (dimensions) critical to floatation and use a classification scheme to predict if a set of bottles of different volumes and masses will sink or float. To classify objects as "floaters" and "sinkers," a student should consider mass, volume, and the interaction of mass and volume.

Observation

Student performs observations and/or models a process that cannot be manipulated. The scoring system is **accuracy-based**—it focuses on the accuracy of the observations performed and the models constructed. For example, in *Daytime Astronomy,* the student has to solve location problems by modeling sun shadows and to describe what shadows look like in different locations. A correct solution to the location problems is obtained when, among other things, the student models the sunlight and the earth's rotation, respectively, by shining the flashlight on the equator and rotating the earth globe to the East.

from the perspective of our conceptual framework.

To produce a sound, cost-effective assessment, each dimension needs to be addressed. The challenges posed by some dimensions cannot be anticipated: They need to be discovered and tackled by refining the task, response format, or scoring system or by making adjustments to the assessment administration procedure (see Table 5).

A considerable amount of work and time may be needed before the challenges posed by a single dimension are properly surmounted. Take the case of predictability in *Bubbles* (see Table 1), an assessment we gave to fifth-grade students before and after learning about *Bubble Science,* an instructional unit on the physics of bubbles (see Solano-Flores, 1994; Solano-Flores & Shavelson, 1994b). In *Bubbles,* students have to determine which one of three soapy solutions makes bubbles with the longest duration. The three soapy solutions used in this assessment had to satisfy four conditions: (a) The duration of the longest-lived bubbles should be short enough that several bubbles could be made for each soapy solution in a reasonable time period; (b) the standard deviation of the bubble durations for a particular solution should be small to prevent students from arriving at erroneous conclusions due to sampling error; (c) the differences between the mean bubble durations across the soapy solutions should be significant ($\alpha = .01$); and (d) the durations of the bubbles with the three different solutions should not overlap. These conditions posed high equipment predictability requirements and were met only after about a hundred soap formulas were carefully made and tested.

Tables 6 and 7 compare seven assessments on equipment and use dimensions. Three facts stand out: (a) Each assessment poses a special set of challenges for development; (b) the magnitude of the challenges posed by these dimensions depends on the specific assessment; and (c) each challenge must be tackled in a different manner. Rather than applying a set of simple rules, then, assessment developers need to identify which dimensions are critical to developing an assessment and how these dimensions interact. Experience and creativity are necessary ingredients.

The Tension Between Assessment Dimensions

The development of an assessment is shaped by the tensions among assessment dimensions. An example of this tension is the cost-benefit relationship involved in constructing equipment for an assessment. Using cheap materials reduces production costs, but cheap materials tend to behave unreliably, introducing measurement error because scores will confound performance quality and random variations in the way the equipment reacts. However, if high-quality materials are used to reduce measurement error, costs may be dramatically increased.

Another example of this tension is the relationship between interrater reliability and scorer training time. Suppose two scoring systems, A and B, are used to score performance on the same SPA (see Druker, Solano-Flores, Brown, & Shavelson, 1996). System A produces a slightly higher interrater reliability than System B, but System B is easier to learn by

Table 3
Shell for Developing Comparative Investigations: Hands-On Investigation Part, Low and High Inquiry Levels

	Low inquiry		High inquiry
Step 1	Provide preparatory knowledge in one of three ways: • Written instruction • Illustration with related task • Illustration with embedded task.	Step 1	Introduce the concepts that will be used in the assessment.
Step 2	Pose a problem or a hypothesis involving one relevant independent variable.	Step 2	Pose a problem or a hypothesis involving one relevant independent variable (A) and one irrelevant independent variable (B).
Step 3	Provide equipment—include independent variable. Introduce variable name.	Step 3	Provide equipment—include independent variable A and independent variable B. Introduce variable names.
Step 4	Tell the students which manipulations should be done and how they should be done.	Step 4	Ask the students to solve the problem or test the hypothesis.
Step 5	Ask students to solve the problem or test the hypothesis.	Step 5	Ask students to report manipulations, measurements, and results.
Step 6	Ask students to report manipulations, measurements, and results. Provide table/chart.		

scorers. Provided that the interrater reliability obtained with scoring System B is reasonably high, the score dependability gained with System A might not justify a substantial increase in the time invested to train scorers.

The tension between dimensions brings uncertainty into the process of assessment development because the actions taken to address one dimension may have unexpected, undesirable consequences on others. Rather than a linear sequence of actions, then, we use a cyclical developmental process that allows assessment developers to test the consequences of their actions (see Figure 2). In this cyclical process, the task, response format, and scoring system are revised on each iteration. When a version of the assessment is finished (Box A), we first try it out with one or two students (Box B). After revision, we then pilot it with a larger sample of students (Box C). Based on the experience gained in the tryout and pilot phases, we revise the assessment (Box D) and prepare a new version, which is also tried out, piloted, and revised.

The information gathered to revise the assessment on each iteration requires: (a) sampling students with different backgrounds and levels of academic skills from different schools and classrooms; (b) observing students and having them talk aloud as they perform the assessment; (c) interviewing them to investigate how well they understand the problem posed, the kind of knowledge and thinking skill they use in solving the problem, and how effectively the response format captures that knowledge and skill; (d) testing how reliably equipment functions; and (e) trying out the scoring system to see if scores reflect the adequacy of the students' responses.

The task, response format, and scoring system are developed in dynamic interaction: Changes made to one component imply changes to the other two. For example, the revision of the scoring system may reveal the need to refine the task (say, by changing or improving some pieces of equipment), which in turn may reveal aspects of performance that should be considered in the response format. Every transformation has practical and methodological implications that affect the quality of the assessment.

We consider an assessment ready to use when, among many other things: (a) students understand properly what the task is about; (b) the equipment reacts to the students' actions as expected; (c) the students' responses captured by the response format reflect the targeted knowledge and skills; (d) a wide variety of responses with varying degrees of accuracy are observed; (e) further changes to the task, response format, or scoring system are minimal; and (f) all student responses can be characterized by the scoring system with ease. Many iterations may occur before these conditions are met.

The Impossibility of Optimizing on All Dimensions

Due to the tensions among dimensions, no assessment can be optimized on all dimensions. Assessment developers should seek a combination of characteristics in which gains are maximized and losses are minimized. To a great extent, the process of developing a SPA consists of learning to trade off the dimensions that are critical to the assessment without jeopardizing its technical quality.

The impossibility of optimizing on all dimensions is quite evident when the final version has been produced and the assessment is ready to use. Unforeseen issues may arise when the assessment is seen from a perspective other than development—such as, shipping the equipment and administering the assessment on a large scale. An appropriate proce-

Table 4
Assessment Development Dimensions

Content
- *Representativeness:* Is the task representative of the knowledge domain addressed?
- *Variety of solutions:* Is the task amenable to many solutions varying in correctness?
- *Prompt complexity:* What actions should students be explicitly asked to report?
- *Effectiveness:* Does the response format elicit the intended type of response?
- *Meaningfulness:* Is the problem engaging enough to catch the students' interest?
- *Clarity:* Do students actually understand the problem?
- *Conciseness:* Is the problem posed in a simple, straightforward manner?
- *Equity and fairness:* Is student diversity (e.g., gender, ethnicity, SES) being considered?
- *Inquiry level:* Does the problem promote the students' active participation?

Equipment
- *Predictability:* Does the equipment react consistently to the students' manipulations?
- *Safety:* Is the equipment harmless to students?
- *Economy:* Are the materials used affordable? Is the equipment easy to replicate?
- *Material availability:* Are the materials easy to find or build with?
- *Usability:* Can the students use the equipment with ease? Are they familiar with it?
- *Resiliency:* Does the equipment endure the students' use?

Use
- *Practicality:* What are the equipment's packaging, storage, transportation, and handling requirements?
- *Set up/put away:* Can the assessment be easily set up and put away in a reasonable time?
- *Physical requirements:* What are the classroom physical conditions (e.g., illumination, water availability) needed to properly administer the assessment?
- *Grouping:* Should students take the assessment individually or in teams?
- *Control:* What activities (e.g., hand-out material, watch students) must be carried out to properly administer the assessment?
- *Completion time:* How much time are students allowed to complete the assessment?
- *Correlated error:* How should the assessment be given to ensure prompt independence?
- *Interrater reliability:* How consistently are students rank ordered by independent scorers?
- *Scorer training time:* How much training time do scorers need to use the scoring system?
- *Scoring time:* How time consuming is scoring?
- *Scoring form usability:* Can scorers use and interpret the scoring form easily?
- *Scoring protocol:* What procedure should scorers use to review the students' responses?

Table 5
Discovering and Tackling the Challenges Posed by Some Dimensions: Examples From Three Assessments

Task
- Dimension: Usability
  - The letters, *S, M,* and *L* were eliminated from the consonant letters used to label the bottles of *Sink and Float.* Although the labels were assigned randomly so the sequence of letters did not provide erroneous ordering clues when students had to sort the bottles by mass and volume, some students sorted the letters *S, M,* and *L* as if they were initials of *small, medium,* and *large.*

Response Format
- Dimension: Prompt Complexity
  - The prompts used in the notebooks of *Daytime Astronomy* provided just blank spaces. Although students were allowed to give their answers with words or drawings, they were less likely to make drawings if the prompts provided spaces bounded by boxes or a combination of a space for drawings and some lines to write on.

Scoring System
- Dimension: Scoring Protocol
  - To ensure interrater reliability, the scorer training for *Bubbles* had to include a set of detailed decision rules for reviewing and scoring the students' responses. Some students reported conflicting results in different parts of their notebooks. Others failed to provide important information where they were explicitly asked to but provided that information somewhere else in their response formats.

Table 6

Examples of Equipment Dimensions in Four Science Performance Assessments

Assessment	Predictability	Safety	Economy	Material availability	Usability
Bubbles	To produce the desired bubble durations, many soapy solution formulas have to be prepared and tested systematically.	No potential harm.	Equipment includes an electronic stopwatch, which may be difficult to find at a reasonable price.	The brand of stopwatches and basters used in the assessment are difficult to find in stores in the quantities needed for 20 kits.	Stopwatches are adapted to prevent students from using features not relevant to the task.
Electric Mysteries	To prevent variations in light intensity, the same brands of bulbs and batteries must be used consistently.	Assessment developers make sure that no electric shock is possible with the number of batteries used.	The best bulb brands—needed to ensure equipment predictability—are expensive.	The boxes are not commercially available.	Some students have difficulty using the clips to connect wires.
Sink and Float	3 or 4 BBs may make the difference between a floater or a sinker. The BBs are carefully weighted to fill the bottles so they float or sink as expected.	Children-proof medicine vials are used to prevent access to BBs. Caps are sealed with plumbers putty and silicon.	Although the materials are not expensive, the production process is extremely laborious.	All the materials are commercially available.	The combination of colors used ensures that color-blind students can distinguish the bottles.
Daytime Astronomy	To prevent light from interfering with shadow observations, the globe is placed inside a carton box.	No potential harm.	The earth globes are expensive if bought with their stands.	Earth globes without their stands have to be obtained from a provider on special request.	The combination of colors used ensures that color-blind students can distinguish the towers.

dure for use and administration, then, needs to be designed to deal with those issues without affecting the assessment's integrity.

Daytime Astronomy (Solano-Flores, Shavelson, Ruiz-Primo, Schultz, Wiley, & Brown, 1997) is a case in point. This assessment was developed to assess the knowledge acquired by students from "Daytime Astronomy," a fifth-grade instructional unit on the observation of the sun and the moon during school time and on the use of models to explain those observations (Hamilton, 1994). In the assessment, students place plastic towers at different locations on a spinning Earth globe inside a carton box and solve location problems by modeling the sun with a pocket flashlight that is used to project the towers' shadows onto the globe (see Figure 1a). The box ensures that the globe inside is dark, so the shadows projected with the flashlight can be readily seen (predictability). Because the box is big, it occupies a volume of 16 cubic inches (practicality). It would be difficult for fifth-graders to model the sun and observe the shadows projected with the flashlight individually (usability). Therefore, the students take the assessment in pairs (grouping); they take turns holding the flashlight, pointing it at the globe, and observing the tower shadows. However, to reduce measurement error on individual level scores due to student interaction, they have to report their investigations individually.

Several changes in the assessment were considered to enhance practicality: for example, using smaller globes, using inflatable globes, or transporting the kits disassembled and assembling them right before assessment administration. Each of these ideas was discarded because it would adversely affect the equipment's predictability and would have, in addition, complicated the assessment's administration. More specifically, smaller globes (and, consequently, smaller boxes) make observation of shadows difficult (predictability). Inflatable globes take considerable time to inflate (set up/put away) and distort the shadows (predictability); in addition, towers are difficult to stick on them (usability). Even if students were asked to inflate their globes, they would have to use their mouths (safety) or costly bike pumps (economy). Finally, transporting the kits in pieces and assembling them before coming to the classrooms would take a great deal of administrator time (practicality and set up/put away).

Any change made in the equipment would affect the technical

Table 7
Examples of Use Dimensions in Four Science Performance Assessments

Assessment	Practicality	Set up/Put away	Physical requirements	Grouping	Control
Bubbles	A little heavy. Soapy solutions must be made exactly 24 hours before administration. Careful handling needed to prevent the solutions from getting foamy.	Set up and put away takes a considerable amount of time. A bit messy: Some water and soapy solution is spilled on desks.	Occupies most of space on the students' desks. Almost no room left for students to write. A sink in the classroom is desirable.	The assessment is administered individually.	Some supervision is required to help students who have problems operating the stopwatches. The assessment is given in two parts on the same session.
Electric Mysteries	Very heavy. Difficult to transport. Batteries have to be checked right before administration.	Easy set up and put away. Neat, clean.	Occupies only some space on the students' desks.	The assessment is administered individually.	Little supervision required. Some batteries or bulbs may go off during administration.
Sink and Float	Heavy and a little bulky.	Set up is a bit time consuming. A bit messy: Some water is spilled on desks.	Occupies only some space on the students' desks. A sink in the classroom is desirable.	The assessment is administered individually.	Some supervision required. The assessment is given in two parts on the same session.
Daytime Astronomy	Light but bulky. Difficult to transport and store. In some classrooms, desks have to be rearranged, so the daylight does not interfere with the students' investigations.	Easy set up and put away. Neat, clean.	Occupies most of the space on the students' desks. Almost no room left for students to write. Blinds in the windows are desirable.	Students collaborate in pairs to use the equipment, but complete their notebooks individually.	Some supervision required. The assessment is given in two parts on the same session.

quality of the assessment's predictability, thus increasing measurement error. Despite practical inconveniences, the integrity of the assessment's configuration of characteristics needed to be maintained. Although the equipment is bulky—a van was needed to transport only 15 kits!—(practicality), this was not really a problem, because we worked with classes of no more than 30 students and the assessment was designed to be taken in pairs (grouping). In addition, the equipment is extremely light, has only a few components, and is easy to handle (practicality). We decided, then, that cost savings should come from administering the assessment, not from the equipment.

Summary and Conclusions

In this article, we discussed conceptual, practical, and logistical issues involved in developing science performance assessments. Our conceptual framework identified three assessment components (task, response format, and scoring system) and conceived SPAs as tasks that attempt to recreate the conditions in

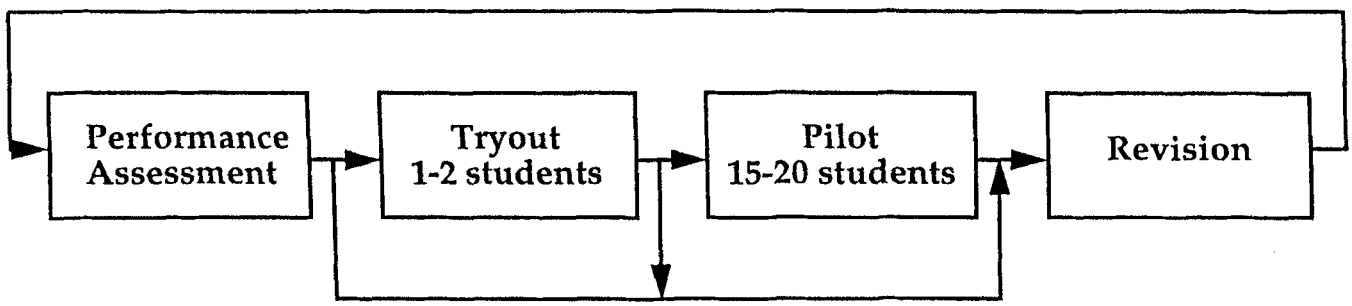


FIGURE 2. *Process of performance assessment development*

which scientists work and to elicit the kind of thinking and reasoning used by scientists to solve problems. The process for iteratively developing SPAs reflects the tensions among content, equipment, and use dimensions as well as the fact that every assessment is unique in the set of challenges that must be overcome. Once a remarkable set of trade-offs among assessment dimensions has been made to produce a reliable, valid, and usable assessment, adaptation to local contexts must be made in logistics and administration.

Seems like we have a good news, bad news conclusion. The good news is that the framework saves a great deal of developmental effort and time in constructing an initial prototype. It provides the departure point for constructing a technology for assessment development. The bad news is that the tension among assessment dimensions means that optimization is impossible. Performance assessments are very delicate instruments. Developing high-quality SPAs is a sophisticated production endeavor. Addressing the conceptual, practical, and logistical issues discussed here may contribute to easing that endeavor.

Notes

Research reported in this article was supported by grants from the National Science Foundation (Nos. ESI 95-96080 and SPA-8751511) and conducted by a team of researchers including, in alphabetical order, Marilyn Bachman, Gail P. Baxter, Janet H. Brown, Katherine Brown, Lynn Cavazos, Stephen Druker, Kimberly Feely, Xiao-hong Gao, Heather Lange, Jerry Pine, Maria A. Ruiz-Primo, Richard J. Shavelson, and Guillermo Solano-Flores. The ideas presented in this article are not necessarily endorsed by the supporting agency or our colleagues. The authors wish to thank three anonymous reviewers for their insightful comments on a previous version of this article.

References

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 305–318.

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

Baxter, G. P., Elder, A. D., & Glaser, R. (1995). *Cognitive analysis of a science performance assessment* (CSE Tech. Rep. No. 398). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: benchmarks and surrogates. *International Journal of Educational Research, 21*(3), 279–298.

Baxter, G. P., & Shavelson, R. J. (1995). *Performance assessments in elementary science classrooms: Questions of rater consistency.* Unpublished manuscript, University of Michigan.

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

Blum, R. E., & Arter, J. A. (Eds.). (1996). *A handbook for student performance assessment in an era of restructuring.* Alexandria, VA: Association for Supervision and Curriculum Development.

Brown, J. H., & Shavelson, R. J. (1996). *Assessing hands-on science.* Thousands Oaks, CA: Corwin.

Dalton, B., Morocco, C.C., Tivnan, T., & Rawson, P. (1994). Effect of format on learning disabled and non-learning disabled students' performance on a hands-on science assessment. *International Journal of Educational Research, 21*(3), 299–316.

Druker, S. L. (1997). *A framework for performance task development: Conducting pilot studies.* Unpublished manuscript.

Druker, S. L., Solano-Flores, G., Brown, J., & Shavelson, R. J. (1996, April). *A comparison of two approaches to scoring science performance assessments.* Paper presented at the Annuual Meeting of the American Educational Research Association, New York City.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303.

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: promises and problems. *Applied Measurement in Education, 7*(4), 323–342.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment.* Washington, DC: National Center for Education Statistics, U. S. Department of Education, Office of Educational Research and Improvement.

Hamilton, E. (1994). Daytime astronomy: Teacher's guide. Pasadena: California Institute of Technology.

Phillips, G. W. (Ed.). (1996). *Technical-issues in large-scale performance assessment.* Washington, DC: National Center for Education Statistics, U. S. Department of Education, Office of Educational Research and Improvement.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41–53.

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

Shavelson, R. J. (1995). *On the development of a science performance assessment.* Stanford University: National Academy of Education.

Shavelson, R. J., & Baxter, G. P. (1992). What we've learned about assessing hands-on science. *Educational Leadership, 49*(8), 20–25.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347–262.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22–27.

Solano-Flores, G. (1994). *A logical model for the development of science performance assessments.* Unpublished doctoral dissertation, University of California, Santa Barbara.

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

Solano-Flores, G., Jovanovic, J., Shavelson, R. J., & Bachman, M. (1997). *On the development and evaluation of a shell for generating science performance assessments.* Manuscript submitted for publication.

Solano-Flores, G., & Shavelson, R. J. (1994a, April). *Binary-based versus weight-based scoring in science performance assessments.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Solano-Flores, G., & Shavelson, R. J. (1994b, April). *Evaluation of a model for generating science performance assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Solano-Flores, G., Shavelson, R. J., Ruiz-Primo, M. A., Schultz, S. E., Wiley, E. W., & Brown, J. H. (1997, March). *On the development and scoring of observation and classification science assessments.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Stiggins, R. J. (1994). *Student-centered classroom assessment.* New York: Macmillan.

Wigdor, A., & Green, B. F., Jr. (1991). *Performance assessment for the workplace, Vol. I.* Washington, DC: National Academy Press.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26–33. ■

Surveys of Programs and Employment in Educational Measurement

Thanos Patelis
Stamford Public Schools
Michael J. Kolen
ACT
Cynthia Parshall
University of South Florida

What is the supply and demand for educational measurement specialists? What types of jobs are more difficult to fill? How can NCME aid recruitment into the profession?

In 1995, the NCME Board of Directors asked the NCME Recruitment of Educational Measurement Specialists Committee to replicate the 1990 surveys of programs and employment in educational measurement (Brennan & Plake, 1990, 1991) to discover whether the demand for measurement professionals was still outstripping the supply. This report presents the results from the surveys conducted during 1996 by a subgroup of the Committee, consisting of the authors of this article.

In addition to the type of information collected in the 1990 surveys, the 1996 surveys allowed for a breakdown of responses by race/ethnicity. Also, in the Employers' Survey, responses were differentiated by job types in order to ascertain whether certain types of jobs were more difficult to fill than others. Finally, a content analysis of open-ended responses was conducted. The results from these content analyses might provide NCME with a basis for developing strategies for recruiting individuals into the profession.

A full report of the results of these surveys (Patelis, Kolen, & Parshall, 1996) is available from the authors and from the NCME Central Office. A brief summary is provided below.

Institutional Survey

In January 1996, a questionnaire was mailed to 152 institutions of higher education with at least one NCME member. This questionnaire was similar to the one mailed in February 1990. However, racial/ethnic data were collected in the 1996 version.

The types of information collected were (a) data on measurement programs and faculty and (b) suggestions for NCME's role in the recruitment of people into the field of measurement. In terms of the data on measurement programs, the numbers of each program were obtained for racial/ethnic category, nationality, and projected year of graduation. In addition, written responses to the open-ended question of reasons for changes in future enrollment were provided. The number of faculty members teaching measurement courses by type (i.e., full-time, part-time or adjunct) were obtained for both current and projected courses.

Responses were obtained from 60 institutions, for a response rate of

Thanos Patelis is a Research Associate at Stamford Public Schools, 888 Washington Blvd., P. O. Box 9310, Stamford, CT 06901. His specialization is educational measurement.

Michael J. Kolen is a Senior Research Scientist at ACT, 2201 N. Dubuque Rd., Iowa City, IA 52243. His specializations are educational measurement and statistics.

Cynthia Parshall is a Psychometrician at the Institute for Instructional Research and Practice, University of South Florida, HMS401, Tampa, FL 33620. Her specializations are computerized testing, psychometrics, and educational measurement.