CHAPTER 7

Performance Assessment: Methods and Characteristics

Contents

Boxes

Figures

Table

CHAPTER 7

Performance Assessment: Methods and Characteristics

Highlights

- Many school districts and States are turning to performance assessment-testing that requires students to create answers or products that demonstrate what they know and can do-as a complement to their traditional testing programs. Thirty-six States now use direct writing samples, and 21 States use other types of performance assessment (in addition to writing samples) on a mandatory, voluntary, or experimental basis.
- Writing samples and constructed-response items, which require test takers to produce an answer rather than select from a number of options, are the most common forms of performance assessment; other methods, such as portfolios of student work, exhibitions and simulations, science experiments, and oral interviews, are still in their infancy.
- Although performance assessment methods vary, they share certain key features. They involve direct observation of student behavior on tasks resembling those considered necessary in the real world, and they shed light on students' learning and thinking processes in addition to the correctness of their answers.
- Performance assessment methods must meet the challenge of producing reliable and valid estimates of student achievement before they can be used for high-stakes decisions involved in system monitoring or selection, placement, and certification. Procedures to reduce subjectivity and eliminate error in human scoring have been developed and used with some success in scoring essays and student writing samples.
- Researchers are developing methods for machine scoring of constructed-response items. Test taking by computer is one approach. Others include having students fill in grids to answer mathematics problems or draw responses on a graph or diagram.
- Advanced information technologies could significantly enhance performance assessment methods: tracking student progress, standardizing scoring, presenting simulations and problems, video recording performance for later analysis, and training teachers are among the most promising possibilities.
- Performance assessment is usually more expensive in dollar outlays than conventional multiple-choice testing because it requires more time and labor to administer and score. However, these high costs might be balanced by the added instructional benefits of teacher participation in developing and scoring tests, and by the closer integration of testing and instruction in the classroom.
- For performance assessment to become a meaningful complement or substitute for conventional testing, educating teachers and the general public will be critical. Teachers need to learn how to use, score, and interpret performance assessments. The public, accustomed to data ranking students on norm-referenced, multiple-choice tests, needs to understand the goals and products of performance assessment.
- *Changing the* format of tests will not by itself ensure that the tests better meet educational goals. However, since what is tested often drives what is taught, testing should be designed to reflect outcomes that are desired as a result of schooling.

Introduction

Springdale High School, Springdale, Arkansas. Spring 1990. Instead of end-of-year examinations, seniors receive the following assignment for a required "Final Performance Across the Disciplines'

Discuss behavior patterns as reflected in the insect world, in animals, in human beings, and in literature. Be sure to include references to your course work over the term in Inquiry and Expression, Literature and the Arts, Social Studies, and Science. This may draw upon works we have studied, including *Macbeth,* Stephen Crane's poetry, Swift's "A Modest Pro-

> posal" and other essays, Mark Twain's fiction, materials from the drug prevention and communication workshop, or behaviors you have observed in school. You may also add references to what you have read about in the news recently. On day 1 of the examination you will be given 4 periods in which to brainstorm, make an outline, write a rough draft, and write a final copy in standard composition form. You will be graded not only on how well you assimilate the material but also how well you reflect our ''student as worker" metaphor and how responsibly you act during the testing period. On day 2 of the examination, you will assemble in villages of three, evaluate anonymous papers according to a set of criteria, and come to a consensus about a grade. Each paper will be evaluated by at least two groups and two instructors. Part of your overall semester grade will reflect how responsibly you act as a member of a team in this task.[1]

Constable Elementary School, South Brunswick, New Jersey. Fall 1990.[2] Every morning, between 10:30 and 11:50, first grade teacher Sharon Suskin settles her class down to a quiet activity supervised by an aide while she calls one student at a time up to her table. With Manuel she says: "I'm going to read you this story but I want you to help me. Where do I start to read? ''As the shy 6-year-old holds the book right side up and points to the print on the first page, she smiles and continues: ''Show me where to start.' She puts a check on her list if he begins at the top left, another if he moves his finger from left to right, another for going page by page. When it is Joanna's turn, she asks her to spell some words: "truck," "dress," "feet." Mrs. Suskin makes a note that, while last month Joanna was stringing together random letters, she now has moved into a more advanced phonetic spelling-' 't-r-k", "j-r-s" and "f-e-t"—representing the sounds in a word. Mrs. Suskin spends anywhere from 2 to 10 minutes with each child, covering about one-half the class each morning, and files the results in each child's portfolio later in the day. When parents come in for conferences, out comes the portfolio. Mrs. Suskin shows Manuel's parents how far he has come in reading skills; Joanna's parents see records of progress rather than grades or test scores. Mrs. Suskin refers to the portfolio regularly, when grouping students having similar difficulties, or when she wishes to check on special areas where an individual child needs help. It's a lot of work, she admits, but she says it gives her a picture of each child's emerging literacy. She laughs: "It makes me put on paper all those things I used to keep in my head.'

All Over California, Spring 1990.[3] All 1.1 million fifth, seventh, and ninth grade students in California were huffing and puffing, running and reaching. They were being tested in five measures of fitness: muscular strength (pull ups); muscular endurance (sit ups); cardiovascular fitness (a mile run); flexibility (sit and reach); and body fat composition (skin fold measurements). Results were tabulated by age and sex, along with self-reported data of other behavior, such as the amount of time spent watching television or engaging in physical activity. The tasks and standards were known in advance, and local physical education teachers had been trained to conduct the scoring themselves. The results were distressing: only 20 percent of the students could complete four or five tasks at the "acceptable" level. The bad news sent a signal to the physical education programs all over the State. Teaching to this test is encouraged as schools work to get better results on the next test administration. The overall goal is more ambitious-to focus awareness on the need for increasing attention to physical fitness for all students, and to change their fitness level for the better.

Why Performance Assessment?

These vignettes are examples of performance assessment, a broad set of testing methods being developed and applied in schools, districts, and sometimes statewide. This concept is based on the premise that testing should be more closely related to the kinds of tasks and skills children are striving to learn. Emotionally charged terms have been applied to this vision of testing. ''Authentic, ' "appropriate," "direct," and even "intelligent" assessment imply something pejorative about multiple-choice tests. This rhetoric tends to ignore that certain multiple-choice tests can provide valuable information about student achievement. OTA uses the more

---

[1]Brown University, The Coalition of Essential Schools, *Horace,* vol. 1, *No. 6,* March 1990, p. 4.

[2]Rem Ruth Mitchell and Amy Stempel, Council for Basic Education, ''Six Case Studies of Performance Assessment, " OTA contractor report, March 1991.

[3]Dale Carlson, ' 'what's New in Large-Scale Performance Testing, ' paper presented at the Boulder Conference of State Testing Directors, Boulder, CO, June 10-12, 1990.

neutral and descriptive term "performance assessment" to refer to testing that requires a student to create an answer or a product that demonstrates **his or her knowledge or skills.**

The act of creating an answer or a product on a test can take many forms. Performance assessment covers a range of methods on a continuum, from short-answer questions to open-ended questions requiring students to write essays or otherwise demonstrate understanding of multiple facts and issues. Performance assessment could involve an experiment demonstrating understanding of scientific principles and procedures, or the creation and defense of a position in oral argument or comprehensive performance. Or it may mean assembling a portfolio of materials over a course of study, to illustrate the development and growth of a student in a particular domain or skill (see ch. 1, box l-D).

Whatever the specific tasks involved, this move toward testing based on direct observation of performance has been described by some educators as ''nothing short of a revolution' in assessment.[4] **Given that performance assessment has been used in businesses and military training for many years, and by teachers in their classrooms as one mechanism to assess student progress, the real revolution is in using performance assessment as a part of large-scale testing programs in elementary and secondary schools.**

The move toward alternative forms of testing students has been motivated by new understandings of how children learn as well as changing views of curriculum. Recent research suggests that complex thinking and learning involves processes that cannot be reduced to a routine,[5] that knowledge is a complex network of information and abilities rather than a series of isolated facts and skills, According to this research, students need to be able to successfully engage in tasks that have multiple solutions and require interpretive and nuanced judgments. This kind of performance in real-world settings is inextricably supported and enriched by other people and by knowledge-extending artifacts like computers, calculators, and texts.[6]

Photo credit: Norwalk *High School, Norwalk, CT*

Performance assessment often involves direct observation of students engaged in classroom tasks. For example, examinations that require students to plan, conduct, and describe experiments reinforce instruction that emphasizes scientific understanding through hands-on activities.

This view of learning challenges traditional views of how to structure curricula and teach, and therefore also how to evaluate students' competence. If knowledge is linked in complex ways to situations in which it is used, then testing should assign students tasks that require interpretation and application of knowledge. If instruction is increasingly individualized, adaptive, and interactive, assessment should share these characteristics. However, educators trying to implement curricular innovations based on this more complex view of learning outcomes have found their new programs judged by traditional tests that do not cover the skills and goals central to their innovations. Many say that school reform without testing reform is impossible. For example, the National Council of Teachers of English recently warned that: ". . . school restructuring may be doomed unless it helps schools move beyond the limitations of standardized tests. '

[4]Jack Foster, secretary for Education and Humanities, State of Kentucky, personal communication, Mar. 11, 1991.

[5]See also ch. 2; and Center for Children and Technology, Bank Street College, "Applications in Educational Assessment: Future Technologies, " OTA contractor report, February 1990.

[6]Additional interest in increased teaching of more complex thinking skills comes not only because of disappointing evidence about students' abilities, but also because of the belief that all workers will require these adaptive capabilities, i.e., the ability to apply knowledge to new situations.

7New York State United Teachers Task Force on Student Assessment, "Multiple Choices: Reforming Student Testing in New York State," unpublished report, January 1991, p. 12: citing the 1990 National Council of Teachers of English, *Report on Trends and Issues*.

Educators advocating performance assessment are also interested in the possibility of making good assessment a more integral and effective part of the learning process. These advocates hope that standardized performance-based testing can become a helpful part of classroom learning rather than a distraction or a derailment of classroom practices. In this view, time spent studying or practicing for tests, or even going through the tests themselves, is no longer seen as time away from valuable classroom learning but rather an integral learning experience.[8]

Indeed, some proponents of performance assessment suggest that its strongest value lies in how it can influence curriculum and instruction by modeling desired educational outcomes. Although ''teaching to the test'' is disparaged when a test calls for selection of isolated facts from a multiple-choice format, it becomes the modus operandi in performance assessment. Perhaps the prime reason for the popularity of performance assessment today stems from the idea that student learning should be guided by clear, understandable, and authentic examples that demonstrate the desired *use* of knowledge and skills. Assessment is then defined as the tool to judge how close the student has come to replicating the level of expertise modeled in the examples. The theory is that performance assessment is an effective method for clarifying standards, building consensus about goals, and delivering a more cohesive curriculum throughout a school system.

As States and districts begin to change their educational goals and curricula, student assessments are also being revised to meet these changing standards and goals. Educators have always recognized that traditional multiple-choice tests do not capture all the objectives valued in the curricula. Some testing programs have attempted to overcome this problem by incorporating some open-ended tasks. However, the increasing stakes attached to traditional test scores has given the tested objectives a great deal of attention and weight in classrooms, often at the expense of objectives that are valued but not directly tested. Policymakers have become interested in tests covering a much wider range of skills and educational objectives, and in various forms of performance assessment that can broaden educational outcomes.

The real policy issue is not a choice between performance assessment and multiple choice, but using tests to enrich learning and understand student progress. Embracing performance assessment does not imply throwing out multiple-choice tests; most States are looking to performance assessment as a means of filling in the gaps. The skills that are not usually evaluated on multiple-choice tests—writing, oral skills, ability to organize material, or perform experiments-have been the first candidates for performance assessments. New York's position is illustrative:

> Student performance assessments should be developed as a significant component of the state's system of assessment. These assessments would include improved multiple-choice tests and incorporate authentic 'real-life' measures of student knowledge. Student performance, judged against clearly defined standards of excellence, would better measure the skills of critical thinking, reasoning, information retrieval and problem solving. Such performance assessments could include portfolios, hands-on problem-solving projects, and demonstrations of ability and knowledge.[9]

State Activities in Performance Assessment

State and local districts have rapidly adopted performance assessment for a range of grade levels and testing objectives. OTA estimates that, as of 1991, 36 States were assessing writing using direct writing samples (see figure 7-l); in addition, 21 States had implemented other types of performance assessment on a mandatory, voluntary, or experimental basis[10] (see figure 7-2). At the present time, most performance assessments are on a pilot or voluntary basis at the State level. When mandated statewide, performance assessments tend to be administered in one or two subjects at selected grade levels.

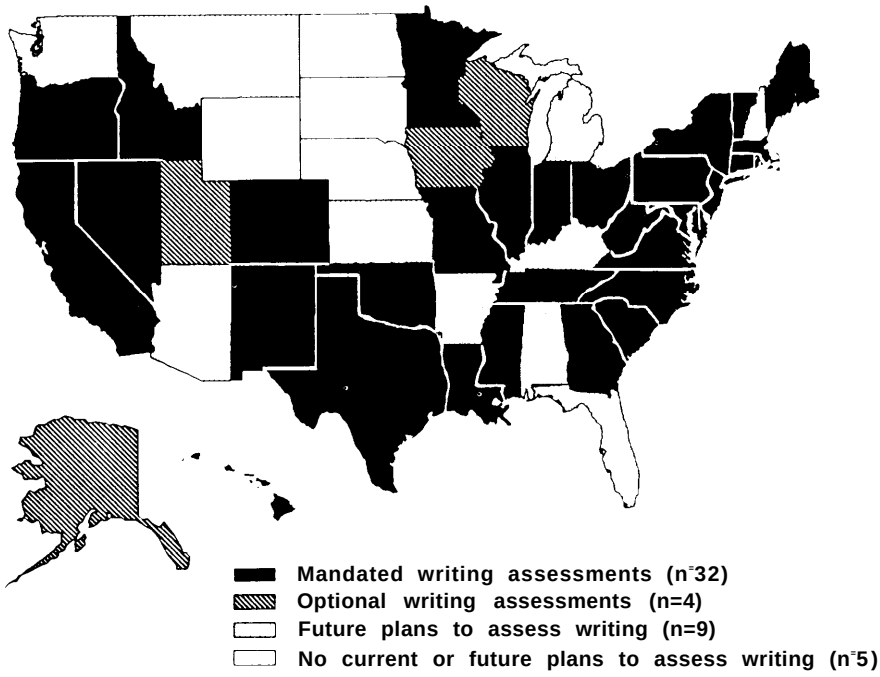
---

[8]This issue has important implications for the estimation of costs associated with alternative testing programs. See discussion in ch. 1.

[9]New York State United Teachers Task Force on Student Assessment, op. cit., footnote 7, p. 4.

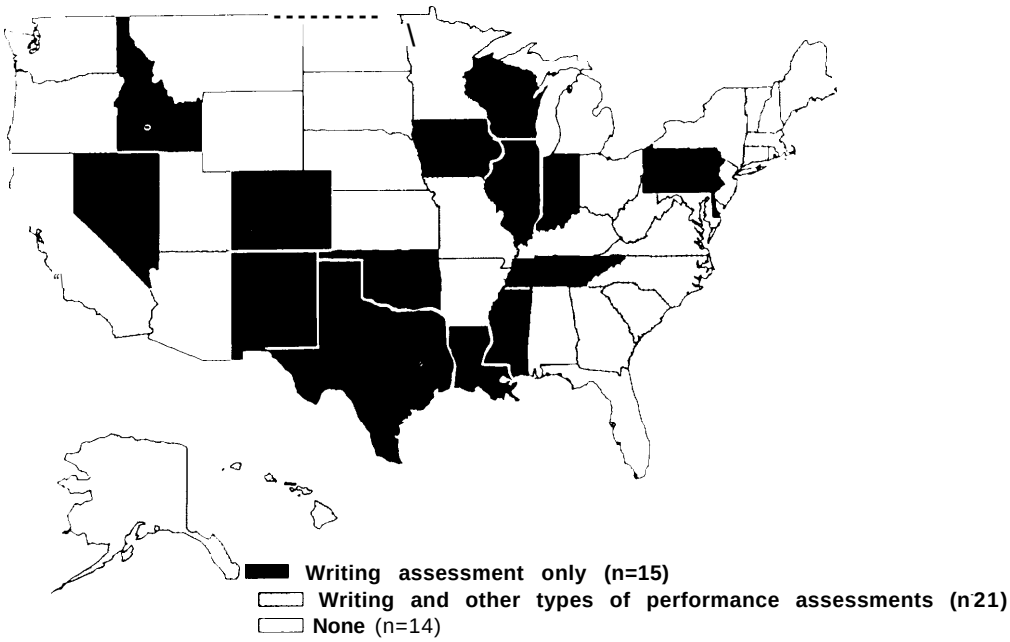
[10]Office of Technology Assessment data, 1991. The category of writing assessments includes just those tests that evaluate student writing skills by asking them to write at some length (paragraphs or essays); other performance assessments reported by States included portfolios, exhibitions or activities, and open-ended paper-and-pencil tests that include student-created answers. This last category includes student essays designed to test knowledge on a particular subject, not testing writing skills per se.

Figure 7-l-State Testing Programs: Direct Sample Writing Assessments, 1991



NOTE: "Future plans" includes current pilot programs.
SOURCE: Office of Technology Assessment, 1992.

Figure 7-2—Statewide Performance Assessments, 1991



NOTE: Map includes optional programs.
SOURCE: Office of Technology Assessment, 1992.

Seven States (Arizona, California, Connecticut, Kentucky, Maryland, New York, and Vermont[11]) are moving their educational evaluation systems toward performance assessment, gradually reducing reliance on norm-referenced multiple-choice testing. Each State has approached the change differently, but they view performance assessment as a tool not only for understanding the progress of individual students, but also for school, district, or State accountability. These State efforts will exert a tremendous influence as comparisons and rankings between schools develop, and policy decisions are made as a result of these new testing results.

The variety of approaches in State testing policies stands in contrast to the traditional State processes for test selection. Historically, State departments of education selected tests with little or no input from teachers or the public. The testing division would invite publishers to bid on the development of a norm-referenced or criterion-referenced test based on the State's curriculum, or, more commonly, shop around and then purchase ''off-the-shelf' tests such as the Iowa Tests of Basic Skills, Stanford Achievement Tests, California Achievement Tests, or other popular norm-referenced achievement tests.[12] This process is changing.

The State profiles in boxes 7-A, 7-B, and 7-C provide a picture of how some States are moving toward greater use of performance assessment in their statewide testing programs. They illustrate the motivation behind these changes, as well as problems and barriers States face in implementing these changes.

The Many Faces of Performance Assessment: Forms and Functions

Performance assessment can take many forms. The central defining element in all performance assessment methods is that the test taker creates an answer or product to demonstrate knowledge or skills in a particular field. From paper-and-pencil, short-answer questions to essays requiring use of knowledge in context, oral interviews, experiments, exhibitions, and comprehensive portfolios with multiple examples of a student's work over a period of an entire year or longer, each type has its own characteristics. Nonetheless, many characteristics are shared. This section describes some of the common *forms* of performance assessment used in K-12 schools today. It is followed by a section that summarizes the common *characteristics* of performance assessment.

Constructed-Response Items

Paper-and-pencil tests designed by teachers have long been a regular feature of the classroom; teachers typically employ a range of item types that include mathematics calculations, geometry proofs, drawing graphs, fill-in-the-blank, matching, definitions, short written answers, and essays. Except for multiple choice and essays, few of these item types have been used for large-scale standardized testing programs, but test developers and educators have begun to consider this possibility.

The term **constructed-response (CR) item** is commonly used to distinguish these items from items such as multiple choice that require selecting a response among the several options presented. CR items require students to produce or construct their own answers.[13]

Several educational advantages might be gained by expanding the use of CR items .14 First, they have higher face validities: they look more like the kinds of tasks we want children to be able to do. Second, these item types may do a better job of reflecting the complexity of knowledge, because they can allow partial credit for partial understanding. Third, these item types may enhance the reliability and validity of scores because they eliminate guessing and other

[11] Vermont did not require statewide testing prior to 1990. The introduction of performance assessment through portfolios in mathematics and writing is the first mandated statewide testing.

[12] See ch. 6 for further discussion of norm-referenced testing.

13A group of researchers at the Educational Testing Service has attempted to describe a framework for categorizing sOme of these item types. These researchers have ordered a number of such item types along an "openness" continuum that includes selection/identification, reordering/rearrangement, substitution/correction, completion and construction. See Randy E. Bennett, William C. Ward, Donald A. Rock, and Colleen LaHart, "Toward a Framework for Constructed-Response Items," ETS research report RR 90-7, 1990.

[14] Ibid.; and James Braswell and J. Kupin, ''Item Formats in Mathematics,' *Construction Versus Choice in Cognitive Measurement,* R.E. Bennett and W.C. Ward (eds.) (Hillsdale, NJ: L. Erlbaum Associates, in press).

Box 7-A—The Arizona Student Assessment Program[1]

Arizona revised its curriculum substantially and then discovered that existing State-mandated tests were no longer appropriate. Teachers carried a heavy annual testing burden, but remained unsure how the various tests corresponded to what they were expected to teach. Describing the old State-mandated testing required in grades 1 through 12 every spring, using the Iowa Tests of Basic Skills (ITBS), Tests of Achievement and Proficiency (TAP), and district testing under the Continuous Uniform Evaluation System (CUES), one teacher expressed frustration:

> We have these CUES tests, pre- and post-test. . . , In one grade we have 135 little skills tests in all of those forms, pre- and post-test. We teach what we think is important to teach. . . until right before our CUES tests. Then we teach students how to do well on the CUES tests. We also give the Iowa Tests of Basic Skills and it takes about a week. We teach what we think is important all year long. . . until right before the ITBS. Then we teach students how to take the ITBS. . . . We get the scores back on the ITBS right before students leave for the summer, and I usually have to follow students out the door on the last day with a stapler in one hand and the test scores in the other so I can staple the score reports onto their report cards. We have an entirely different group of students over the next year so that it doesn't do much good to analyze the test scores over the summer. . . . I feel confused. What are we supposed to teach? What is valued? It seems to me we are spending a great deal of time getting ready for two measures that are at odds with what we have agreed in my district is important to teach.[2]

Statewide curriculum frameworks, known as Essential Skills Documents (ESDs), were developed starting in 1986, to outline broad competencies and goals at the elementary, middle, and high school levels across the State.[3] Most teachers enthusiastically embraced the documents but some lamented: "That's the way I'd like to teach . . . if it weren't for the way we test. ' Reflecting this concern, the State legislature set up a joint committee in 1987 to review the overall teaching and assessment program in the State, looking especially to see if the skills and processes identified in the Essential Skills curriculum frameworks were being successfully acquired by Arizona students.

An independent committee analyzed whether the skills required in the ESDs were being assessed in the ITBS and TAP. Results for mathematics, reading, and writing indicated that only 20 to 40 percent (with an average of 26 percent) of the Essential Skills were assessed by the ITBS and TAP. Thus, even with annual testing for all grades, Arizona was only receiving information on how well students were mastering one-quarter of the content of the new curriculum. As one teacher said:

> The teachers in Arizona can't serve two masters. If they want the teachers to do a good job of teaching math they can use the Essential Skills Documents . . . and throw out the ITBS tests, or teach the ITBS tests and throw out the Essential Skills Documents.[5]

With the support of teachers, school boards, administrators, and the business community, the legislature passed State Law 1442 by a landslide. The act required the Arizona Department of Education to create an assessment plan that would do a better job of testing the Essential Skills. Thus the Arizona Student Assessment Program (ASAP) was born in the spring of 1990, setting a new approach to State testing.

ASAP is an umbrella program composed of new performance measures, continuing but reduced emphasis on norm-referenced testing, and extensive school, district, and State report cards. Riverside Publishing Co., the same company that produces the TAP and ITBS, was selected to produce the new assessments at the benchmark grades of 3, 8, and 12 in each of the three subject areas. To best match the goals of the ESDs, the new tests were to be performance- and curriculum-based assessments. The language arts assessment is an interesting example. Paralleling the way writing is taught under the language arts framework the assessment is a two-step process. On the first day of testing, students engage in the steps that make up the "rewriting" process (e.g., brainstorming, listing, mapping, or "webbing" ideas) and creating a first draft; on the second day of testing, they reread the draft,

---

[1]Much of this discussion is taken from Ruth Mitchell and Amy Stempel, Council for Basic Education, "Six Case Studies Of Performance Assessment" OTA contractor report, March 1991.

[2]Lois Brown Easton, ' 'Developing Educational Performance Tests for a Statewide Program, " *Educational Performance Assessment*, Fred L. Finch (cd.) (Chicago, IL: Riverside Publishing Co., 1991), p. 47.

[3]The language arts framework was published in 1986 and the mathematics framework in 1987; by the end of 1990, Essential Skills Documents were available in 12 subjects including, in addition to the above, frameworks in science, health, social studies, and the arts. Mitchell and Stempel, op. cit., foomote 1.

[4]Easton, op. cit., foomote 2.

[5]Arizona Department of Education, *Arizona Essential Skills for Mathematics* (Phoenix, AZ: July 1987), p. i.

Continued on next page

Box 7-A—The Arizona Student Assessment Program[1]--Continued

revise, and write a second draft. Similar performance-based assessments have been created for mathematics and reading, with science and social studies assessments also under development.

The first official assessment will be implemented in March 1992 and scored by teachers at regional scoring sites, none more than an hour's drive from any district. Classroom teachers are being trained and certified as scorers, and will receive a small stipend and graduate credit for their work In pilot scoring sessions, scoring was found to be reliable between readers as well as consistent when a reader was given the same paper to score more than once. Scoring also took less time than expected.[6] Having the classroom teachers score the examinations is seen as a positive staff development activity, as teachers become involved in setting common quality standards and in sharing the review process with their colleagues from around the State.

Norm-referenced tests (NRTs) are being continued as a way to compare Arizona's student achievement against a national testing reference. However, their influence is being reduced Students will take only a part of the ITBS and TAP each year (i.e., subtests, rather than the full test battery), reducing test-taking time overall by one-half to two-thirds.[7] The norm-referenced testing will be moved from spring to fall, further reducing their impact Scores derived from spring testing had been considered a reflection of what the teachers taught over the past year, even if the test content did not always correspond to what was actually taught. Teachers often felt pressured to spend considerable time preparing students for the spring tests. With fall testing, both teachers and students should face the tests with more equanimity, and there will be less pressure to "prep" students. Fall testing also means that scores will be returned in time to be used for that year's instructional planning.

The third component of ASAP changes the way school and district achievement will be reported. Previously, each July things got "hot" in Arizona, as newspaper stories listed every school in a district alongside their test scores on the TAP and ITBS. Little interpretative information was provided and the message was implicit-the higher the score, the better the school. The new reporting system will try to paint a more realistic picture of achievement at the school, district, and State level. These annual "Arizona Report Cards' will report Essential Skills scores, NRT scores, and other factors that reflect achievement (e.g., numbers of students in advanced courses, science fair winners, and special award winners). However, to set these in context, factors that *affect* achievement are also reported, such as student socioeconomic status, mobility rate, percentage of students with limited English proficiency, and faculty turnover rates. Although it is assumed that school and district comparisons will continue to be made, it is hoped that these comparisons will be made on a more meaningful and realistic cluster of factors.

When the new program was introduced to teams of 850 teachers from across the State at a 3-day conference in October 1990, teacher reaction was mixed. Although many were pleased with the new approach, they were concerned with the difficulty of putting the new system into place. As one said: "The staff development needs are incredible. We need staff development on pedagogy, on writing, on logic, everything. To do this in the timeframe we have, we need big bucks. "

Assessment costs are difficult to determine because the change in assessment is aligned to changes across the system-especially curriculum development and professional development. Money saved from less ITBS and TAP testing will be used for all three parts of the ASAP in coming years--NRTs, performance assessments, and nontest indicators. Nevertheless, costs for the program (the request for proposal for developing the new performance-based assessments, the statewide teacher conference, preparing teacher scorers, and training all teachers in the new system) will be substantial. While perhaps an expensive gamble, the State commitment to move forward indicates the priority Arizona legislators and educators have placed on introducing a new approach to assessment throughout the State.

---

[6] Easton, op. cit., footnote 2, P. 56.

[7] Ibid., p. 57.

"back door" approaches, such as strategies of elimination or getting cues from incorrect choices. Fourth, some of these items can use scoring methods that recognize the correctness of a variety of different answers, representing the complexity of understanding and knowledge. (This suggests the potential diagnostic value of CR items. These items can reveal the processes used by the learner; e.g., a scorer can examine the student's problem-solving steps and detect errors in reasoning or misconceptions). And, finally, one of the most often cited (but least documented) assumptions is that these items

Box 7-B—Kentucky's Performance Assessments and Valued Outcomes[1]

Kentucky is fundamentally redesigning its State educational system. When the 1990 Kentucky Education Reform Act is fully implemented, the State will have the first system that measures student achievement entirely b y performance-based testing. **It will also be unique in the emphasis placed** *on these* **tests: schools will be rewarded and punished based on test results.**[2]

In rethinking basic educational practices and premises, Kentucky educators hope to give classroom teachers a larger voice and improved ability to report on what they believe **a student has achieved. They hope to move away from the common model that values the results of State-administered norm-referenced tests more highly than classroom-based testing and teacher's grade cards. The goal is to integrate teaching with assessment so it is almost invisible to the student, minimizing the use of external instruments as much as possible. The Kentucky approach will require extensive training of teachers as well as a backup system to ensure quality control.**

Under the guidance of a Council on School Performance Standards, 11 task forces involving some 1,000 educators are working to identify the activities needed to define expected student outcomes and set the level of proficiency desired at three "anchor points": the 4th, 8th, and 12th grades. Teachers will continually evaluate students on a less formal basis in the interim grades to be sure progress is being made by all students as they prepare for the benchmark performance levels. Additionally, as younger children watch the performance of older peers, they will be encouraged to model themselves on the older students and see how close they are to that level of proficiency. This approach is based on a sports metaphor, with the students participating in "scrimmages" that involve practice tests at earlier grade levels. Younger students are similar to the "junior varsity" as they become motivated by and learn from watching the "varsity," older students at higher levels of performance.

Benchmark grades will be tested each year but reported every other year for accountability purposes, Successful schools will receive monetary rewards from the State; unsuccessful schools will be required to develop plans for improvement. If a school is particularly unsuccessful, it may be declared a "school in crisis" and its students may be permitted to transfer to more successful schools or administrators may be replaced and "distinguished educators" may be brought into help.[3]

In the summer of 1991, a contractor was selected to create the 1995-96 performance assessments in language arts, science and technology, mathematics, social studies, arts and **humanities,** practical living, and vocational studies. Development costs over the first 18 months are estimated to be approximately $3.5 million. An interim testing program administered to a sample of students during the 1991-92 school year will provide baseline data for school success during 1993-94. The interim test has been controversial because of its traditional nature; some fear it could sidetrack implementation of the full program of performance-based measures.

---

[1]Much of this box is taken from Kentucky Department of Education, "Request for proposals to Implement an Interim and Full-Scale Student Assessment Program for the Commonwealth of Kentucky," March 1991; and Jack Foster, secretary of Education, Kentucky, personal communication, June 1991.

[2]"Update," 10, No. 40, July 31,1991, p. 33.

[3]Mary Helen Miller, Kevin Noland, and John Schaff, *to the Kentucky of 1990* (Frankfort, KY: Legislative Research Commission, April 1990), p. 5.

tap more sophisticated reasoning and thinking processes than do multiple-choice items.

California has been a pioneer in the effort to use open-ended CR items. In 1987-88, the State piloted a number of open-ended mathematics problems as part of the 12th grade State test. Some of the questions were intentionally structured to be broad to allow ''. . . students to respond creatively, demonstrate the full extent of their mathematical understanding, and display the elegance and originality of their thought processes."[15] One such question, along with representative answers, is pictured in figure 7-3. As the sample answers suggest, some students demonstrated a high degree of competence in mathematical reasoning while others displayed misconceptions or lack of mathematical understanding. Sixty-five percent of the answers to this question were judged to be inadequate, leading the developers to surmise that: ". . . the inadequate

[15]California State Department of Education, ''A Question Of Thinking: A First Look at Students' Perfo rmance on Open-Ended Questions in Mathematics," unpublished report, 1989, p. 3.

Box 7-C—The California Assessment Program: Testing That Goes "Beyond the Bubble"[1]

The California Assessment Program (CAP) was created in 1974-75 as part of an early school reform program. **It has evolved over the years to reflect changes in curricula, student population, and pressures for accountability, but CAP continues to be seen as a model for other States, primarily due to two factors: the State carefully defined curricular objectives as the starting point for assessment, and devoted considerable research and support to the development of new forms of assessment.**

Bringing education reform to a State as large as California, larger in population than many European countries, has been a monumental task. The main vehicle for change has come with the creation of statewide curriculum frameworks--documents developed starting in 1983 in response to a major school reform bill. These curriculum guidelines and frameworks have been modified over time and now center on developing students' ability to think, to apply concepts, and to take responsibility for their own learning. The frameworks mandate a curricula that is . . . literature-based, value-laden, culturally rich, and integrated across content areas."[2] Writing across the curriculum, cooperative learning, experiential learning, problem **solving are emphasized Although the frameworks are not mandated, they are the basis for the madated CAP assessments, creating indirect pressure on districts to align the curriculum and instruction.**

It became **clear that much of what was to be taught with the new frameworks would not be taught or assessed appropriately if student achievement was evaluated with existing multiple-choice tests. A shift to performance assessment was sought to bring curriculum and instruction in line with the frameworks. The first performance assessment component, a direct writing assessment, was developed by teachers and put into place in 1987. Each year several hundred teachers gather over a 4- to 6-day period at four sites across the State to score the essays. Teacher scoring is emphasized to enhance the connection between instruction and assessment.**

The success of the effort seems to validate this connection and meet expectations. One report suggests that: " . . . "educators throughout California have expressed the belief that no single program has ever had statewide impact on instruction equal to that of the writing assessment. "[3] A study at the completion of the first year of the writing assessment found that 78 percent of the teachers surveyed reported they assigned more writing, and almost all (94 percent) assigned a greater variety of writing tasks.[4] The percentage of students who reported that they wrote 11 or more papers in a 6-week period jumped from 22 to 33 percent. The writing assessment has also motivated a huge increase in staff development, with the California Writing Project training over 10,000 teachers in support of improved instruction in writing.[5]

In December 1989, California held an Education Summit, in response to the National Education Summit of the Nation's Governors in Charlottesville, Virginia. In seeking areas most likely to produce significant change ("targets of opportunity"), and building on the strengths of the California system, the educators called for statewide performance goals that would be measured through a strengthened assessment system. The report stated:

> *The* fundamental objectives of educational testing in California schools are far from fulfilled. The *methods and formats not only fail to support the kind of teaching and learning that the state and national curriculum reform movement calls for, but actually retard that movement in California.* **Students, teachers, and parents are not** getting the necessary information to gauge the educational system's progress, detect strengths and weaknesses, improve instruction, and judge overall effectiveness. . . . The current approach to assessment of student achievement which relies on multiple choice student response must be abandoned because of its deleterious effect on the educational process. An assessment system which measures student achievement on performance-based measures is essential for driving the needed reform toward a thinking curriculum in which students are actively engaged and successful in achieving goals in and beyond high school.[6]

---

[1] **Ruth Mitchell and Amy Stempel, Council for Basic Education, "Six Case Studies** of Performance **Assessment," OTA contractor report,** March 1991.

[2] **North Central Regional Educational Laboratory and Public Broadcasting Service, "Multidimensional Assessment Strategies for Schools," Video Conference 4, 1990, p. 27.**

[3] **California Assessment Program, "California: The State of Assessment," draft report, Apr. 3,** 1990P. **8.**

[4] **An evaluation of the grade eight writing assessment** by **the National Center** for the **Study** of Writing at the University **of California, Berkeley, cited in ibid.**

[5] **Ibid., p. 8.**

[6] **California Department of Education,** **final** report **(Sacramento,** CA: February 1990).

The direct writing assessment was cited as an example of the kind of assessment needed to drive program improvements. The summit thus gave support and further stimulus for continuing research and piloting of new methods.

In the past, statewide testing used matrix sampling, in which each student takes only a portion of the test and scores are reported on the school or district level, but not for individual students. However, recent legislation mandates that beginning in 1992-93 individual testing will be conducted statewide in grades 4,5,8, and 10 in basic skills and content courses. The use of direct writing assessment and other performance-based assessments is encouraged. Districts can also choose their own student tests at other grade levels. All testing is to be aligned to the California curriculum frameworks, with reporting based on common performance standards. The new program gives special emphasis to end-of-course examinations for secondary school subjects. These will be based on the existing Golden State Examinations, which students now take on a voluntary basis at the completion of Algebra, Geometry, Biology, Chemistry, U.S. History, and Economics. Districts may require that all students take one or more Golden State Examination. Finally, the integrated student assessment system will also include a portfolio for all students graduating from high school. The portfolio will contain documentation of performance standards attained on the grade 10 test (or other forms of the test taken in grades 11 and 12), on end-of-course Golden State Examinations, and on vocational certification examinations, as well as evidence of job experience and other valued accomplishments. [8]

This represents a big jump in required testing. Performance-based components are defined as building blocks for all the tests, both CAP and district-administered. CAP has indirectly influenced the testing done at the district level by ". . . opening the door. . . giving permission to go ahead with performance assessment."[9] CAP also has pilot projects for portfolios in writing and mathematics, and research studying the impact on instruction of open-ended mathematics questions.

Developing performance-based assessments is not a simple task. At the 1987 "Beyond the Bubble" conference on testing, educators grappled with the issue of developing new ways to produce alternative assessments that more directly reflect student performance. A suggestion to support grassroots efforts by teachers with assistance from assessment experts eventually led to the Alternative Assessment Pilot Project. In 1991, the Governor authorized $1 million to implement its provisions, and two consortia of California school districts (one in the north and one in the south of the State) have been given grants totaling over $965,000 to begin the project. Each consortium will develop, field test, and disseminate alternatives to standardized multiple-choice tests for assessment of student achievement. At the school level, teachers will develop their own materials and strategies and pilot them with their own classrooms and schools, sharing information with other teachers across the State. A cost-benefit analysis of the local use of current performance-based assessment systems will also be conducted.[10]

Because of the scope of these endeavors, many other States are looking to the California experiment as a guide to their own efforts to realign testing and curriculum.

[7]Chapter 760, California Statutes of 1991 (SB 662; Hart).

[8]Superintendent Honig, California State Department of Education, "New Integrated Assessment System," testimony before the State Assembly Education Committee, background information Aug. 21, 1991.

[9]Ruben Carriedo, director of Planning, Research and Evaluation Division, San Diego City Schools, cited in Mitchell and Stempel, op. cit., footnote 1, p. 17.

[10]California Department of Education News Release, Aug. 2, 1991.

responses of a large number of students occurred primarily because students are not accustomed to writing about mathematics. ' '[16]

The National Assessment of Educational Progress (NAEP) has also successfully utilized a variety of open-ended items. In the 1990 NAEP mathematics assessment, about one-third of the items included open-ended questions that required students to use calculators, produce the solution to a question, or explain their answers. The 1990 reading test, which also employed text passages drawn from primary sources, including literary text, informational text, and documents, used a number of short essays to assess the student's ability to construct meaning and provide interpretations of text. The 1985-86 NAEP assessment of computer competence included some

[16]Ibid., p. 6.

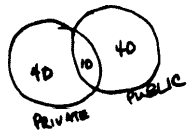
Figure 7-3--Open-Ended Mathematics Item With Sample Student Answers

QUESTION: James knows that half of the students from his school are accepted at the public university nearby. Also, half are accepted at the local private college. James thinks that this adds up to 100 percent, so he will surely be accepted at one or the other institution. Explain why James may be wrong. If possible, use a diagram in your explanation.

Good Mathematical Reasoning: Sample Answers

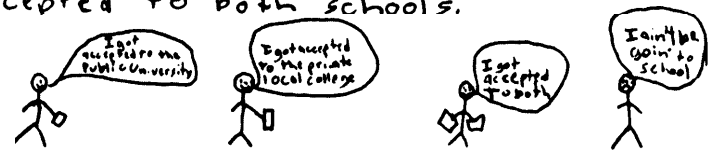
James THINKS this adds up to 100%, but there may be students who are accepted to both institutions, thus leaving James out in the cold.

For example, say there are 100 people at his school. Half get accepted at Public U., & the other half at Private U.



50 were accepted at Private U. 50 were accepted at Public U. But this only accounts for 90% of the students. Jim could be in the other 10%, thus not getting accepted

Some of the people accepted may have been accepted to both schools.



Misconceptions: Sample Answers

If 175 students apply, and ½ are accepted to the public university and ½ are accepted to the local private college.

½ of 175 is 87. (175)
87 go to the university 87 x 2 = 174
87 go to the college
leaving one student out, which can be James

Thats wrong because everyone doesnt go to college I think its

15% doesn't go
45 That goes to local college
40 That goes to private college

NOTE: Used in the 1967-86 version of the 12th grade California Assessment Program test, this logic problem assesses a student's ability to detect and explain faulty reasoning. Answers are scored on a O to 6 point scale. The student must give a dear and mathematically correct explanation of the faulty reasoning. For the highest score, responses must be complete, contain examples and/or counter examples of overlapping sets, or have elegantly expressed mathematics. A diagram is expected.

SOURCE: California State Department of Education, *A Question of Thinking: A First Look at Students' Performance on Open-Ended Questions in Mathematics* (Sacramento, CA: 19S9), pp. 21-28.

open-ended items asking students to write short computer programs or indicate how the "turtle" would move in response to a set of computer commands; students were given partial credit for elements of a correct response.

Scoring: Machines and Judges

Researchers and test developers are now considering ways to streamline available methods for scoring the more open-ended CR items. One promising area involves new types of CR items that can be entered on paper-and-pencil answer sheets and scanned by machines.[17] One such item type for mathematics problems is the grid-in format. Students solve the problem, write their solution at the top of a grid, and then fill in a bubble corresponding to each number in the column under that number (see figure 7-4). Questions that have more than one correct answer are possible, and the format allows for the possibility of answering in either fractions, decimals, or integers.[18]

"Figural response" items, which require drawing in a response on a graph, illustration, or diagram, were field tested in the 1989-90 NAEP science assessment (see figure 7-5). The feasibility of machine scoring of these items was also tested by using high-resolution image processors to score the penciled-in answers. Some initial technological difficulties were encountered with the scanning process—many student answers were too light to be read and the ink created some interference. However, the researchers express optimism that the scanning mechanism can be made to work.[19]

Researchers are working on technologies of handwriting recognition that will eventually result in printed letters and numbers that can be machine scanned from answer sheets, but these technologies are still far from reliable except under optimal conditions--the letters must be cleanly printed and properly aligned. Systems that can read cursive handwriting are in a more experimental stage; whether the ". . . scrawl likely to be produced under the pressure of examinations' could ever be read by a computer is questionable.[20]

CR items vary considerably in the extent to which they can be scored objectively. More objective items will have scoring rules that are very clear and involve little or no judgment. Other responses, such as short written descriptions or writing the steps to a geometry proof, are more complicated to score-in part because there are multiple possibilities for

Figure 7-4--Machine-Scorable Formats: Grid-In and Multiple-Choice Versions of a Mathematics Item

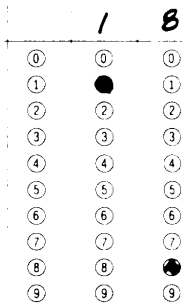
The Question:

Section I of a certain theater contains 12 rows of 15 seats each. Section II contains 10 rows, but has the same total number of seats as Section 1. If each row in Section II contains the same number of seats, how many seats are in each row?

Test 1, Multiple Choice Version

(A) 16
(B) 17
(C) 18*
(D) 19
(E) 20

Test 2, Grid Version



NOTE: This item was designed for high school juniors and seniors.

SOURCE: Educational Testing Service, Policy information Center, ETS *Policy Notes*, voi. 2, No. 3, August 1990, p. 5.

[17] Many of the problems involved in machine scanning are solved if constructed-response items can be delivered via computer. If the students take a mathematics computation test via computer, they can simply type in the correct numbers; a short essay can be written on the keyboard. As a result, the computer is in many ways a more "friendly' system for the delivery of many constructed-response type items, because problems related to scanning in the answer are solved. The machine-scanning problem is much less tractable for items delivered via paper-and-pencil tests. See ch. 8 for further discussion of the issues involved in administering tests via computers.

[18] James Braswell, "AU Alternative to Multiple-Choice Testing in Mathematics for Large-Volume Examination Programs," paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April 1990. Grid-in items for mathematics are currently under development for both the SAT and the ACT college admissions examinations. Preliminary results with college-bound students are encouraging:

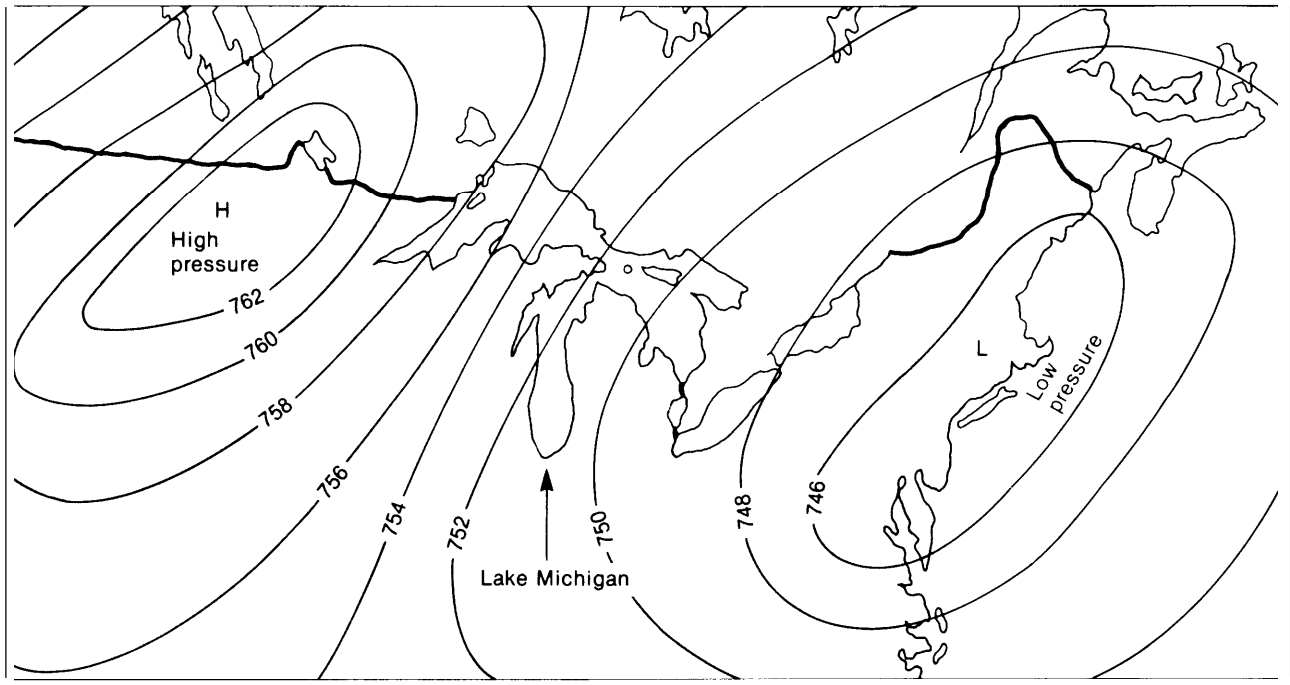
> Guessing and back door approaches to solving mathematics questions are virtually eliminated and the range of answers that students offer to individual questions is great and frequently does not match well with the distracters provided in multiple-choice versions of the same items. As one would expect, the grid-in format requires more time. (p. 1)

[19] Michael Martinez, John J. Ferris, William Kraft, and Winton H. Manning, "Automated Scoring of Paper-and-Pencil Figural Responses," ETS research report RR-90-23, October 1990.

[20] Leslie Kitchen, "What Computers Can See: A Sketch of Accomplishments in Computer Vision, With Speculations on Its Use in Educational Testing," *Artificial Intelligence and the Future of Testing*, Roy Freedle (cd.) (Hillsdale, NJ: L. Erlbaum Associates, 1990), p. 134.

Figure 7-5--Figural Response Item Used in 1990 NAEP Science Assessment

The map below shows a high-pressure area centered over North Dakota and a low-pressure area centered over Massachusetts. Draw an arrow (—►) over Lake Michigan that shows the direction in which the winds will blow.



KEY: NAEP = National Assessment of Educational Progress.
NOTE: This item was used with 8th and 12th graders.

SOURCE: Michael E. Martinez, "A Comparison of Multiple-Choice and Constructed Figural Response Items," paper presented at the annual meeting of the American Educational Research Association, Boston, MA, April 1990.

correct or partially correct answers. Machine scoring of even more complex products, such as the steps in the solution of algebra word problems or computer programming, proves to be much more complicated; preliminary work drawing on artificial intelligence research suggests that automated scoring can eventually be developed. However, the time and cost required to develop such a program is very high. "In both instances, the underlying scoring mechanism is an expert system—a computer program that emulates one or more aspects of the behavior of a master judge."[21]

One of the more difficult and long-term problems of developing artificial intelligence models to score constructed responses is building their capacity for error detection. Programming machines to recognize correct answers is far easier than programming them to detect errors, grade partial solutions, and provide evaluation of error patterns.[22] When questions that allow for more than one right answer are used, programming of the scoring can get quite complicated.[23] Yet one of the highly desirable features of CR items is their potential for diagnosis of misconceptions, errors, and incorrect strategies.[24]

Although most CR items still require human scoring, procedures exist that can eliminate error and make this scoring more reliable. Development of clear standards for judging student answers and

[21] Randy Bennett, "Toward Intelligent Assessment: An Integration of Constructed Response Testing, Artificial Intelligence, and Model-Based Measurement," ETS research report RR-90-5, 1990, p. 5. For a description of artificial intelligence applied to a constructed-response computer programming problem, see Henry I. Braun, Randy E. Bennett, Douglas Frye, and Elliot Soloway, "Scoring Constructed Responses Using Expert Systems,' *Journal of* vol. 27, No. 2, summer 1990, pp. 93-108.

[22] Roy Freedle, "Artificial Intelligence and Its Implications for the Future of ETS's Tests," in Freedle (cd.), op. cit., footnote 20.

[23] Braswell and Kupin, Op. cit., footnote 14.

[24] See Menucha Birenbaum and Kikumi Tatsuoka, "Open-Ended Versus Multiple-Choice Response Formats-It Does Make a Difference for Diagnostic Purposes," *Applied Psychological Measurement*, vol. 11, No. 4, 1987, pp. 385-395.

intensive training of judges until they reach acceptable levels of agreement are important components of establishing high *inter-rater reliability (see* discussion inch. 6). Preliminary indications are that most CR items can be scored with inter-rater reliability equal to or better than that achieved by judges grading essays. The process of training judges to grade essays reliably has been successfully developed in some large-scale testing programs; in addition, many commercial publishers and other companies now offer commercial grading services to schools that want independent and technically supervised rating procedures.

The feasibility of scoring geometry proofs on a large scale has recently been demonstrated by the State of North Carolina. Because an important objective of the high school geometry curriculum in North Carolina was for students to learn to develop complete proofs, the State assessment program included such proofs in the new assessment. All 43,000 geometry students in the State were given two geometry proof questions in the spring of 1989. Over 400 teachers from throughout the State were trained to score the proofs. Drawing on the lessons from the scoring of writing assessments (e.g., the importance of developing scoring criteria and training), high levels of scorer agreement were achieved. Actual time devoted to training was less than 3 hours.[25]

Constructed-Response Items as Diagnostic Tools

One of the features of CR items that makes them attractive to educators is that they allow closer examination of learners' thinking processes. When students write out the steps taken in solving a proof, or a list of how they reached their conclusions, the students' thinking processes can be examined and scored. Results of one study have suggested that CR-type items may be more effective than multiple-choice items for diagnostic purposes; i.e., for uncovering the processes of learners in ways that might help a teacher better understand students' errors or misconceptions.[26]

Not only might errors and misconceptions be more readily uncovered, but students' abilities to generate and construct meaning in complex tasks can also be assessed. The methods for developing these more complex scoring systems are not yet well established or understood. Cognitive research methods (see ch. 2) are beginning to be applied to the development of scoring rubrics for CR-type items. ''Think aloud" methods, where children are closely observed and interviewed while solving open-ended problems, can provide a rich source of information to help build scoring rubrics. Early efforts to generate scoring criteria based on comparing the performance of experts and novices also have been encouraging.[27] One of the challenges for researchers in this area is to develop scoring criteria that have general utility across a number of tasks, instead of being specific to a particular test question or essay prompt.[28]

Although the relative virtues of multiple-choice and CR items have been debated in the educational literature since early in this century, there are few comprehensive empirical studies on the topic. Thus, although there is considerable "textbook" lore about the differences between the two types of items, few generalizations can be made with confidence about differences in student performance.[29] CR items have not been widely field tested in large-scale testing programs. Very few researchers have collected data that allows direct comparison of CR with multiple-choice items.

It is fair to say that no one has yet conclusively demonstrated that CR items measure more ''higher order" thinking skills than do multiple-choice items. "All the same, there are often sound educa-

[25]Zollie Stevenson, Jr., Chris P. Averett, and Daisy Vickers, "The Reliability of Using a Focused-Holistic Scoring Approach to Measure Student Performance on a Geometry Proof, " paper presented at the annual meeting of the American Educational Research Association Boston, MA, April 1990.

[26]Birenbaum and Tatsuoka, op. cit., footnote 24.

[27]See, for example, Kevin Collis and Thomas A. Romberg, "Assessment of Mathematical Performance: An Analysis of Open-Ended Test Items, ' and Eva L. Baker, Marie Freeman, and Serena Clayton, "Cognitive Assessment of History for Large-Scale Testing," Testing *and Cognition*, Merlin C. Wittrock and Eva L. Baker (eds.) (Englewood Cliffs, NJ: Prentice Hall, 1991).

[28]Baker et al., op. cit., footnote 27.

[29]See R.E. Traub and K. MacRury, ''Multiple-Choice vs. Free-Response in the Testing of Scholastic Achievement,' *Tests and Trends 8: Jahrbuch der Padagogischen Diagnostic,* K. Ingenkamp and R.S. Jager (eds.) (Weinheim and Basel, German y: Beltz Verlag, 1990), pp. 128-159; Ross Traub, "On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests,'' in Bennett and Ward (eds.), op. cit., footnote 14; and Thomas P. Hogan, ''Relationship Between Free-Response and Choice-Type Tests of Achievement: A Review of the Literature, ' ERIC document ED 224811 (Green Bay, WI: University of Wisconsin 1991).

297-933 0 - 92 -- 15 : QL 3

tional reasons for employing the less efficient format, as some large-scale testing programs, such as AP [Advanced Placement], have chosen to do. ' '[30]

Essays and Writing Assessment

Essays, particularly when used to assess writing proficiency, are the most common form of performance assessment. In fact, the noun ''essay' is defined as ''trial, test" and the verb as ". . . to make an often tentative or experimental effort to perform."[31] Essays are a relatively well understood testing format, in part because they have been used for many years. An essay is an excellent example of performance assessment when used to assess students' ability to write. Essay questions for assessing content mastery are also a form of performance assessment, because they require student-created products that demonstrate understanding. The problem arises in scoring subject matter essays—are students' understanding of content being masked by a difficulty in written expression? In that case, writing skill can confound scoring for content knowledge.

Essays as Assessments of Content Mastery

Student understanding of a subject has long been assessed by requiring the student to write an essay that uses facts in context. Essay questions have been central to some large-scale testing programs overseas (see ch. 5); they also makeup approximately 60 percent of the questions on the **Advanced Placement examinations administered by the College Board. The essay to show content mastery is in fact the hallmark of classical education; student writing about a subject reveals how fully the student has grasped not only the obvious information but the relationships, subtleties, and implications of the topic. The use of writing as an instructional and testing device is familiar to scholars, and its use by all students is increasingly understood to help develop thinking skills as well as communications skills.**

Students have different expectations about different types of tests. For example, one study found that students report a preference for multiple-choice over essay tests ". . . on the grounds that these tests are easier to prepare for, are easier to take, and hold forth hope for higher relative scores."[32] Other studies have suggested that students study differently for essay tests than they do for multiple-choice tests. For example, one study found that students ". . . consider open questions a more demanding test than a multiple-choice test. . .' and use more study time to prepare for it.[33] However almost no data exist about what students actually do differently when studying for different kinds of tests and evidence is ambiguous regarding whether these different study strategies affect actual achievement.[34]

Essays as Tests of Writing Skill

Many large-scale testing programs have begun the move toward performance assessment by adding a direct writing sample to their tests. One reason for this shift is a concern that the wrong message is sent to students and teachers when writing is not directly tested. According to one researcher of writing ability:

> A test that requires actual writing is sending a clear message to the students, teachers, parents, and the general public that writing should be taught and tested by having students write. Although it may be that a test that includes a writing sample will gain little in psychometric terms over an all-multiple-choice test, the educational gains may be enormous. The English Composition Test, administered as part of the College Board Achievement Tests, contains one 20-minute essay section in the December administration only. At that administration approximately 85,000 students write in response to a set topic, and each of the 85,000 papers must be scored twice. That scoring may cost in the neighborhood of $500,000. The increase in predictive validity for the test is minimal. Admissions officers and others who use the scores are probably not seeing a dramatic increase in the usefulness of scores despite the expenditure of the half million dollars.

---

[30] R.E. Bennett, Donald A. Rock, and Minhwei Wang, "Free-Response and Multiple-Choice Items: Measures of the Same Ability?" ETS research report RR-908, 1990, p. 19.

[31] *Webster's Ninth New Collegiate Dictionary* (Springfield, MA: Miriam Webster, Inc., 1983), p. 425.

[32] Traub and MacRury, op. cit., footnote 29, p. 42.

[33] Gery D'Ydewalle, Anne Swerts, and Erik De Corte, "Study Time and Test Performance as a Function of Test Expectations," *Contemporary Educational Psychology, vol. 8,* January 1983, p. 55. See also Gordon Warren, "Essay Versus Multiple Choice Tests," *Journal of Research in Science Teaching*, vol. 16, No. 6, January 1979, pp. 563-567.

[34] Mary A. Lundeberg and Paul W. Fox, "Do Laboratory Findings on Test Expectancy Generalize to Classroom Outcomes?" *Review of Educational Research, vol. 61, No.* 1, spring 1991, pp. 94-106; and Traub and MacRury, op. cit., footnote 29.

thousands of English teachers in the United States consider the money well spent. The political clout that a writing sample provides for teaching writing and for emphasizing writing across the curriculum has no monetary equivalent.[35]

Of 38 States that currently assess student writing skills, 36 use direct writing samples in which students are given one or more ''prompts' or questions requiring them to write in various formats. An additional nine States have plans to add a direct writing assessment. Many districts also use writing assessments (see figure 7-l). These tests are used for a variety of purposes: some are required to certify students for graduation or to identify students who need further instruction, while others are used for district accountability measures.

For example, in order to identify students who need extra help in writing instruction prior to graduation, all ninth graders in the Milwaukee, Wisconsin public schools write two pieces each spring-a business letter and an essay describing a solution to a problem in their own life. The assessment helps reveal strengths and weakness in writing instruction among the district's schools and teachers. It is a standardized procedure, with all students given the same set of instructions and a set time limit for completing both pieces. Scoring is done by the English teachers during a week in June. The training process and the discussions that follow the scoring are valued by the teachers as an important professional activity, guiding them to reflect on educational goals, standards, and the evaluation of writing. The central office staff finds this one of the best forms of staff development; by clarifying the standards and building a consensus among teachers, the writing program can be more cohesively delivered throughout the district.[36]

The testimony of practitioners like the Milwaukee teachers supports the positive effects of tests using writing samples on writing instruction. It also appears that the positive effects of direct writing assessments on instruction are enhanced when teachers do the scoring themselves. In 19 of the 36 States currently assessing writing with direct writing samples, teachers from the home State score the assessments .37

A recent survey of the teachers involved in the California Assessment Program's (CAP) direct assessment of student writing found that, as a result of the direct writing assessment, over 90 percent of them made changes in their own teaching-either the amount of writing assigned, variety of writing assigned, or other changes.[38] Most report that they believe the CAP writing assessment will increase teachers' expectations for students' writing achievement at their school and that the new assessment will strengthen their school's English curriculum. Finally, there was almost unanimous agreement with the position that: ". . . this testis a big improvement over multiple choice tests that really don't measure writing skills. ' '[39] (See also box 7-C.)

An informal survey of practitioners using direct writing samples found these effects: increased quality and quantity of classroom writing instruction, changed attitudes of administrators, increased in-service training focused on teaching writing, use of test results to help less able pupils get ''real help," and improvement in workload for English teachers.[40] However, some practitioners noted possible negative effects as well, including the increased pressure on good writing programs to narrow their focus to the test, tendencies of some teachers to teach formulas for passing, and fears that the study of literature may be neglected due to intense focus on composition.

Because essays and direct writing assessments have been used in large-scale testing programs, they provide a rich source of information and experience

---

[35]Gertrude Conlan, " 'Objective' Measures of Writing Ability," *Writing Assessment: Issues and Strategies,* Karen L. Greenberg, Harvey S. Wiener, and Richard A. Donovan (cd...) (New York, NY: Longman, 1986), pp. 110111, emphasis added.

[36]Doug A. Archbald and Fred M. Newmann, *Beyond Standardized* Testing (Reston, VA: National Association of Secondary School Principals, 1988).

[37]The 19 States in which teachers participate as scorers are: Arkansas (voluntary), California, Connecticut, Georgia, Hawaii, Idaho, Indiana (voluntary), Maine, Maryland, Massachusetts, Minnesota, Missouri, Nevada, New York, Oregon, Pennsylvania, Rhode Island, Utah (voluntary), and West Virginia. In two-thirds of these States, teachers are trained by State assessment personnel. In the other one-third, they are trained by the contractor.

[38]California Assessment Program, ''~pact of the CAP Writing Assessment on Instruction and Curriculum: A Preliminary S ummary of Results of a Statewide Study by the National Center for the Study of Writing, ' draft report, n.d. The study sampled 600 teachers at California's 1,500 junior or middle schools in May 1988, just after the second statewide administration of the California Assessment program's grade eight writing test.

[39]Ibid.

@Charles Suhor, ''Objective Tests and Writing Samples: How Do They Affect Instruction in Composition?' *Phi Delta Kappan*, vol. 66, No. 9, May 1985, pp. 635-639.

for new attempts at performance assessment. Many practical issues, such as scoring and cost, are often raised as barriers to the large-scale implementation of performance assessment. The lessons drawn from the history of essays and direct writing assessments are illustrative--both for their demonstrations of feasibility and promise as well as their illumination of issues that will require further attention and care. These issues are discussed further at the end of this chapter.

Interviews and Direct Observations

Oral examin ations were the earliest form of performance assessment. The example best known among scholars is the oral defense of the dissertation at the Master's and Ph.D. levels. There are many varieties and uses of oral examinations at all school levels. University entrance examinations in a few countries are still conducted through oral examinations. Foreign language examinations often contain a portion assessing oral fluency. Other related methods allow teachers or other evaluators to observe children performing desired tasks, such as reading aloud.

The systematic evaluation of speaking skills has been incorporated into the College Outcome Measures Program (COMP) for the American College Testing Program (ACT). This test was designed to help postsecondary institutions assess general education outcomes. For the speaking skills portion of the assessment, students are given three topics and told to prepare a 3-minute speech on each. At an appointed time they report to a test site where they tape record each speech, using only a note card as a speaking aid. At some later time, trained judges listen to the tapes and score each speech on attributes related to both content and delivery.

Methods that use interviews and direct observations are particularly appropriate for use with young children. Young children have not yet mastered the symbolic skills involved in communicating through reading and writing; thus most paper-and-pencil-type tests are inappropriate because they cannot accurately represent what young children have learned. The best window into learning for the very young may come from observing them directly, listening to them talk, asking them to perform tasks they have been taught, and collecting samples of their work. This approach uses adults' observations to record and evaluate children's progress in km-

Photo credit: Educational Testing Service

Paper-and-pencil tests are often inappropriate for young children. This teacher, in South Brunswick, New Jersey, keeps a portfolio of her observations as she records each child's developing literacy skills.

guage acquisition, emphasizing growth over time rather than single-point testing.

Several States (i.e., Georgia, North Carolina, and Missouri) have developed statewide early-childhood assessments designed to complement developmentally appropriate instruction for young children. Most of these developmentally appropriate assessments are based on an English model, the Primary Language Record (PLR) developed at the Center for Language in Primary Education in London. The PLR is a systematic method of organizing the observations teachers routinely make. It consists of two parts, a continuous working record and a summary form, completed several times a year. The working record includes observations of the child's literacy behavior, such as **"running** records' of reading aloud, and writing samples, as well as a list of books the child can read either in

English or the language spoken at home. The summary record includes an interview with the parents about what the child likes to read and do at home and an interview with the child about his or her interests. The interviews take place at the beginning and end of each school year. The summary record goes with the child to the next grade, throughout primary school. The South Brunswick (New Jersey) schools have recently incorporated this approach into a teacher portfolio for assessing each student's learning in kindergarten through second grade (see box 7-D).

One assessment technique, used in South Brunswick as well as many other schools, is known as ''reading miscue analysis. ' The teacher sits with an individual student, listens to him read aloud, and systematically records the errors he makes while reading. From this analysis, which requires training, teachers can determine what strategies each child uses while reading. This can be a very useful assessment technique for all children, and especially in programs focused on improving reading skills in disadvantaged children.

The Georgia Department of Education has recently developed a new kindergarten assessment program (see box 7-E). One important component of this assessment is repeated and systematic observations of each child by the kindergarten teacher in many skill areas throughout the year. In addition, each kindergarten teacher receives a kit containing a number of structured activities that resemble classroom tasks. A teacher spends individual time with each student conducting these activities, which assess the child's skills in a number of areas. For example, one of the identified skills in the logical-mathematical area is the child's ability to recognize and extend patterns. The teacher presents the child with a task consisting of small cut-out dinosaurs in a variety of colors. Following a standardized set of instructions, the teacher places the dinosaurs in a sequenced pattern and asks the child to add to the sequence. Several different patterns are presented so that the teacher can assess whether the child has mastered this skill. If the child does not successfully complete the task, the teacher will know to work on related skills in the classroom; later in the year the teacher can use another task in the kit, this time using cut-out trucks or flowers, to reassess the child's skill in understanding patterns. Through this process, in which the teacher works directly with the child in a structured situation, the teacher is able to obtain valuable diagnostic information to adjust instruction for the individual child.

Exhibitions

Exhibitions are designed as inclusive, comprehensive means for students to demonstrate competence. They often involve production of comprehensive products, presentations, or performances before the public. They usually require a broad range of competencies and student initiative in design and implementation. The term has become popularized as a central assessment feature in the Coalition of Essential Schools (CES), a loose confederation of over 100 schools (generally middle and high schools) that share a set of principles reflecting a philosophy of learning and school reform that emphasizes student-centered learning and rigorous performance standards.

The term exhibition has two meanings as used in the Essential Schools. The most specific is the ''senior exhibition, ' a comprehensive interdisciplinary activity each senior must complete in order to receive a diploma. In this regard they are similar to the "Rite of Passage Experience" initiated by the Walden III Senior High School in Racine, Wisconsin. In order to graduate from Walden III, all seniors must demonstrate mastery in 15 areas of knowledge and competence by completing a portfolio, project, and 15 presentations before a committee consisting of staff members, a student, and an adult from the community .41

The CES senior exhibitions mirror some of these requirements, and typically fall into two main categories: the recital mode, which is a public performance or series of performances; and the "comprehensive portfolio" or ''exhibition portfolio,' a detailed series of activities, projects, or demonstrations over the school year that are cumulatively assembled and provide an aggregate picture of a student's grasp of the central skills and knowledge of the school's program.

There is also a general use of the term "exhibition" to mean a more discrete performance assessment when the student must demonstrate that he or she understands a rich core of subject matter and can apply this knowledge in a resourceful, persuasive,

41 Archbald and Newmann, op. cit., footnote 36, p. 23.

Box 7-D—South Brunswick Teacher Portfolios: Records of Progress in Early Childhood Learning[1]

How do you know if young children are developing critical language skills (reading, writing, and speaking) if you do not give them tests? This is the predicament facing many schools as educators become increasingly disenchanted with giving standardized paper-and-pencil tests to young children. When the South Brunswick New Jersey schools adopted anew, more developmentally appropriate curriculum it became necessary to develop anew method of assessment consistent with this teaching approach. Teachers worked with district personnel to create a teacher portfolio that drew on several models, including the Primary Learning Record used in England and Wales. Teachers piloted the portfolios over the 1989-90 school year, and revised them in the summer of 1990 for use the following school year.

The purpose of the portfolio is to focus on language acquisition in young students, grades K through 2. Teachers view the portfolio as a tool to promote instruction. It gives them a picture of the 1earning strategies of each child, which can be the basis of developing activities that will stress students' strengths while providing practice and help with weaknesses.

Each portfolio consists of 10 parts, plus one optional part:

- Self Portrait-The child is asked to "draw a picture of yourself" at the beginning and the end of the school year. The portraits are generally placed on the front and back covers of a manila folder.
- Interview--This maybe conducted several times during the year and includes the child's answers to such questions as: What is your favorite thing to do at home? Do you watch TV? Sesame Street? Do you have books at home? What is your favorite book? Do other people at home like to read? What do they read? Does someone read to you at home?
- Parent questionnaire--Parents complete this before their first conference with the teacher. It includes questions about the child's reading interests as well as any concerns the parent has about the child's language or reading development.
- Concepts about print test-This check list measures the child's understanding of significant concepts about printed language, such as the front of the book, that print (not the picture) tells the story, what a letter is, what a word is, where a word begins and ends, and big and little letters. This is a nationally normed test and is also used to identify children in need of compensatory education.
- Word awareness writing activity-This records the level at which children begin to comprehend the rules of forming words in their writing. Progress is recorded along a five-stage scale: precommunicative (random spelling or scribbling); semiphonetic (some sounds represented by letters, e.g., the word "feet" might be rendered as "ft"); phonetic (letters used appropriately for sounds, e.g., "fet"'); transitional (some awareness of spelling patterns, e.g., "fete"); or mostly comet (10 out of 13 words correctly spelled).
- Reading sample-This is taken three or more times a year. The teacher may use a "running record" or "miscue analysis.' The running record is used with emergent readers, children who mimic the act of reading but do not yet know how to read. It records what a young child is thinking and doing while "reading."
- Writing sample--This is a sample of the student's free writing, "translated" by the student for the teacher if invented spelling and syntax make it difficult to read easily.
- Student observation forms (optional).
- Story retelling form.
- Diagnostic form.
- Class record-This class profile helps the teacher identify those children who may need extra attention in certain areas. It is a one-page matrix with yes-no answers to the following five questions: Does the child pay attention in large and small groups? Interact in groups? Retell a story? Choose to read? Write willingly? This is the only element of the portfolio not a part of the child's individual record.

Because of Federal requirements for determiningg eligibility for compensatory education, the South Brunswick schools also use norm-referenced, multiple-choice tests. However, teachers report that these tests are not useful because they do not assess development in the instructional approach adopted by the South Brunswick schools.

---

[1]Much of the material in this box comes from Ruth Mitchell and Amy Stempel, Council for Basic Education, "Six Case Studies of Performance Assessment," OTA contractor_ March 1991.

> The tests go from part to whole, and our programs go from whole to part. Those tests are basically for basals [i.e., reading textbooks], and to assess kids that have learned a whole language by basals (when the South Brunswick students used children's literature as texts)--it makes no sense at all.[2]

The portfolios provide a different approach to the question of student retention. While a student may have been held in grade before because of low test scores, research has suggested that having a child repeat a year in grade may in fact cause more harm than good.[3] In South Brunswick, when there is a question about retention or special education labeling in the early grades, the portfolio record is consulted to see if the child has made progress. If progress can be shown, then the student is promoted on the assumption that every child develops at his or her own rate and can be monitored closely until he or she reaches the third grade. If no progress is apparent at that point, the child is promoted but is identified for compensatory education.

One of the purposes of the portfolio is to help the teacher provide a clearer picture of student progress to parents than is possible from standardized test scores. Yet a tension remains between the old and the new. The numbers that are derived from norm-referenced, multiple-choice tests are familiar and understandable. The new developmentally appropriate methods of teaching and testing do not have the perceived rigor or precision of the old tests. Some parents assume that only norm-referenced tests can be objective, and worry about subjectivity in recording progress on the portfolios. Some want traditional test scores that assure that their children are learning what everyone else in the country is learning--or can be measured against children in other communities. Until this tension is resolved, full acceptance of a portfolio system maybe slow. As one teacher said:

> The next step is to educate the parents. We need workshops for parents. That is the big issue, after we get all the teachers settled in using the portfolio. This is basically not going to be acceptable until these children get older and everyone can see that we're graduating literate kids and that's not going to be until many, many years from now.[4]

Standardization of the portfolio assessment was not an issue for the teachers, because of its primary role as an instructional information tool. Since the teachers were involved in the initial design and remain involved in modifications, and as they have attended workshops on its use, there is implicit standardization. Although the South Brunswick portfolio is primarily meant as a feedback mechanism to improve instruction, it also is being used as an accountability instrument. The Educational Testing Service (ETS), working with the teachers, has produced a numerical literacy scale based on the portfolio. The scale provides a means of aggregating data from the portfolios. Central office staff, working with a consultant from ETS, examined literacy scales and will rank children's literacy as evidenced by the portfolio on these scales. Teachers in one school rank the portfolios based on these scales, in order to evaluate how well the system communicates standards. The "South Brunswick-Educational Testing Service scale" for evaluating children's progress in literacy is now being used in all district schools. The literacy scales replace the first grade standardized reading test. The existence of aggregatable data will clearly enhance the scoring and the overall value of the portfolio in the South Brunswick public schools.

There are additional approaches to standardizing the portfolio. Some of the contents, such as the Concepts of Print test and the Word Awareness Writing Activity, can be scored using a key. Running record and miscue analysis can also be scored consistently. Those aspects that cannot be scored using a key-e. g., the writing sample--can be graded by a group of teachers developing a rubric from each set of papers. These could also be standardized by exchanging a sample of portfolios among teachers, so that each reads about 10 percent from each class and discusses common standards. This is the method used by the New York State Department of Education to ensure standardization of the results of their grade four science manipulative skills test. It is also used in several European school systems.

The issue of bias has not been raised, since the teachers record each student's growth against himself or herself, not in comparison with other students in the class or school. However, this issue will be more prominent if achievement levels are set and there are differing success rates in meeting these standards, or if the portfolio is used for school accountability or for student selection, two goals not currently planned.

---

[2]Willa Spicer, director of Instruction, South Brunswick Public Schools, New Jersey, personal communication December 1991.

[3]Lorrie Shepard and Mary Lee Smith, *Flunking Grades: Research and Policies on Retention* (London, England: The Falmer Press, 1989).

[4]Mitchell and Stempel, op. cit., footnote 1, p.17.

Box 7-E—Testing in Early Childhood: Georgia's Kindergarten Assessment Program

In recent years, many educators and policymakers have been reducing or eliminating the use of standardized paper-and-pencil tests in the early grades. Many of these tests were being used to make decisions about kindergarten retention and whether children were ready to begin first grade. The issue of retention in the early grades, as well as the role of tests in making such decisions, is receiving increasing scrutiny and many policies are changing. The Texas State Board of Education recently barred the retention of any pupils in prekindergarten and kindergarten.[1] The legislatures in both Mississippi and North Carolina have eliminated State-mandated testing in the early grades.[2] At least two States, Kentucky, and Florida, are encouraging ungraded primaries (K-3) which loosen the rigid boundaries between the early grades and allow children to move according to individual progress.

In a policy running somewhat counter to these trends, the Georgia Legislature in 1985 mandated that all 6-year-olds must pass a test in order to enter first grade. During the first 2 years of this policy, a standardized paper-and-pencil test was used. However, the use of such a test quickly brought to public attention concerns about this approach to readiness assessment, including:

1. the appropriateness of a paper-and-pencil test for children who are five to six years of age.
2. the concern that a focus on tests narrows the curriculum . . .
3. the need to consider not just the child's cognitive skills, but the development of social, emotional, and physical capacities as well.
4. the need to consider the teacher's observations of the child throughout the course of the school year.[3]

In response to these concerns the Georgia Department of Education embarked on a large project to design a developmentally appropriate model of assessment. The Georgia Kindergarten Assessment Program (GKAP), piloted during 1989-90, uses two methods of assessment--observations by kindergarten teachers and individually administered standardized tasks that resemble classroom activities. GKAP assesses a child's capabilities in five areas: communicative, logical-mathematical, physical, personal, and social. This assessment program is designed to help teachers make multiple, repeated, and systematic observations about each child's progress during the year. Behavioral observations in all five areas are made in three time periods throughout the year. In addition, a set of structured activities have been designed to assess each child's communicative and logical-mathematical capabilities. The teacher conducts each of these activities individually with a child If a child cannot successfully complete the task, teachers can plan activities to help the child work on that skill in the classroom; a second activity, assessing that same skill, can be given by the teacher later in the year. These tasks involve toys, manipulative, and colorful pictures.

Each kindergarten teacher in Georgia receives a GKAP kit that contains manuals for administration, manipulative, and reporting forms. Training and practice are required prior to the use of GKAP. A self-contained video training program developed for this purpose has been provided to each school.

The education department anticipates that this assessment program will serve a number of important functions:

> A significant use of GKAP results is to provide instructionally relevant diagnostic information for kindergarten teachers. In the process of collecting GKAP information, teachers gain insights regarding their students' developmental status and subsequent modifications which may be needed in their instructional programs. In addition, when forwarded, this information will also be useful to the child's teacher at the beginning of the first grade year. Another use of GKAP results is communication with parents about their child's progress throughout the kindergarten year.
>
> The results of the GKAP are also to serve, along with other information about the child, as a factor in the decision regarding whether to promote the child to the first grade. GKAP results, by themselves, should not be used as the sole criterion for promotion/retention (placement) decisions.[4]

---

[1] Deborah L. Cohen, "Texas Board Votes to Forbid Retention Before the 1st Grade," *Education Week*, vol. 90, No. 1, Aug. 1, 1990.

[2] Mississippi stopped** kindergarten children and North Carolina banned testing of first and second graders. See Adria Steinberg, "Kindergarten: Producing Early Failure?" *Principal*, vol. 69, May 1990, pp. 6-9.

[3] Werner Rogers and Joy E. Blount, "Georgia's First Grade Readiness Assessment: The Historical Perspective%" paper presented at the annual meeting of the American Educational Research Association Boston, MA, April 1990, p. 3.

[4] Susan P. Tyson and Joy E. Blount, "The Georgia Kindergarten Assessment Program: A State's Emphasis on a Developmentally Appropriate Assessment," paper presented at the American Educational Research Association Boston, MA, April 1990, p. 7.

and imaginative way. It is a creative and difficult concept to put into place, however, and requires that the teacher create assignments that take students beyond the surface of a subject. For example, one history teacher suggested: "Under the old system, the question would be 'Who was the King of France in 800?' Today, it is 'How is Charlemagne important to your life?' "[42] While the exhibition format could be an essay or research paper, it might also call for a Socratic dialog between student and teacher, an oral interview, debate, group project, dramatic presentation, or combination of multiple elements, partly in preparation for the more comprehensive senior exhibitions. Clearly, developing and evaluating successful exhibitions can be as big a challenge to the teachers as it can be for the students to perform well on them.

Exhibitions can also be competitions, some at the individual level, like the Westinghouse Science Talent Search, or in groups, like the Odyssey of the Mind, a national competition requiring groups of students to solve problems crossing academic disciplines. Group competitions add group cooperation skills to the mix of desirable outcomes.

One interesting group competition is the Center for Civic Education's" We the People . . ." program on Congress and the Constitution. It is a national program, sponsored by the Commission on the Bicentennial of the U.S. Constitution and funded by Congress. Students in participating schools study a specially developed curriculum and compete with teams from around the country. In the competition they serve as panels of "experts" testifying before a mock congressional committee. The curriculum can be used as a supplement to American history or civics classes and has materials that are appropriate for three levels (upper elementary, middle school, and high school). The text centers on the history and principles of the U.S. Constitution. When students have completed the curriculum the entire class is divided into groups, each responsible for one unit of the curriculum. Each group presents statements and answers questions on its unit before a panel of community representatives who act as the mock congressional committee members. Winning teams from each school compete at district, State, and finally a national-level competition. Training for judges at each level is conducted through videotapes and training sessions in which the judges evaluate each group on a scale of 1 to 10, on the criteria shown in figure 7-6.

Experiments

Science educators who suggest that students can best understand science by doing science have promoted hands-on science all across the science curriculum. Similarly, they maintain that students' understanding of science can best be measured by how they do science--the process of planning, conducting, and writing up experiments. Thus, science educators are seeking ways to assess and measure hands-on science. A number of States, including New York, California, and Connecticut, have pioneering efforts under way to conduct large-scale hands-on assessments in science.

In 1986, NAEP conducted a pilot project to examine the feasibility of conducting innovative hands-on assessments in mathematics and science. Working closely with the staff of Great Britain's Assessment of Performance Unit, 30 pilot tasks using group activities, work station activities, and complete experiments were field tested. School administrators, teachers, and students were enthusiastic and encouraging about these efforts. As part of the pilot project, NAEP has made available detailed descriptions of these 30 tasks so that other educators can adapt the ideas.[43] A sample experiment used with third graders and scoring criteria are pictured in figure 7-7.

New York Elementary Science Program Evaluation Test

In 1989, the New York State Department of Education, building on the NAEP tasks, included five hands-on manipulative skills tasks as an important component of their Elementary Science Program Evaluation Test (ESPET). Used with fourth graders, the test also included a content-oriented, paper-and-pencil component. It was the intent of the

[42] James Charleson, Hope H@ School, Providence, RI, quoted in Thomas Tech and Matthew Cooper, ''hSSOm From the Trenches, " *U.S. News & World Report,* vol. 108, No. 8, Feb. 26, 1990, p. 54.

[43] See Educational Testing Service, *Learning* by *Doing: A Manual for Teaching and Assessing Higher Order Thinking in Science and Mathematics* (Princeton, NJ: May 1987); or the full-report, Fran Blumberg, Marion Epstein, Walter MacDonald, and Ina Mullis, *A Pilot Study of Higher Order Thinking Skills: Assessment Techniques in Science and Mathematics, Final Report* (Princeton, NJ: Educational Testing Service, November 1986).

Figure 7-6-Scoring Sheet for the "We the People" Competition

Student teams act as witnesses before a 'Congressional Committee" and answer questions on the U.S. Constitution (history, law, and current applications). Each group is scored on a scale of 1-10 on the criteria listed below.

1-2 .poor 3-4 .fair 5-6 .average 7-8 .above average 9-10 ▪ excellent

	Score	Notes
1. Understanding: To what extent did participants demonstrate a clear understanding of the basic issues involved in the questions?		
2. Constitutional Application: To what extent did participants appropriately apply knowledge of constitutional history and principles?		
3. Reasoning: To what extent did participants support positions with sound reasoning?		
4. Supporting Evidence: To what extent did participants support positions with historical or contemporary evidence, examples, and/or illustrations?		
5. Responsiveness: To what extent did participants' answers address the questions asked?		
6. Participation: To what extent did most group members contribute to the group's presentation?		
Group total		

Judge: _____ Date: _____

Congressional District: _____ Tie breaker* _____

.P l ease award u p to 100 points for this group's overall performanc e. (Bonus points will only be used in the event of a tie.)

SOURCE: Center for Civic Education, Calabasas, CA.

test designers to align classroom practices with the **State** objectives reflected in the syllabus.[44]

The manipulative test consists of five tasks, and each student is given 7 minutes to work on each of the tasks. At the end of each timed segment, the teacher organizes a swift exchange of desks, or stations, moving the front row children to the back of the column and the others each moving up one desk, somewhat like a volleyball rotation. Test **stations are separated by cardboard dividers and are arranged so that** adjacent stations do not have the **same apparatus. Four classes of about 25 children** each can be **tested comfortably in a school day. The skills assessed by the five stations include measurement** (of volume, length, mass, and **temperature),** prediction from observations, classification, hypothesis formation, and observation.

The examinationswere scored by their teachers, but student scores were not reported above the school level. School scores were reported in terms of the items on which students had difficulty. The ESPET is currently being evaluated for use in other grades.

Connecticut Common Core Science and Mathematics Assessments

Connecticut has been a leader in the development of a set of mathematics and science assessments that

[44] Sally Bauer, Sandra Mathison, Eileen Merriam, and Kathleen Toms, ' 'Controlling Curricular Change Through State-Mandated Testing: Teacher's Views and Perceptions," paper presented at the annual meeting of the American Educational Research Association, Boston, MA, Apr. 17, 1990, p. 7.

call on group skills and performance activities.[45] Under a 45-month grant from the National Science Foundation, Connecticut has assembled teams of high school science and mathematics teachers working jointly on Connecticut Multi-State Performance Assessment Collaborative Teams (COMPACT). CoMPACT is made up of seven State Departments of Education (Connecticut, Michigan, Minnesota, New York, Texas, Vermont, and Wisconsin), CES, The Urban District Leadership Consortium of the American Federation of Teachers, and Project Re:Learning.

The COMPACT group has designed and developed 50 performance assessment tasks, 31 across 8 areas of high school science (biology, chemistry, Earth science, and physics) and 19 in mathematics (general or applied mathematics, algebra, geometry, and advanced mathematics). After pulling together the experiences of COMPACT teachers trying out these tasks, Connecticut will convene committees of expert judges to establish ''marker papers' and common scoring standards. These scoring standards will be used during 1991-92 on the frost administration of the Connecticut Common Core of Learning Assessments in high school science and mathematics across the State. A key element of the entire endeavor will be the assessment of student attitudes toward science and mathematics, and the demonstration of teamwork and interpersonal skills in these real-life testing contexts.

Each task has three parts that require individual work at the beginning and end, and group work in the middle (see figure 7-8). First, each student is presented with the task and asked to formulate a hunch, an estimate of the solution, and a preliminary design for a study. This portion of the task has several goals-it focuses the student's preliminary thinking, becomes a springboard for student group discussion, gives the teacher a feel for where the students are in their thinking, and serves as a record that the student can revisit throughout the assessment.

The middle section involves the longest phase. Here students plan and work together to produce a group product; teamwork is emphasized throughout. Evidence of deepening understanding is recorded through a variety of assessment tools such as written checklists, journals, logs, or portfolios. Oral or visual records such as videotapes of group discussions and oral presentations are also maintained. Teachers can rate individual performance on a subset of objectives in the group task. The ability to infer levels of individual contribution on collective work is one of the largest assessment challenges.

The third part of the task consists of individual performance on a related task. These tasks consist of similar activities that attempt to assess some of the same content and processes as the group task. The transfer task provides each student with an opportunity to synthesize and integrate the learning that occurred in the group experience and apply it in a new context. It also provides teachers, parents, and policymakers with a summative view of what each student knows and can do at the end of a rich set of learning and assessment opportunities.

Several evaluations of the project have been completed to date. Teacher perceptions are quite positive. Through the participation of the Urban District's Leadership Consortium, students in 16 large urban school systems tried out the performance tasks during the 1990-91 school year, demonstrating the feasibility of this type of assessment in schools with large populations of African-American and Hispanic students.[46]

Portfolios

Portfolios are typically files or folders that contain a variety of information documenting a student's experiences and accomplishments. They furnish a broad portrait of individual performance, collected over time. The components can vary and can offer multiple indicators of growth as well as cumulative achievement. As students assemble their own portfolios, they evaluate their own work, a key feature in performance assessment. Proponents suggest that this process also provides students a different understanding of testing, with the following positive effects:

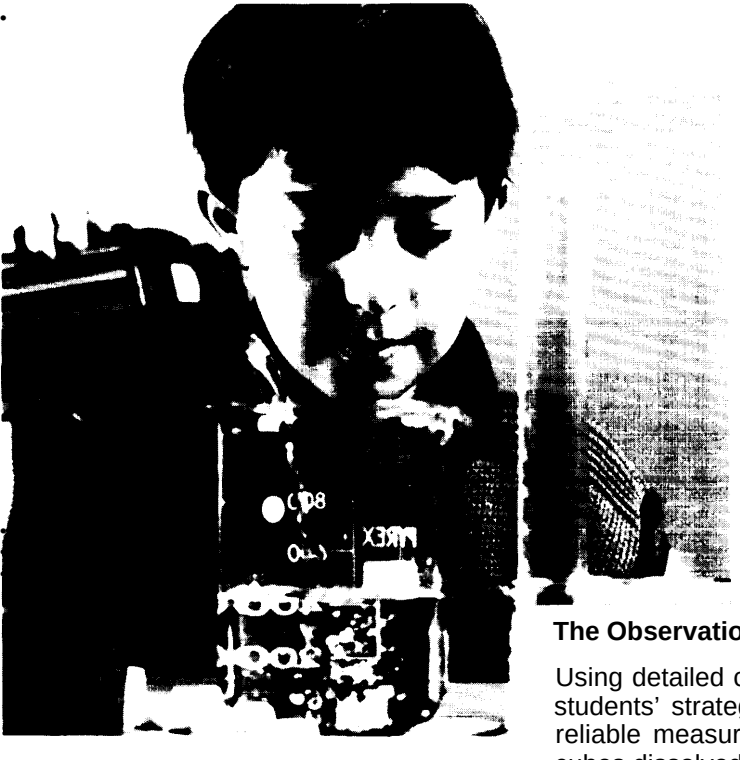
[45]See Pascal D. Forgione, Jr. and Joan Boykoff Baron, Connecticut State Department Of Education, ''Assessment of Student Performance in High School Science and Mathematics: The Connecticut Study,' paper presented at the Seminar on Student Assessment and Its Impact on School Curriculum, Washington DC, May 23, 1990.

[46]Joan Boykoff Baron, Connecticut Department of Education, personal communication, November 1991.

Figure 7-7—"Sugar Cubes": A NAEP Hands-On Science Experiment for 3rd Graders

The Experiment

Students are given laboratory equipment and asked to determine which type of sugar, granulated or cubed, dissolves faster **when placed in warm water that is stirred and not stirred, respectively. To complete this investigation, students** need to identify the variables to be manipulated, controlled, and measured. **They also** need to make reliable and accurate measurements, record their findings, and draw conclusions. Examples of written conclusions are presented on the next page.



The Observation

Using detailed checklists, NAEP administrators recorded students' strategies for determining-wit h accurate and reliable measurements-whether loose sugar or sugar cubes dissolved at a faster rate.

NAME: _____
CODE: _____
SCHOOL DISTRICT: _____

Sugar Cubes Behavioral Checklist

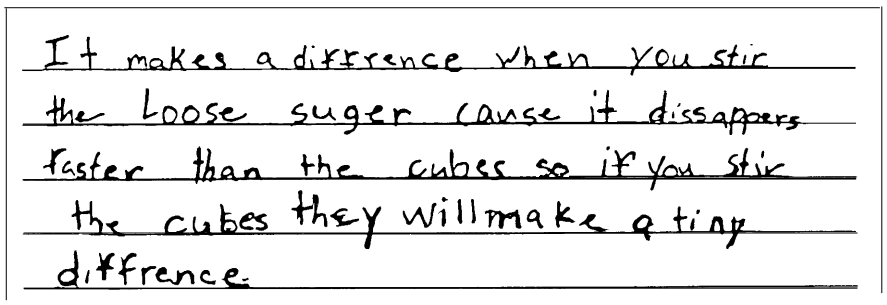
NOT STIRRING	1. Loose sugar tested	☐
	2. Cube sugar tested	u
SET-UP	3. Volume of water measured-by eye	❑
	4. by ruler	
	5. by cylinder	
	6. Volume used< 10 cc	
	7. Volume used >10 cc	
	8. Volume same for both types	
	9. Mass same for both types	❑
MEASUREMENT	10. No apparent measurement	
	11. Qualitative measurement	
	12. Clock used	☐
	13. within +-3 sees. of start point	☐
	14. within +-3 secs. of end point	1
	15. Timed-until all dissolved	1
	16. until partially dissolved	❑
	17. no dear end point	❑
	18. Fixed time--notes amount remaining	1
RESPONSE SHEET	• 19. Reports results consistent with evidence	❑
STIRRING	20. Stirring not tested-sugar type not controlled	1
	21. Loose sugar tested	❑
	22. Cube sugar tested	1
	23. Stirring tested-by counting number of stirs	c1
	• 24. by timing	1
	25. Stirring at regular intervals	☐
	26. Stirring rate-constant	☐
	• 27. ran&m	1
SET-UP	28. Volume of water measured-by eye	n
	• 29. by ruler	1
	• 30. by cylinder	1
	31. Volume used < 10 cc	❑
	• 32. Volume used >10 cc	1
	33. Volume same for both types	☐
	34. Mass same for both types	u
MEASUREMENT	35. No apparent measurement	☐
	36. Qualitative measurement	u
	37. clock used	1
	38. within +-3 specs. of start point	1
	39. within +-3 sees. of end point	1
	40 Tired — until all dissolved	☐
	olved	☐☐
	nount remaining	☐
	istent with evidence	1
	both trials	☐
	ck findings	☐☐
	or minimal)	☐

*48. Acknowledges that procedures' could be improved if experiment repeated-aware that certain variables

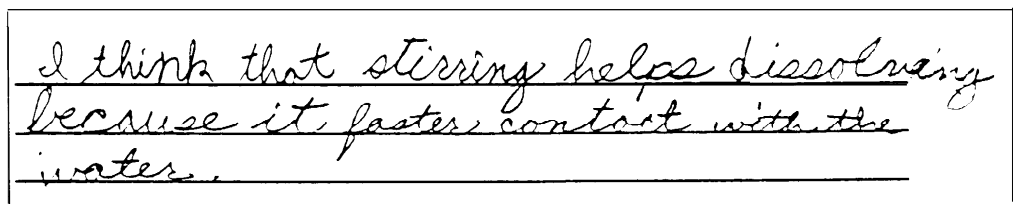
FIND OUT **IF STIRRING MAKES ANY DIFFERENCE IN** HOW FAST THE SUGAR CUBES AND LOOSE SUGAR DISSOLVE.

B) Use the space below to answer the question in the box.

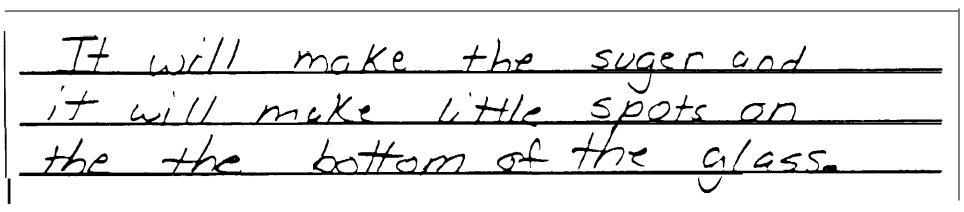
Score received



5 point answer



3 point answer



1 point answer

Scoring of Written Answers

5 points = response states that both types of sugar dissolve faster but loose sugar dissolves the fastest.
4 points = response states that the loose sugar dissolves faster than the cube and that stirring is the cause of it.
3 points = response states that stirring makes a difference only or how or why an effect upon the sugar is found only.
2 points = response states that one type of sugar dissolves faster than another only.
1 point = incorrect response.
0 points = no response.

KEY: NAEP = National Assessment of Educational Progress.

SOURCE: Educational Testing Service, *Learning by Doing: A Manual for Teaching and Assessing Higher Order Thinking in Science and Mathematics (Princeton,* NJ: May 1987); and Fran Blumberg, Marion Epstein, Walter MacDonald, and Ina Mullis, *A Pilot Study of Higher Order Thinking Skills: Assessment Techniques in Science and Mathematics, Final Report* (Princeton, NJ: Educational Testing Service November 1986)

Figure 7-8-Connecticut Science Performance Assessment Task: "Exploring the Maplecopter"

OVERVIEW: This task was designed for high school physics classes, and includes both individual and group work. Students study the motion of maple seeds anddesign experiments to explain their spinning fright patterns. Curriculum topics include laws of motion, aerodynamics, air resistance, and the use of models in explaining scientific phenomena. Equipment needed: maple seeds, paperboard, stopwatches, and scissors. The suggested length of time for the task is 3 to 5 class periods.

Part 1: Getting Started by Yourself

1. Throw a maple winged seed up in the air and watch it "float" down to the floor. Describe as many aspects of the motion oft he pod as you can. You may add diagrams if you wish.

2. One of the things you probably noticed is that the seed spins at it falls, like a little helicopter. Try to explain how and why the seed spins as it falls.

Part II: Group Work

The criteria that will be used to assess your work are found on the Objectives Rating Form - Group. Each member of your group will also fill out the Group Performance Rating Form.

1. Discuss the motion of the winged maple seed with the members of your group. Write a description of the motion, using the observations of the entire group. You may add diagrams if you wish.

2. Write down the variables that might affect the motion of the maple seed.

3. Design a series of experiments to test the effect of each of these variables. Carry out as many experiments as necessary in order to come up with a complete explanation for the spinning motion of the winged seed.

Using Models in Science

4. Sometimes using a simplified model (or a simulation) might help one to understand more complex phenomena. A paper helicopter, in this case, might serve as a simplified model of the seed.

   a. Construct a paper helicopter following the general instructions in figures 1 and 2.

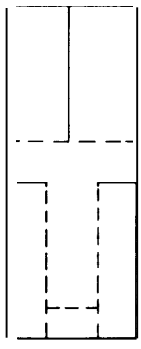


Figure 1

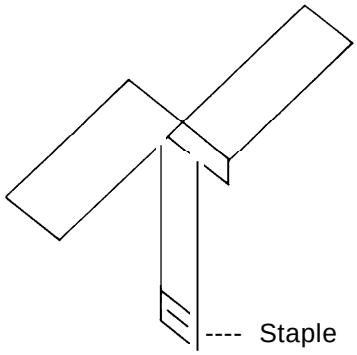


Figure 2

   b. Throw the paper helicopter in the air and observe its motion.
   c. Try changing various aspects of the paper helicopter to test the effect of the variables your group chose.
   d. Experiment with different types of paper helicopters until you feel t hat you have a complete understanding of how the variables you identified affect the motion.
   e. Summarize your results with the help of a chart or a graph.

5. Based on what you've learned from the paper helicopters, design and perform additional experiments with the maple seeds.

6. Describe your group's findings from all your experiments. Raw data should be presented in charts or graphs, as appropriate and summarized by a short written statement.

7. Now, after you have completed all the necessary experiments, try to explain again the motion of the maple seed. Try to include in your explanation t he effect of all the variables that you observed in your experiments. You may add diagrams if you wish.

8. in this activity you used simplified models to help explain a more complicated phenomenon. Describe the advantages and disadvantages of your paper helicopter as a model of a winged maple seed.

9. What are the biological advantage(s) of the structure of the maple seed? Explain fully.

Part ill: Finishing by Yourself

THE GRAND MAPLECOPTER COMPETITION

Your goal is to design a helicopter, from a 4" X 8" piece of paperboard, that will remain in the air for the longest time when dropped from the same height.

a. Design the "helicopter."
b. Write down factors related to your design.
c. Cut out the "helicopter."
d. Mark the helicopter with your name.
e. Good luck and have fun!

SOURCE: Connecticut State Department of Education, 1991.

GROUP PERFORMANCE RATING FORM

Student Name _____

Student ID # _____

	Almost always	Often	Some-times	Rarely
A. GROUP PARTICIPATION				
1. Participation in group discussion without prom				
2. Did his or her fair share of the work				
3. Tried to dominate the group - interrupted oth				
4. *Participated in the group's activities*				
B. STAYING ON THE TOPIC				
5. Paid attention, listened to what was being				
6. Made comments aimed at getting the gro				
7. Got off the topic or changed the subject				
8. *Stayed off the topic*				
C. **OFFERING USEFUL IDEAS**				
9. Gave ideas and suggestions that help				
10. Offered helpful criticism and comme				
11. Influenced the group's decisions ar				
12. *Offered useful ideas*				
D. CONSIDERATION				
13. Made positive, encouraging rem				
14. Gave recognition and credit to				
15. Made inconsiderate or hostile				
16. *Was considerate of others*				
E. **INVOLVING OTHERS**				
17. Got others involved by askir				
18. Tried to get the group worki				
19. Seriously considered the ic				
20. *Involved others*				
F. COMMUNICATING				
21. Spoke clearly, was easy to hea				
22. Expressed ideas clearly and effectively				
23. *Communicated clearly*				

Part II: OBJECTIVES RATING FORM (GROUP)

Title of the test _____

Student ID numbers (1) _____ (2) _____ Teacher ID# _____ (3) _____

The group should be able to:	E	G	N.I.	U
1. describe the flight of a maple seed pod based on observation				
2. describe the variables that might affect the motion of a maple pod				
3. design and perform experiments to help explain the motion of a maple pod				
4. design and perform experiments to help explain the motion of a paper helicopter				
5. design and perform new experiments based on transfer from the model				
6. describe findings from experiments in words, charts, and graphs				
7. explain the motion of a maple pod based on experimental data				
8. describe the advantages and disadvantages of paper helicopters as models of maple pods				
9. explain the biological advantage of the design of the maple pod				
10. communicate effectively through written means				
11. collaborate effectively				

E=excellent G=good N.I.=needs improvement U=unacceptable

A Sample of Other Science Performance Tasks Under Development

BOILING POINT LABORATORY: Students are asked to design and carry out a controlled experiment to determine the mixture of antifreeze and water that has the highest boiling point and is thus the most effective in keeping cars running smoothly in extreme temperatures.

OUTCROP ANALYSIS: Students are given a variety of information, including videotapes, pictures, and rock samples, from a site in Connecticut and are asked to determine if it is a good site on which to build a nuclear power facility. Students may be asked to investigate other factors, such as population, waste disposal, weather, politics, etc. in determining if it is a good site.

WEATHER PREDICTION: Students are asked to predict the weather based on their knowledge of meteorology, data they collect, and observations that they are able to make. Students may be asked to make simple weather instruments or create a weather forecasting segment as it would appear on a television newscast.

- testing becomes a personal responsibility;
- students realize that they need to demonstrate a full range of knowledge and accomplishments, rather than a one-shot performance;
- they begin to learn that frost draft work is never good enough; and
- they appreciate that development is as important as achievement.[47]

A small but growing number of States have embraced portfolios as an educational assessment tool. As of 1991, five States (Alaska, California, North Carolina, Rhode Island, and Vermont) had implemented portfolios as a mandatory, voluntary, or experimental component of the statewide educational assessment program. Four additional States (Delaware, Georgia, South Carolina, and Texas) are considering implementing portfolios for this purpose. At the State level, portfolios have been implemented mostly in mathematics and writing at grade levels ranging from 1st to 12th but concentrated in the early grades.[48] The Vermont experience with portfolios is noteworthy (see box 7-F). Michigan's portfolio project, begun on a pilot basis in 22 districts during 1990-91, focuses on the skills that high school graduates are expected to have in order to be productive workers. As described in box 7-G, this use of portfolios aims at providing both students and prospective employers with information on workplace skill competencies.

Research on effectiveness of portfolios is being assembled by the project Arts PROPEL, a 5-year cooperative effort involving artists, researchers from Harvard University's Project Zero, the Educational Testing Service (ETS), and teachers, students, and administrators from the Pittsburgh and Boston public school systems. Supported by a grant from the Rockefeller Foundation, Arts PROPEL seeks to create a closer link between instruction and assessment in three areas of the middle and secondary school curriculum: visual arts, music, and imaginative writing.[49] The primary purpose of the assessment is not for selection, prediction, or as an institutional measure of achievement. Instead, it is focused on understanding individual student learning as a way of improving classroom instruction. The goal is to create assessments that provide a learning profile of the individual on as many dimensions as possible, as well as showing student change over time.[50] The two sources of assessment are portfolios and what is called the ''domain project," an instructional sequence that focuses on central aspects of a domain and provides opportunities for multiple observations of the student. Domain projects function as self-contained instructional units central to the arts curriculum, and are graded by the classroom teacher.

The portfolio is the central defining element in Arts PROPEL. It is intended to be a complete process-tracking record of each student's attempts to realize a work of art, music, or writing. It also serves as a basis for students' reflection about their work, a means for them to identify what they value in selecting pieces for inclusion, and a vehicle for conversations about that work with teachers. A typical portfolio might contain initial sketches, drafts, or audiotapes; self criticisms and those of teachers and other students; successive drafts and reflections; and examples of works of others that have influenced the student. A final evaluation by the student and others is included, along with plans for successive work. Researchers and school district personnel are attempting to find methods of assessing artistic growth and of conveying this information effectively-through scores or other summary indicators—to administrators, college admissions officers, and others.

Like writing assessments, the use of portfolios is not new. For 19 years it has been the major component of the Advanced Placement (AP) studio art examination, administered by ETS[51] (see box 7-H).

---

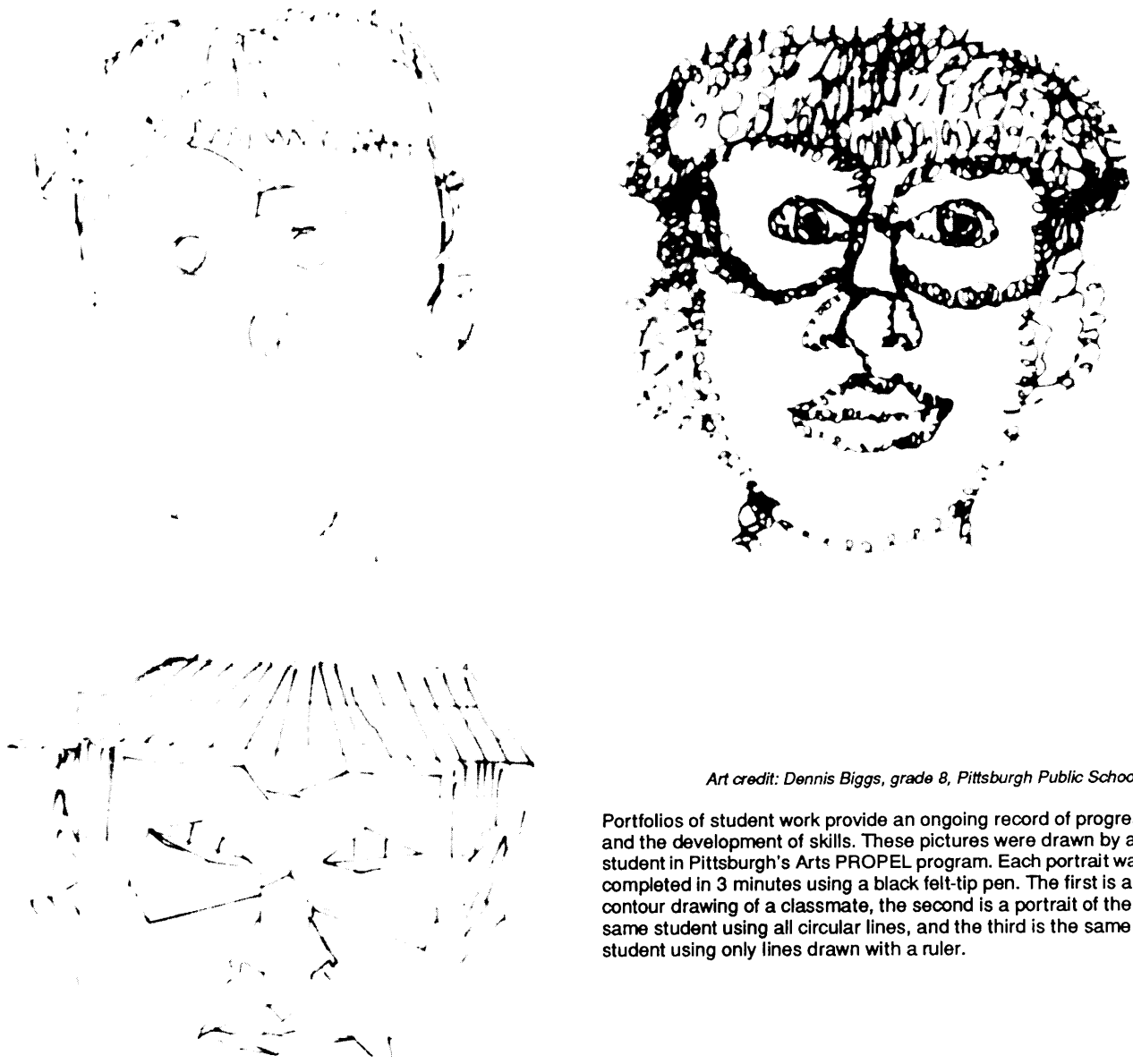
[47] From Dennie Palmer Wolf, ''Portfolio Assessment: Sampling Student Work." *Educational Leadership,* vol. 46, No. 7, April 1989, pp. 35-36.

[48] OTA data, 1991.

[49] Roberta Camp, "Presentation on Arts PROPEL Portfolio Explorations," paper presented at the Educational Testing Seminar on Alternatives to Multiple-Choice Assessment, Washington, DC, Mar. 30, 1990, p. 1.

[50] Drew H. Gitomer, "Assessing Artistic Learning Using Domain Projects, " paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 1988, p. 4.

[51] Mitchell and Stempel, op. cit., footnote 2.



Art credit: Dennis Biggs, grade 8, Pittsburgh Public Schools

Portfolios of student work provide an ongoing record of progress and the development of skills. These pictures were drawn by a student in Pittsburgh's Arts PROPEL program. Each portrait was completed in 3 minutes using a black felt-tip pen. The first is a contour drawing of a classmate, the second is a portrait of the same student using all circular lines, and the third is the same student using only lines drawn with a ruler.

Common Characteristics of Performance Assessment

Although there is great variety in the kinds of measures that fall under the umbrella of performance assessment, certain common characteristics distinguish their use and implementation in school systerns.

Performance tests require *student-constructed* responses as opposed to student-selected responses. While it is not certain that these two responses involve different cognitive processes, creating a

297-933 0 - 92 -- 16 : QL 3

Box 7-F—"This is My Best": Vermont's Portfolio Assessment Project

Prior to 1990, Vermont was one of the few States with no mandated statewide testing program. Districts could conduct standardized norm-referenced testing for their own purposes. However, change came to Vermont when the legislature approved funds for a statewide assessment program to be integrated with classroom instruction. The first piece of the plan, piloted in the 199091 school year in one-quarter of the schools across the State, focused on writing and mathematics in grades four and eight. Eventually all the major academic disciplines will be covered. Each assessment has three parts: a uniform test, "best pieces" exemplifying the student's highest achievement in the judgment of the student and teacher, and a portfolio showing development throughout the year.

The mathematics assessment includes a standardized test that contains multiple-choice, open-ended, and longer computational problems. Each student is also responsible for assembling a mathematics portfolio, a collection of some 10 to 20 entries of problems and projects completed Five to seven of these are pieces the student and teachers have chosen as best pieces, accompanied by a letter the student writes to the evaluator, explaining why these were selected. All this conferring, questioning, reviewing, and writing about mathematics is aimed at better understanding and communication about mathematical reasoning, logic, and problem solving. The mathematics portfolios are designed to foster an attitude of responsibility for learning on the part of the student, reveal the student's feelings about mathematics, and provide a means of showing growth in areas not well suited to standardized tests[1] (see figure 7-F1).

The writing assessment is made up of a uniform writing prompt and an interdisciplinary writing portfolio.[2] The writing assessment is similar to that used in other States, with students given a uniform prompt and 90 minutes to respond. The students are encouraged to think through ideas first and write rough drafts, using dictionaries and thesauruses provided in the testing room, and then produce a finished product. The prompt used for the 1990-91 pilot was:

> Most people have strong feelings about something that happened to them in the past. Think about a time when you felt happy, scared, surprised, or proud. Tell about this time so that the reader will understand what happened, who was involved, how the experience made you feel, and why it was important to you.[3]

Students also answered 12 general information questions that accompanied the writing assessment. Their responses were correlated to levels of writing performance and illuminated several issues the State found important. These included: the negative impacts of television viewing, positive effects of reading, and support for teaching of writing as a precess and writing across the curriculum. The analysis was conducted by an outside contractor, also responsible for scoring the uniform writing assessments.

The writing portfolio can contain pieces from grades prior to the fourth and eighth grade "snapshot" years; works in various stages of revision; several other writing samples, including a poem, short story, play, or personal narration; a personal response to an event, exhibit, book, issue, mathematics problem, or scientific phenomenon; and prose pieces from any curricular area outside of English. As in the mathematics portfolio, the student also chooses one best piece, and writes a letter to the evaluators explaining why the piece was selected and the process of its composition.

The writing portfolios are scored by teachers. In the pilot year, approximately 150 fourth and eighth grade teachers from the sample schools did this scoring. Each portfolio and best piece was assessed by two teachers (using the writing benchmarks shown in table 7-F1) and the process took 2 days. Although it was an intense experience, the teachers' reactions were generally positive:

> . . . &spite the work load, this was an invigorating and inspiring couple of days. A few things impressed me: the uniformity of the grading; the joy of discovering various "nuggets" of good stuff; the variety and the quality of eighth grade writing.
>
> I learned a hell of a lot. The experience confirmed the prevailing sense among the writing community that language can be the close, personal ally of every self, regardless of ability, age, or station.
>
> what was most useful about this process was that teachers from an over the state saw the variety and talked about it.[4]

---

[1] Vermont Department of Education, *Looking Beyond "The Answer": Vermont's Mathematics Portfolio* (Montpelier, 1991).

[2] See Vermont Department of Education, *My Writing* VT: 1991), p. 7.

[3] Ibid., p. 19.

[4] Ibid., pp. 13-14.

Figure 7-F1—Portfolios as a "Window" on Student Feelings About Mathematics

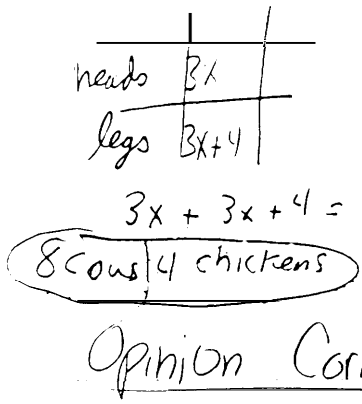
Students keep copies of their mathematics problems as well as their feelings and opinions about mathematics in their portfolios. This student's current frustration is reflected in his entry:

as you can see in this problem, and every problem in Stupid idiotic math class! I really stink in math. I have no stupid brain for it. Math is the dumbest, Studdest class ever. The person who invented it should be drug out into the street and shot! I can't ever plan on be a High-Tech mathmatitian, or an Engineer. That all Bull Crud. Sorry I am in such a bad mood but I just got a 45 on my math test. I studied hard, I want in for extra help but you know what? That all didn't do a thing for me!

Later in the year, he was faced with the following problem:

In a group of cows and chickens, the number of legs is four more than three times the number of heads. What is the least number of cows and chickens in the group?

What **follows** is **his solution, and his reaction, in what he called his "opinion corner":**



SOURCE: Vermont Department of Education, *Looking* ***Beyond*** *"The Answer": Vermont's Mathematics* ***Portfolio*** *Assessment Program,* ***Pilot*** *Year* ***Report 1990-91*** **(Montpelier VT: 1991), p. 31.**

Continued on next page

Box 7-F—"This is My Best": Vermont's Portfolio Assessment Project-Continued

In 1991-92, all Vermont schools are required to use the assessments in the target grades. Local teachers will assess the writing portfolios in their own schools, after a series of professional development sessions. They have the option of working alone, assessing only their own students' portfolios, or working cooperatively with other teachers in their schools. In late spring they will bring a sample of five portfolios to a regional meeting, where teachers from others schools will score their sample portfolios to determine a rate of reliability. A sample of portfolios from each regional meeting will be assessed at a statewide meeting to ensure that common standards are applied statewide.[5] Aware of the importance of training teachers to use new assessment tools as levers for instructional change, the State has committed 40 percent of the assessment budget to professional development.[6]

The reporting **system has also been carefully considered. Building on Vermont's tradition of town meetings, each district declares an annual Vermont School Report Day each spring. At this time community members and the press go to their schools for an analysis of assessment results and to discuss the district's response to a list of questions prepared by the State board to encourage discussion about local schooling goals and successes.**

[5]Ibid., p. 8.

[6]Ross Brewer, presentation at "Educational Assessment for the Twenty-First Century: The National Agenda," sponsored by the National Center for Research on Evaluating Standards and Student Testing, Manhattan Beach, CA, Mar. 9, 1991.

Table 7F-1—Vermont Writing Assessment Analytic Assessment Guide

Five dimensions of writing are rated on the following levels of performance: extensively, frequently, sometimes, rarely (criteria for each of these are listed)	
Purpose	The degree to which the writer's response: • establishes and maintains a clear purpose; . demonstrates an awareness of audience and task; • exhibits clarity of ideas.
Organization	The degree to which the writer's response illustrates: . unity; . coherence.
Details	The degree to which the details are appropriate for the writer's purpose and support the main point(s) of the writer's response.
Voice/tone	The degree to which the writer's response reflects personal Investment and expression.
Usage, mechanics, grammar	The degree to which the writer's response exhibits correct: . usage (e.g., tense formation, agreement, word choice); . mechanics--spelling, capitalization, punctuation; . grammar; . sentences; as appropriate to the piece and grade level.

SOURCE: Vermont State Board of Education, *This is My Best": Vermont's Writing Assessment Program, Pilot Year Report 1990-91* (Montpelier, VT: 1991), p. 6.

response may more closely approximate the real-world process of solving problems. Most performance tasks require the student to engage in a complex group of judgments; the student must analyze the problem, define various options to solve the problem, and communicate the solution in written, oral, or other forms. Furthermore, often a solution requires balancing "tradeoffs" that can only be understood when the person making the choices explains or demonstrates the rationale for the choice. Performance assessment tasks make it possible to trace the path a student has taken in arriving at the chosen solution or decision.

Performance assessment attempts as much as possible to assess desired behavior directly, in the context in which this behavior is used. Tasks chosen for testing must sample representatively from the desirable skills and understandings: demonstrating ability to write a persuasive argument might be reflected in asking students to write a paragraph convincing the teacher why an extension is needed on an assignment; demonstrating an understanding of experimental design might involve designing and conducting an experiment to find out if sow bugs prefer light over dark environments; showing one's facility with the French written language might involve translating a French poem into English. In each of these cases, it is possible to conduct other kinds of tests that can accompany the performance task (e.g., vocabulary tests, lists of procedures,

Box 7-G—Michigan's Employability Skills Assessment Program

In an effort to ensure that Michigan's high school graduates acquire skills necessary to remain competitive in an increasingly technological workplace, the Governor's Commission on Jobs and Economic Development convened the Employability Skills Task Force in 1987. The Task Force, made up of leaders from business, labor, and education, was charged with identifying the skills Michigan employers believe important to succeed in the modem workplace. The Task Force concluded that Michigan workers need skills in three areas:

- *Academic skills*, such as the ability to read and understand written materials, charts, and graphs;
- *Teamwork skills,* such as the ability to express ideas to colleagues of a team and compromise to accomplish a goal; and
- *Personal Management skills*, such as the ability to meet deadlines and pay attention to details.[1]

The Task Force also served as a policy advisory group on the development of Michigan's Employability SkiIls Assessment Program for the State's high schools. The Task Force concluded that student portfolios would best describe the strengths and weaknesses of individual students in the skill groups, and could serve as the basis for planning an individual skills development program for each student.

The portfolio program was piloted during the 1990-91 school year in 22 school districts. Districts were encouraged to apply the program to a cross section of students in order to emphasize that the program was designed for everyone, not just noncollege-bound youth.

To help students, the State provided several tools including three portfolios (one in each skill area), a portfolio information guide for the student, a parent guide for the student's parents, a personal rating form to be filled out by students, teachers, and parents, and a work appraisal form for employers to complete.

Each of the three portfolios, Academic, Teamwork and Personal Management, stresses skills considered important in that particular area. Students are responsible for updating their portfolios with sample work and information about grades, awards, and recommendations. For example, the captain of the school track team might ask her coach for a letter of recommendation to place in her Teamwork portfolio as proof of her leadership ability. If students feel they are lacking in a particular skill category, they can seek out an activity designed to help them master that skill. In this way students are expected to *discover, develop,* and *document their* 'employability skills." It is envisioned that the portfolios will serve as 'resume builders. When applying for jobs, students will use their portfolios to demonstrate employability skills.

It is difficult to assess the results of the Employability Skills Assessment since the program is so new. The few collected responses have been mixed. Schools that have taken the program to heart, contacting local businesses and informing them of the program, have been enthusiastic. Some schools have even invited local business managers to assess individual student's portfolios. Other schools, however, have been less satisfied. Some are resisting suggested changes because they appear incompatible with other reform efforts; others are hesitant to involve business in what is viewed primarily as the job of the schools. Michigan law now requires every school to design a portfolio system to assess ninth graders beginning in the 1992-93 school year. The State's Department of Education plans to continue piloting the Employability program.

[1]A similar emphasis on the blend of academic, cooperative, and personal skills underlies a recent U.S. Department of Labor report. See U.S. Department of Labor, Secretary's Commission on Achieving Necessary Skills (SCANS), *What Work Requires of Schools* (Washington, *DC:* June 1991).

[2]Edward D. Roeber, Michigan Department of Education, personal communication, Oct. 22, 1991.

questions about content), but in performance assessment direct performances of- desired tasks are evaluated.

Performance assessments focus on the process and the quality of a product or a performance. Effectiveness and craftsmanship are important elements of the assessment; getting the ''right answer' is not the only criterion.[52] The process as well as the results are examined in solving a geometric proof, improving one's programming skills, or formulating a scientific hypothesis and testing it.

[52]Grant Wiggins, ' 'Authentic Assessment: Principles, Policy and Provocations,' paper presented at the Boulder Conference of State Testing Directors, Boulder, CO, June 1990.

Box 7-H—Advanced Placement Studio Art Portfolios

While the idea of portfolios for large-scale testing is considered a novel idea, portfolios have been the heart of the Advanced Placement (AP) examination on for studio art for nearly 20 years.[2] The purpose of the AP Studio Art Portfolio Evaluation is to certify that a high school student has produced works that meet the achievement level expected of first year college students in studio art. The cost to the student is $65, the same as for other AP examinations. There are several points that make the assessment of particular interest:

- The assessment is conducted entirely through evaluation of the work contained in the student portfolio. There are no essays, no questions to answer, no standard paper-and-pencil examination.
- It is a considered "high-stakes" assessment, for, like all AP examinations, students must receive a passing grade (a score of 3 or higher on a 1 to 5 ranking) to earn college level credit for the course.[3]
- Despite the fact that the topic is a "subjective" one like art, administration and scoring are standardized and conducted in an objective manner.
- There is no set curriculum; teachers have great flexibility in their choice of approach, organization, assignments, and so forth
- A high degree of student initiative and motivation is required.
- The program has won the respect of teachers and students at both the high school and college level and there is little controversy surrounding it.

Standardization of Portfolio Submissions

Students submit a portfolio based on the work they have created during the year-long AP studio art course.[4] A student can choose one of two evaluations: the drawing portfolio or general portfolio evaluation. In the *drawing portfolio, there* must be six original works no larger than 16 inches by 20 inches, and from 14 to 20 slides on an area of special concentration. The concentration is a single theme (e.g., self portraiture) developed by the student. Some of the concentrations chosen as exemplary in recent years have included cubist still-life drawings, manipulated photographs, wood relief sculptures, still lifes transformed into surreal landscapes, and expressionist drawings that serve as social commentary.[5] Another 14 to 20 slides illustrate breadth. The *general portfolio* is set up in much the same format.[6] Film and videotapes maybe submitted in the concentration section.

Standardizing Artistic Judgment

In June 1991, nearly 5,000 portfolios were submitted for the evaluation. These were graded by a panel of 21 *readers (scorers)* assembled at Trenton State College in Trenton, New Jersey. The readers all teach either AP studio art or analogous college courses; scoring took 6 days.

Each grading session began with a standard-setting session. A number of portfolios were presented to the assembled readers, roughly illustrating all the possible scores. These examples were chosen beforehand by the *chief reader* for the whole evaluation and the *table leader* for each section; their selection and judgment were guided by their experience of teaching. There was no general scoring rubric per se; no analytic scales of *primary traits* as there are in the evaluation of writing. As one former chief reader suggested:

---

[1] Much of this discussion comes from Ruth Mitchell and Amy Stempel, Council for Basic Education, "Six Case Studies of Performance Assessment," OTA contractor- March 1991.

[2] Studio art was added to the Advanced Placement (AP) program in two stages—the general portfolio in 1972 and the drawing portfolio in 19S0. A separate AP art history course is also offered; its examination has a more typical format of multiple-choice and free-response items.

[3] Colleges have varying polices regarding AP credits. Some grant exemption from freshman-level courses, while others require students to take the introductory courses, but grant a certain number of elective credits. In general, students can reduce the number of courses required to graduate from college by passing these AP college-level courses in high school. Thus there is a strong financial incentive to succeed on the AP examination.

[4] Not all schools offer a separate AP course. A separate AP studio art course is "almost a luxury"; in some schools, a small number of AP students work alongside other students m regular classes, while other studenta submit work done independently during their museum courses. Alice Sims-Gunzenhauser, Educational Testing Service consultant, AP studio art, personal communication, November 1991.

[5] Ibid.

[6] Only four works are required in the original work portion. The breadth section specifies that eight slides illustrate drawing skill, with four each in three other categories (color, design, and sculpture).

> Factors that are included in assessing quality include imagination; freshness of conception and interpretation; mastery of concepts, composition, materials, and techniques; a distinct sense of order and form; evidence of a range of experience; and, finally, awareness of art-historical sources, including contemporary artists and art movements. It is not expected that every student's portfolio will reflect all of these considerations to the same degree. . . . What you're really after is a mind at work, an interested, live, thinkin g being. You want to see engagement. Recognition of it comes from long experience and you intuit it.[7]

In commenting on how this approach related to judgments in other disciplines, he noted:

> There are more things that join us together than separate us. You can make those judgments as accurately as you can in mathematics or in writing or in any other subjects. These other subjects frequently have much more difficulty than we do in the visual arts in agreeing on standards.. . . You get a sense for copied work, a sense when there's engagement, when inspiration, belief, direct involvement are present or absent.[8]

The portfolios chosen to exemplify each grade remained on display throughout the scoring as references for comparison. The readers assigned scores to each part separately, on a scale from 1 to 4. Originality of work was scored independently by three readers; concentrations and breadth by two readers. The scores were manipulated by computer to arrive at a raw score (1 to approximately 100) to which the three sections (original work, concentration, and breadth) contribute equally. If discrepancies of 2 or more points between two readers' evaluations of the same section occurred, the chief reader reviewed the section and reconciled the scores. The chief reader might speak with a reader and use the models to reinforce the agreed standard.

After all portfolios had been evaluated, *cutoff scores were* determined and the total scores then converted to the AP grades on a scale of a high of 5 to a low of 1. Although assigning the cutoff scores (i.e., determining the lowest total score to receive an AP grade of 5 on down) is the chief reader's responsibility, there was input from a long debriefing meeting of all readers and from statistical information supplied by the computer, historical data regarding previous years' cutoff scores, composite and raw scores for present year's candidates, and tables showing the consequences of choosing certain cutoff scores, in terms of percentage of students receiving 5,4, 3, and so on. The scores overall were roughly distributed in a bell curve, with most receiving a 3, but fewer 1s than 5s. (Colleges do not usually accept either 2 or 1 scores, so a 2 can perform the same function as a 1 (i.e., denying the awarding of college credit) without making such a negative judgment of a student's work.)

Impacts on Students

In the process of creating portfolios for AP studio art, students begin to develop artistic judgment about their own work and that of their fellow students. Students are taught to criticize each other's work constructively. As they learn how to select works for their own portfolios, they also learn to communicate with each another about areas that need improvement. This climate of reflection is an important byproduct of portfolio assembly.

Another key factor is motivation. As one teacher suggested, the course is a test of students' self motivation.[9] For example, students must have the ability to envision a concentration project and then work steadily toward completing it for 8 or 9 months, solving problems as they arise. The work on all three sections must be timed so that the entire portfolio is ready at the deadline. Pieces have to be photographed for slides and final selections made for the collection of original works.

Broad Public Acceptance

Another important point is the relative lack of controversy surrounding judgment of a subject traditionally considered subjective. This respect comes from the long history of the evaluation and the refinements the Educational Testing Service has made to the jury method of judging works of art, based on collective, but independent, judgments by teachers who are involved in the day-today teaching of students like those being assessed. These teachers are well trained in the objectives of the course as well as the performance standards for each level, and their judgment is valued and respected.

Walter Askin, *Evaluating the Advanced Placement Portfolio in Studio Art* (Princeton, NJ: Advanced Placement Program, 1985), pp. 2S.

[9] Ibid.

Raymond Campeau, AP studio art teacher in Bozeman, MT, in Mitchell and Dempel, op. cit., footnote 1.

The product or record of a performance assessment is scored by teachers or other qualified judges. In classroom testing this observation is done by the teacher, but in large-scale assessments, products, portfolios, or other records of work are scored by teams of readers. How much psychometric rigor is required in making these qualitative and complex judgments varies with the purpose of the assessment; less rigor is acceptable for use within the classroom for diagnostic purposes than would be acceptable in large-scale testing programs where comparability is essential. What is important is that performance tests are not '(beyond standardized testing"; they should be standardized whenever comparability is required.[53]

The criteria **for judging performance assessments are clear to those being judged.** Criteria for judging successful performance must be available and understood by teachers and students. The tasks and standards must allow for thorough preparation and self-assessment by the student,[54] if the test is to be successful in motivating and directing learning, and in helping teachers to successfully guide practice.[55] The goal in performance assessment is to provide tasks that are known to the student—activities that not only can but should be practiced. Performance assessment tasks are intended to be ''taught to, ' integrating curriculum and assessment into a seamless web. Practice required for good performance is understood to increase and stimulate learning.

Performance assessment may take place at one point or over time. Typically it examines patterns of student work and consistency of performance, looking at how an individual student progresses and develops. This is particularly true of portfolios, which are collections of student work over time.

While multiple-choice and other paper-and-pencil examinations are almost exclusively taken by an individual student, some performance assessments can be and are often conducted as group activities. This group activity reflects increasing interest in student team work and cooperation in solving tasks as a valued outcome of the educational process. Proponents suggest that, if teamwork is a valued skill, it should be assessed. However, the problems associated with inferring individual effort, ability, and achievement from group performances are significant. Individual performance and performance as a member of a group are often **scored as two** separate pieces **of the** assessment.

Performance assessments **are** generally criterion-referenced, rather than norm-referenced. Although it **is important to collect** information on how **a** wide **range of** students respond **to performance** assessment tasks, the **primary** focus is on scoring students relative **to standards of** competence **and mastery.** Developers **of performance** assessment are seeking test-based indicators that portray individual performance with respect **to** specific educational goals rather than those **that** simply compare an individual's performance **to a** sample **of other test takers.**

Performance Assessment Abroad

The standardized, machine-scored, norm-referenced, multiple-choice **tests so common in this country** for **large-scale testing are** rarely used in other countries. In fact, these are often referred to generically as "American **tests."** Instead, **examinations like the** French *Bac, the German Abitur,* or the English General Certificate of Secondary Education or ''A levels, " generally require students to create rather than **select answers,** usually in the format of short-answer **or** longer **essay** questions or, in some **cases,** oral examinations. These examinations share several **of the** characteristics noted above regarding performance assessments in American schools: they are typically **graded by** teachers, the content **is** based on a common curriculum or syllabus for which students prepare **and** practice, **and the** questions are made public **at the** end of the examination period.

It is important to note, however, **as** discussed in chapter 4, that these tests are most commonly used for selection **of** students **into** postsecondary education rather than for classroom diagnosis or school accountability. Consequently, several of **the characteristics noted in American** performance assessments are not present in these examinations. That is, the examinations **are** usually individual assessments, **with no** opportunity for group activities; they

[53]Frederick L. Finch, ''Toward a Definition for Educational Performance Assessment,' paper presented at the ERIC/PDK Symposium, Alternative Assessment of Performance in the Language Arts, Bloomington IN, Aug. 23, 1990.

[54]Wiggins, op. cit., footnote 52.

[55]Center for Children and Technology, Op. cit., footnote 5, p. 3.

do not involve self assessment or student involvement in evaluation; the examinations are timed rather than open-ended; and, even when administered over several days, they do not involve tasks that take several testing periods or longer time periods to complete.

Nevertheless, European experience can be informative. For example, the national assessment in Holland structures performance-based assessments for students by designing comprehensive problems for the year-end examinations. A committee of teachers in art history, for example, selects a unifying subject (e.g., ''revolution"). Students are provided with information packages to guide their study of art throughout the year in ways that help them to critically develop the theme (e.g., readings and lists of museums). Teachers are encouraged to work with students to help them develop individual interpretations and points of view. This assessment approach supports students in doing individualized in-depth work in a context of shared ideas, procedures, and problems.[56]

The United Kingdom is the furthest along of European countries using performance assessment for national testing. The Education Reform Act of 1988 set in place a national curriculum, which has at its core a set of attainment targets for each of the 10 foundation subjects to be taught to all students. These statements of attainment provide the basis for the criterion-referenced assessment system. Teachers have been given detailed, clearly defined Standard Assessment Tasks (SATs) to use with all students at or near the completion of four levels or "key stages" of schooling: ages 7, 11, 14, and 16. Each SAT carries with it levels of attainment and the tasks for determining levels, described in manuals provided to all teachers. The tasks involve one or more components of every aspect of performance: reading, writing, speaking, listening, investigating, demonstrating, drawing, experimenting, showing, and assembling. The tasks were developed through research conducted at schools across the United Kingdom by the National Foundation for Education Research in England and Wales.

Following a 2-day teacher training period, three sets of SATs were piloted in May 1990, testing 6,219 students in level one (age 7) in schools throughout England and Wales. Each was constructed around an overall theme hoped to engage the interest of 7-year-olds: Toys and Games, Myself, and The World About Me.

Evaluation data and recommendations reflect widespread concern with the extremely detailed and directive nature of the assessment system:

> In view of the issue of time and workload . . . an inescapable conclusion must be that future SATs should be significantly shorter than those piloted. SEAC [School Examinations and Assessment Council] are likely to recommend that the SAT is to be carried out in a ***three week period, and to take not more than half the teacher's time during those three weeks. . . . The*** number of activities that can be fitted in will need to be reduced to about six in order to be sure that these time constraints can be observed. . . . The model of a SAT covering all or most, or even half of the ATs has now been proven to be unworkable in light of the number and nature of ATs included in the final statutory orders. . . . The SAT should still offer teachers the opportunity to embed the assessments within a coherent cross-curricular theme.[57]

How far the United Kingdom will be able to move forward on this ambitious assessment plan that requires so much teacher time is still under debate. However, the close tie to the national curriculum strengthens the likelihood that the SATs will be maintained as centerpieces for assessment.

Finally, some countries are experimenting with the use of portfolios for large-scale testing activities, and many are looking to the United States for guidance in this field. Because the United States is widely respected as a leader in psychometric design, many other countries are watching with interest how we match psychometric rigor to the development of performance assessment techniques.

Policy Issues in Performance Assessment

Various direct methods of assessing performance have long been used by teachers as a basis for making judgments about student achievement within the classroom. Teachers often understand intuitively their own potential for errors in judgment and the

[56] Center for Children and Technology, op. cit., footnote 5, p. 8.

[57] National Foundation for Educational Research/Bishop Grosseteste College, Lincoln Consortium, The *Pilot Study of Standard Assessment Tasks for Key Stage 1—Part 1: Main Text & Comparability Studies* (Berkshire, England: March 1991), p. 10, emphasis added.

ways in which student performance can vary from day to day. As a result they use daily and repeated observations over time to formulate judgments and shape instruction. An error in judgment on one day can be corrected or supplanted by new observations the next.

The stakes are raised when testing is used for comparisons across children, classrooms, or schools, and when test results inform important decisions. As noted by several experts in test design and policy:

> . . . . when direct measures of performance take on an assessment role beyond the confines of the classroom--portfolios passed on to next year's teacher, district wide science laboratory tasks for program evaluation, or state-mandated writing assessments for accountability are just a few examples--whatever contextual understanding of their fallibility may have existed in the classroom is gone. In such situations, a performance assessment, like any other measurement device, requires enough consistency to justify the broader inferences about performance beyond the classroom that are likely to be based on it. Most large-scale performance assessments are being proposed today for fundamentally different purposes from those of classroom measurement, such as monitoring system performance, program and/or teacher evaluation, accountability and broadly defined educational reform. Even though none of these uses typically involves scores for and decisions about individual students, each is a high stakes application of an educational measurement to the extent that it can effect a wholesale change in a school program affecting all students.[58]

The feasibility and acceptance of the widespread use of performance assessment by policymakers must rest on consideration of a number of important issues. In addition, the purpose of a particular test will, in large part, determine the relative importance or weight that should be given to each of these issues.

Standardization of Scoring Judgments

One of the first concerns about the applicability of performance assessment to large-scale testing is the extent to which human judgment is required in scoring. Variability across judges and potential for bias in scoring could create impediments to using these methods for high-stakes testing. For scores to yield meaningful inferences or comparisons, they must be consistent and comparable. A student's score should reflect his or her level of achievement, and should not vary as a function of who is doing the judging. A key feature of performance assessment is the complexity of judgment needed for scoring; however, this very complexity, some suggest, may be a barrier to its widespread implementation in situations where comparability matters.

For performance assessment to fulfill its promise, it must meet challenges regarding reasonable standards for reliable scoring, whether this scoring is done by individuals, teams, or by machines programmed to simulate human judgment. This is an area where test publishers have experience and expertise to offer school districts and States considering performance assessments. As noted above, Arizona has hired the Riverside Publishing Co., in part because of experience with the Arizona educators and their curriculum and past testing activities (the Iowa Tests of Basic Skills and the Tests of Achievement and Proficiency programs), but also because the publishers claim expertise in field testing items or tasks and providing scales that meet previous standards for reliability.

Because there has been considerable research by curriculum experts and the research community on developing and scoring essays and writing assessments, they present a model that students, teachers, and the general public can appreciate. Scoring has been made more systematic and reliable by a number of procedures. Scoring criteria are carefully written to indicate what constitutes good and poor performance; representative student papers are then selected to exemplify the different score levels. Panels of readers or scorers are carefully trained until they learn to apply the scoring criteria in a manner consistent with other readers. In most large-scale writing assessments, each essay is read by two readers. When significant scoring discrepancies occur, a third reader (often the ''team leader' reads and scores the essay. Various scoring systems can be employed from holistic (a single score is given for the quality of the writing) to more fine-grained analytic scores (each essay is rated on multiple criteria). Table 7-1 presents an example of one analytic scoring system that focuses on rating five aspects of the student's writing: organization, sen-

[58] Stephen Dunbar, Daniel Koretz, and H.D. Hoover, "Quality Control in the Development and Use Of Performance Assessments,' paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 1991, p. 1.

Table 7-I-Criteria for Analytical Scoring

Scale:	1	2	3	4	5
Organization:	Little or nothing is written. The essay is disorganized, incoherent, and poorly developed. The essay does not stay on the topic.	The essay is not complete. It lacks an introduction, well-developed body or conclusion. The coherence and sequence are attempted, but not adequate.	The essay is well-organized. It contains an introductory supporting and concluding paragraph. The essay is coherent, ordered logically, and fully developed.
Sentence structure:	The student writes frequent r unons or fragments.	The student makes occasional errors in sentence structure. Little variety in sentence length or structure exists.	The sentences are complete and varied in length and structure.
Usage:	The student makes frequent errors in word choice and agreement.	The student makes occasional errors in word choice or agreement.	The usage is correct. Word choice is appropriate.
Mechanics:	The student makes frequent errors in spelling, punctuation, and capitalization.	The student makes an occasional error in mechanics.	The spelling, capitalization, and punctuation are correct.
Format:	The format is sloppy. There are no margins or indentations. Handwriting is inconsistent.	The handwriting, margins, and indentations have occasional Inconsistencies--no title or inappropriate title.	The format is correct. The title is appropriate. The handwriting, margins, and indentations are consistent.

SOURCE: Adams County School District No. 12, Northglenn, CO.

tence structure, use of language, mechanics, and format.

In the California Assessment Program's writing assessments, essays and answers are read by a single reader, but there are a variety of techniques used to maintain consistency of grading. Marked papers already read are circulated back into the pile to see if they get the same grade again; the table leaders randomly reread papers to make sure that readers are consistent; examples of graded papers are kept available for comparison as ''anchors.' Using these techniques, the inter-rater reliability for the CAP writing assessment is about 90 percent in a single year, although less high for the same question across years. This remains an unsolved problem for CAP and other States and districts using group grading if they want to make longitudinal comparisons.[59]

Other scoring questions related to design have yet to be solved. One of these is the time allotted for producing a composition. A 15-minute essay, with no chance for revision, may not be a true test of the kind of writing that is valued. Thus, testing time affects how reliably the writing sample reflects writing skill. Additionally, specifying scoring criteria and rating scale format are no easy matters. Although research has recently provided some empirical analysis of the features of writing that distinguish skilled from unskilled writing, some suggest that the criteria applied to a particular assessment may represent arbitrary preferences of the group designing the scale. It is difficult but necessary to come to a consensus on these issues.

Photo credit: Educational Testing Service

Essays and writing samples can be graded consistently if teachers are trained to apply scoring criteria based on common standards. In this example, the Educational Testing Service has assembled experienced teachers to read and score essays written by students across the country on their Advanced Placement examinations.

[59]Mitchell and Stempel, op. cit., footnote 2.

Policy Implication

Writing assessments, essays, and courses like AP studio art have a proven track record of assessing performance in a standardized and reliable fashion. Whether these same procedures for obtaining consistency in scoring can be applied to other forms of performance assessment (e.g., portfolios, exhibitions, oral examinations, and experiments) is as yet largely unexplored. Moreover, although inter-rater reliability is relatively high (for judging essays), it still contains some variation that may add error to scores. What degree of error in measurement is acceptable depends, in part, on the purposes of the test. Careful development **of scoring criteria and intensive training of judges are key to establishing consistency of judgment.**

General inability of Scores: Are the Tests Valid Estimates of What Students Know?

Most students, current and former, can remember taking an essay test and feeling 'lucky' because the questions just happened to hit topics they knew well; a high score, perhaps higher than their study and knowledge actually deserved, was the result. More likely, they remember the time they "bombed" on a test, unjustly they felt, because the essays covered areas they had not understood or studied as well. One of the advantages of item-based tests is that a large number of items can be given in a limited amount of testing time, thereby reducing the effect of a single question on the overall score.

When only a few tasks are used there is a much higher risk that a child's score will be associated with that particular task and not generalize to the whole subject area that the test is meant to cover. Writing assessment provides a particularly good example of the problem of generalizing results from a single question. In many cases a 30-minute essay test is given to students in order to estimate something about their overall ability to write well. However, a number of different kinds of writing tasks can be given. The National Council of Teachers of English lists five methods of communication in writing--narrating, explaining, describing, reporting, and persuading-that provide the framework for much of the classroom instruction in writing.[60] When tests are given, the essay question (or prompt) can be in any of these modes of discourse.

Two kinds of information are needed to make essay test results generalizable. First, would two different essays drawn from the same mode of discourse result in the same score? Results of several studies cited in a recent review suggest that agreement between two essays written by the same child in the same writing mode is not very high (reliability scores range from 0.26 to 0.46).[61] Second, are scores for essay prompts from different modes of writing similar? For example, if a student is asked to write a narrative piece, will the score for this prompt be similar to a score the same child receives for writing a persuasive piece? Results of several investigations of writing assessments indicate that correlations across tasks are low to moderate.

Other factors such as the topic of the essay, the time limit, and handwriting quality have been shown to affect scores on essay tests.[62] Preliminary results suggest that a number of tasks would need to be administered to any given child (and scores aggregated across tasks) before a sufficiently high level of reliability could be achieved to use these tests for making decisions about individuals. One investigation of these issues has suggested that six essays, each scored by at least two readers, would be needed to achieve a level of score reliability comparable to that of a multiple-choice test.[63]

One of the particular problems faced by performance assessment is that of substantiating that similar generalizations to the whole domain can be made on the basis of a few tasks. Very little research exists that can shed light on the extent to which different performance assessment tasks intended to assess the

[60] A.N. Hieronymus and H.D. *Hoover, University* of Iowa, *Writing: Teachers Guide,* Iowa Tests of Basic Skills, Levels 9-14 (Chicago, IL: Riverside Publishing Co., 1987).

[61] Dunbar et al., op. Cit., footnote 58. See also peter L, Cooper, *The Assessment of Writing Ability: A Review of Research,* ' GRE Board research report GREB No. 82-15R (Princeton, NJ: Educational Testing Service, May 1984).

[62] Cooper, *op.* cit., footnote 61"

[63] H.M. Breland, R. *cop,* R.J. Jones, M.M. Morns, and D.A. Rock, ' 'Assess@ Writing Skill, " research monograph No. 11, prepared for the College Entrance Examition Board, 1987, cited in Wayne Patience and Joan Auchter, "Monitoring Score Scale Stability and Reading Reliability in Decentralized Large-Scale Essay Scoring Programs," paper presented at the annual meeting of the National Testing Network in Writing, Montreal, Canada, April 1989.

same set of skills produce similar scores. Data from writing assessments suggest, for example, that a child who produces a superior essay in one format may write only a mediocre one on a different day in a different format.

The issue of generalizability--whether a child's performance on one or two tasks can fairly represent what he or she knows in that area-is an important one that greatly influences the conclusions that can be made from tests. Establishing generalizability is particularly critical if a test is going to be used to make decisions about individual students. Again the experience of writing assessment offers important lessons for other forms of performance assessment:

> It has long been known that neither an objective test nor a writing sample is an adequate basis for evaluation of an student, whether for purposes of placement, promotion or graduation. [One author] . . . noted that a reliable individual evaluation would require a minimum of four writing samples, rated blindly (i.e., without knowledge of the student's identity) by trained evaluators. It is a continuing scandal of school testing programs that patently inadequate data are used for placement and categorization.[64]

Policy Implications

Issues of task generalizability present an important challenge to policymakers and test developers interested in expanding the uses of performance assessment. If individual scores are not required, however, sampling techniques can mitigate these issues. For example, many large-scale assessments of writing administer multiple prompts in each mode but each individual child only answers one or two of a larger number of prompts. The large number of children answering any one prompt, however, allows generalizable inferences to be made within and across modes about levels of writing achievement for students as a whole. The use of sampling techniques can allow policymakers and administrators to make generalizable inferences about schools or districts without having to administer prohibitively long or costly tests to every student (see box 7-I).

costs

The costs of performance assessment represent a substantial barrier to expanded use. Performance assessment is a labor-intensive and therefore costly alternative unless it is integrated in the instructional process. Essays and other performance tasks may cost less to develop than do multiple-choice items, but are very costly to score. One estimate puts scoring a writing assessment as 5 to 10 times more expensive as scoring a multiple-choice examination,[65] while another estimate, based on a review of several testing programs administered by ETS, suggests that the cost of assessment via one 20- to 40-minute essay is between 3 to 5 times higher than assessment by means of a test of 150 to 200 machine-scored, multiple-choice items.[66] Among the factors that influence scoring costs are the length of time students are given to complete the essay, the number of readers scoring each essay, qualifications and location of readers (which affects how much they are paid, and travel and lodging costs for the scoring process), and the amount of pretesting conducted on each prompt or question. The higher these factors, the higher the ratio of essay to multiple-choice costs. The volume of essays read at each scoring session has a reverse impact on cost—the greater the volume, the lower the per item cost.[67]

Is performance-based assessment worth the significantly higher direct costs of scoring? First, it is important to recall that high direct costs may overestimate total costs if the indirect costs are not taken into account. As explained in chapter 1, comparison of two testing programs on the basis of direct costs alone is deceiving. Because performance assessment is intended to be integrated with instruction, its advocates argue that it is less costly than it

---

[64] Suhor, op. cit., footnote 40. The author referred to is Paul Diederich, *Measuring Growth in English* (Urbana, IL: National Council of Teachers of English, 1974),

[65] John Fremer, "What Is So Real About Authentic Assessment?" paper presented at the Boulder Conference of State Testing Directors, Boulder, CO, June 10-12, 1990.

[66] The testing programs reviewed included: ". . . the Advanced Placement Program, several essay assessments we operate for the state of California, the College Level Examination program, the Graduate Record Exam, NAEP, the National Teacher Examination Programs, and the English Composition Test with Essay of the Admissions Testing program. . . ." Penny Engle, Educational Testing Service, Washington, DC, personal communication% June 101991. Multiple-choice tests are scored for $1.20 per student; in contrast, scoring of the Iowa Tests of Basic Skills writing test costs $4.22 per student. Frederick L. Finch, vice president, The Riverside Publishing Co., personal communication, March 1991.

[67] Engle, op. cit., footnote 66.

Box 7-1—Assessing Hands-On Science Skills

The National Science Foundation has supported a research project that attempts to explore reliability, transferability, and validity issues affecting performance tasks for large-scale science assessments.[1] The researchers first developed three different hands-on laboratory tasks for children to solve. Each requires students to conduct an experiment and manipulate equipment. In the "PaperTowels" experiment, students had to determine which of three kinds of paper towels soaked up the most water. The second task required students to figure out the contents of a number of "mystery boxes" containing wires, batteries, and/or light bulbs. The third assessment had students determine what kinds of environments sow bugs prefer (e.g., dark or light, dry or damp). Students were observed by experts while they performed the experiments; the experts scored students according to the procedures they used as well as the findings of the investigation.

Evidence about the validity of these measures was obtained by giving the participating students a traditional multiple-choice standardized test of science achievement, in order to compare the scores they obtained on their hands-on experiments with the scores received on multiple-choice tests. In addition, the performance of students who had been taught using a hands-on approach to science was compared to those studying under a more traditional approach.

Results provide some encouragement and some warnings. Among the findings of these initial development efforts with fifth and sixth graders were the following:

- **Hands-on investigations can be reliably scored by trained judges.**
- **. Performance on any one of the tasks was not highly related to that on the others. A student could perform well on one hands-on task and quite poorly on another. This suggests that a substantial number of tasks will be needed unless matrix" sampling can be used**
- **Hands-on scores were only moderately related to student's scores on the traditional multiple-choice science test, suggesting that different skills are being tapped.**
- **Students who had been taught with a hands-on approach did better on these tasks than did students from a traditional science classroom, suggesting that the tests are sensitive to classroom instruction.**

[1]Richard J. Shavelson, Gail P. Baxter, Jerome Pine, and Jennifer Yure, "New Technologies for Large-Scale Science Assessments: Instruments of Educational Reform," symposium presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 1991.

appears. Resolution of this issue requires agreement on the degree to which any given testing options under consideration are integrated with regular instruction.

Second, although a performance assessment may provide less data than a typical multiple-choice test, it can provide richer information that sheds light on student capacities not usually accessible from multiple-choice tests. Even in an externally scored writing assessment, for example, teachers can gain insight into students' writing difficulties by looking not just at the raw scores, but at the writing itself. Similarly, some outcomes that cannot be measured on multiple-choice tests (e.g., ability to work cooperatively in a group) can be assessed in performance tasks.

Finally, many educators maintain that the staff development that accompanies performance assessment is in itself a valuable byproduct. For example, when teachers gather to discuss what distinguishes **a weak piece of writing from an acceptable or an excellent piece of writing, they learn from one another and internalize the teaching standards.**

The major problem in approaching an analysis of the costs of performance assessment is a lack of a common base for the information. When the Council of Chief State School Officers compiled a chart of performance assessments in the States in order to make comparisons, they asked for reporting under the category of "costs." As the data came in, the numbers fluctuated dramatically, because different respondents thought of costs differently: some reported costs of development ($2 million in one case), some costs of administration ($5 per student), and some combined them. In the end, the researchers decided to eliminate the question altogether because it could provide no meaningful information and

Hands-on assessments like this are costly in time, equipment, and human resources. Because of this, these investigators also sought "surrogate tasks" that might provide much of the information obtained from hands-on tasks but at considerably lower cost. To this end they created the following surrogates for the three experiments, listed in order of "conceptual verisimilitude" (similarity to the hands-on experiments):

- **laboratory notebooks students kept as a record of their experiments;**
- computer simulations;
- short-answer, paper-and-pencil questions based on the experiments; and
- multiple-choice items based on the **hands-on procedures.**

The researchers then examined the extent to which these various surrogates were exchangeable for the hands-on benchmark tasks. If simpler, less costly methods can provide the same information, why not use them? Preliminary findings from these investigations suggest the following:

- **Laboratory notebooks** provide the best surrogate for the hands-on investigation and can acceptably be used in lieu of direct observation.
- . In the computer simulations, the computer saved all the child's moves, so they could be replayed and scored by the evaluator. The average time required for grading was about one-tenth of that needed for observing hands-on investigations-suggesting that computer simulations can offer a big savings in skilled personnel time.
- . Neither the computer simulation nor the paper-and-pencil measures appeared to be adequate substitutes for the **benchmark hands-on procedure. The computer simulation showed considerable variability for individual students-some individuals appear to do very well on this type of test while others do not.**
- **The students** enthusiastically participated in the hands-on procedures as well as the computer simulations.

As investigators throughout the country begin to develop new performance assessments, they will need to collect data like this in order to evaluate the technical quality of their new measures. As one of the investigators involved in the above study concludes: ". . . these assessments are delicate instruments that require a great deal of piloting to fine tune them."[2] Because so many investigators are experimenting in uncharted testing and statistical territories, research support will be needed to encourage the collection of test data and the dissemination of results so that others can learn from data that are innovative, instructive, and yet costly to obtain.

[2]Richard J. Shavelson, "Authentic Assessment: The Rhetoric and the Reality," paper presented at a symposium at the annual meeting

would require extensive explanation no matter what it included.[68]

In light of these uncertainties about the relative costs of testing programs, some school systems are striving for improved definitions and better cost data. In California, for example:

> The lead consortium is required to develop a cost-benefit analysis of existing vs. various types of alternative assessment for consideration by the California Department of Education and the State Board of Education. The cost-benefit analysis should consider payoffs, tradeoffs and advantages or disadvantages of alternative vs. existing assessment practices. The testing costs of alternative assessments, especially the staff development component, should be considered as a part of overall curriculum costs. Teachers' renewed motivation and commitment to the Curriculum Frameworks should be viewed as a major element in the cost-benefit analysis.[69]

Policy Implications

In considering the costs of performance assessment, policymakers may wish to adopt a more inclusive cost-benefit model than has typically been considered for testing. Benefits in the areas of curriculum development and teacher enhancement (staff training) may offset the higher costs associated with performance assessment. However, little data has been collected to date; a broader and deeper analysis will be required before judgments can be made.

[68]Mitchell and Stempel, op. cit., footnote 2, p. 11.

[69]California Department of Education, California Assessment Program, "Request for Applications for the Alternative Assessment Pilot Project," unpublished document, 1991.

Fairness

There has long been a concern about the effect of background factors such as prior experience, gender, culture, and ethnicity on test results. Achievement tests, for example, need to eliminate the effect of background factors if they are to measure learning that has resulted from instruction. A combination of statistical and intuitive procedures have been developed for conventional norm-referenced tests to eliminate or reduce background factors that can confound their results. Little is known, however, about how background factors may affect scores on performance assessments.

In addition, judgments about fairness will depend a great deal on the purposes of the test and the interpretations that will be made of the scores. For example, on a test that has no significant personal impact on a student, such as the National Assessment of Educational Progress, it is reasonable to include problems that require the use of calculators even though student access to calculators may be quite inequitable. On the other hand, equitable access would be an important consideration if the assessment were one that determined student selection, teacher promotions, or other high-stakes outcomes .[70]

Performance assessments could theoretically lead to narrowing the gap in test scores across those who have traditionally scored lower on standardized multiple-choice achievement tests. By sampling more broadly across skill domains and relying less heavily on the verbal skills central to existing paper-and-pencil tests, proponents hope that these differences might be minimized. Performance assessments, by providing multiple measures, may be able to give a better and therefore fairer picture of student performance.

On the other hand, performance assessments could exacerbate existing differences between groups of test takers from different backgrounds. Some minority group advocates, for example, fear that tests are being changed just when students from racially diverse backgrounds are beginning to succeed on them. They worry that the rules are being changed just as those who have been most hurt by testing are beginning to learn how to play the game.

The President of the San Diego City Schools Board of Education voiced the apprehensions of the minority community:

> We have a long way to go to convince the public that what we're doing is in the best interests of children. . . . When we talk about the issue of equity, the kind of assessments we're talking about require much more faith in individuals and the belief that people can actually apply equity in testing. Most of the time with a normed test you think of something that has some subjectivity in the development of the instrument, but then in the final result you know what the answer is. When you start talking about some of the assessments we're doing--portfolios--it's all subjective.[71]

Research on the effects of ethnicity, race, and gender on performance assessment is extremely limited. Most existing research has explored group differences on essay test scores only. Moreover, almost all the subjects in this research were college-bound students, limiting its generalizability considerably. Results of studies that examine the performance of women relative to men suggest that women perform somewhat better on essays than they do on multiple-choice examinations.[72]

Studies that report results for different minority groups are even more scarce. Results are mixed but tend to suggest that differences on multiple-choice tests do not disappear when essays are used. For example, data from NAEP indicate that black/white differences on essays assessing writing were about the same size as those observed on primarily multiple-choice tests of reading comprehension.[73] Similarly, adding a performance section to the California Bar Examination in 1984 did not reduce

---

[70] Robert Linn, Eva Baker, and Stephen Dunbar, "Complex, Performance-Based Assessment: Expectations and Validation Criteria," *Education/Researcher, in* press.

[71] Shirley Weber, remarks at Panasonic Partnerships Conference, Santa Fe, NM, June 1990, cited in Mitchell and Stempel, op. cit., footnote 2, California Assessment Program Case Study, p. 15.

[72] H.M. Breland and P.A. Griswold, *Group Comparison for Basic Skills Measures* (New York, NY: College Entrance Examination Board, 1981); Cooper, op. cit., footnote 61; S.B. Dunbar, "Comparability of Indirect Assessment of Writing Skill as Predictors of Writing Performance Across Demographic Groups,' unpublished manuscript, July 1991; Brent Bridgeman and Charles Lewis, "Predictive Validity of Advanced Placement Essay and Multiple Choice Examinations,' paper presented at the annual meeting of the Nationat Council on Measurement in Education, Chicago, IL, April 1991; and Traub and MacRury, op. cit., footnote 29.

[73] Cited in Linn et al., op. cit., footnote 70.

the difference in passing rates between blacks and whites. On the contrary, some studies have suggested that ethnic group differences actually increase with essay examinations.[74]

On the other hand, another study showed that minority college students in California actually performed better on tests that were direct measures of writing ability (the California State University and Colleges English Placement Test Essay Test or EPT) than on a multiple-choice test of English usage and sentence correction (the 50-question, multiple-choice formatted Test of Standard Written English or TSWE). In this study, score distributions on the TSWE and the EPT were similar for white students. Among African-American, Mexican-American, and Asian-American students, however, the two tests generated different score distributions. For these groups, the TSWE rendered a much more negative judgment of their English proficiency than the EPT.[75]

Policy Implications

Because of the limited research on the differing subgroup performance on new assessment instruments, Congress and other policy makers should approach these changes with caution. Data on the impacts of performance assessment on varying groups is needed in considering extension to more high-stakes applications. Careful planning, including representatives of groups traditionally negatively affected by testing, will be required in developing, administering, and scoring performance assessments for school accountability, student certification, or other selection purposes.

Role of Teachers and Teacher Training

In performance assessment, the role of the teacher in administering and scoring tests is much greater than with multiple-choice tests. Although some performance assessments still rely on outsiders to conduct the scoring of papers, in the future, classroom teachers are likely to have greater responsibility.

Although teachers observe performance all day, most have not been involved in defining and determining standards of performance common to those of their colleagues. In Sweden and several other countries a process called "moderation' refers to the development of a standardized scoring approach among multiple teacher readers. The procedure is similar to scoring of the Advanced Placement tests and other examinations relying on panels of scorers. It requires an intensive effort to agree on standards of performance. How does excellent work vary from that which is only fair or is not acceptable at all? This process is based on a shared understanding of curriculum, respect for teacher judgment, compromise, shared values, and a strong dose of common sense. This may be easier to manage in those countries where there is a common curriculum and a more homogeneous teaching population that has been prepared under a central system of teacher training institutions. It is not clear that this can be adopted in the U.S. system. One educator suggested: ''If we can trust our teachers to teach, we should be able to trust them to assess students. '[76]

Teachers in this country receive little formal training in assessment. A recent survey found that fewer than one-third of the States require new teachers to have demonstrated competence in educational measurement.[77] A survey of the six States in the Pacific Northwest reported that only Oregon explicitly requires assessment training for certification.[78]

One reason for the neglect of assessment training may be the assumption on the part of educators that the quality of assessments in the classroom is assured from outside the classroom; that is, most assessment is "teacher proof," beyond the control of the teacher.[79] Textbooks come with their *own*

[74] Breland and Griswold, op. cit., footnote 72; Dunbar, op. cit., footnote 72; Ina Mullis, "Use of Alternative Assessment in National Assessments: The American Experience, " paper presented at the Office of Educational Research and Instruction conference on the Promise and Peril of Alternative Assessment, Washington, DC, Oct. 30, 1990.

[75] Edward M. White and Leon L. Thomas, ''Racial Minorities and Writing Skills Assessment in the California State University and Colleges, ' *College English,* vol. 43, No. 3, March 1981, pp. 276-283.

[76] Jack Webber, teacher, Samantha Smith Elementary School, Redmond, WA, personal communication, 1991.

[77] ''Testing,'' *Education Week, vol.* 10, No. 27, Mar. 27, 1991, p. 9.

[78] Richard J. Stiggins, "Teacher Training in Assessment: Overcoming the Neglect,'' *Teacher Training in Assessment,* vol. *7 in the* Buros Nebraska Symposium in Measurement and Testing, Steven Wise (cd.) (New York, NY: L. Erlbaum Associates, in press).

[79] Ibid., p. 6.

worksheets and quizzes, unit tests, and even computerized test items, so teachers feel little responsibility for developing their own. Yet many of these text-embedded tests and quizzes are in fact developed in the absence of quality control standards. Furthermore, the tests that teachers know will be the **ultimate** judge of student proficiency are seen as beyond the teacher's responsibility. Finally, the courses on testing are often seen as irrelevant to the classroom.[80] There is very little treatment of assessment as a teaching tool. Teachers regularly use assessments to communicate achievement expectations to students, using assignments both as practice and as assessments of achievement, involving students in self and peer evaluation to take stock of their own learning with practice tests. This important area is neglected in teacher training.[81]

The inservice training situation is not much different.[82] However, if standard teacher courses in measurement are irrelevant, there is no reason to try to get more teacher candidates or practicing teachers to take them. On the other hand, if teachers are trained in new curriculum frameworks that have been the basis for much of the move to performance assessment, the techniques of teaching and assessing should be taught as a whole. This is the approach being taken in California, Arizona, and Vermont, and envisioned for Kentucky.

Technology can be a means to fast and efficient delivery of teacher training, as in Kentucky, where the educational television network provides satellite downlinks to every school in the State, making it possible to get the word out to all teachers simultaneously. And, if administrators are to understand the role of assessment in curricular change, and be able to communicate with the public about **school attainment of intended outcomes, they too need training in changing methods and goals of classroom and large-scale assessment.**

Policy Implications

If performance assessment is given a larger role in testing programs around the country, teachers will need to be involved in all aspects: designing tasks, administering and scoring tests, and placing test results into context. Teacher training will need to accompany these efforts. Redesigning the tests will not change teaching unless teachers are informed and involved in the process. The tests themselves could block educational progress unless classroom teachers are given a larger sense of responsibility for them.

Research and Development: Sharing Experience and Research

Performance assessment has been spurred primarily by State Departments of Education as they endeavor to develop tests that better reflect their particular curricula goals. Yet there are many common goals and concerns that have led them to come together to share experience with each other. In an effort to encourage the development of alternative methods of assessment, the U.S. Department of Education has supported the development of a State Alternative Assessment Exchange. The goal is to create a database of new forms of assessment, develop guidelines for evaluating new measures, and help prevent States from making costly mistakes. This collaborative effort, led by the Department's Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the Council of Chief State School Officers, is aimed at facilitating development work, not at creating a new test.

The National Science Foundation (NSF) has also played an important role in supporting research leading to new approaches to assessment in mathematics and the sciences. NSF supported NAEP in the development and pilot testing of hands-on assessment tasks in mathematics and science. Several of these tasks were adopted by the State of New York for their hands-on science skills test for fourth graders. More recently, NSF has committed $6 million for 3 years to support projects in alternative assessment approaches in mathematics and science.

---

[80] Ibid.

[81] Ibid., p. 8.

[82] There are some exceptions, however. For example, the Northwest Regional Educational Laboratory has created a video-based training program that places critical assessment competencies within reach of all teachers and administrators. They have also created "trainer-of-trainer" institutes that will make it possible for attendees to present to teachers and others a series of workshops on such topics as understanding the meaning and importance of high-quality classroom assessment; assessing writing proficiency, reading proficiency, and higher order thinking in the classroom; developing sound grading practices; understanding standardized tests; and designing paper-and-pencil assessments and assessments based on observation and judgment. Northwest Regional Educational Laboratory, *The Northwest Report* (Portland, OR: October 1990).

Assessment research remains a small part of the overall Department of Education research budget.

Greater effort should be directed toward monitoring the development of performance assessment and sharing information about models and techniques to facilitate implementation, prevent duplication of effort, and foster collaboration.[83]

Policy Implications

Because performance assessment is at a developmental stage, encouraging States and districts to pool experience and resources is an appropriate policy goal. Expanding research and comparing results requires a thoughtful atmosphere and adequate time. Although States are making progress in redesigning testing to serve educational goals, pressures for quick implementation of low-cost tests could present a barrier to this goal. Commitment to research projects and careful weighing of outcomes is essential to an improved testing environment.

Public Acceptance

One of the greatest problems with tests is the misuse of data derived from them. There is no reason to believe this would not also be true with performance assessment.

Because performance assessments aim to provide multiple measures of achievement, it may be difficult for parents, politicians, and school officials to understand its implications. The public has grown familiar with test results that rank and compare students and schools; it may be difficult to appreciate the information derived from tests that do not follow this model. Some attempts are being implemented to improve public understanding of the goals and products of performance assessment, through such vehicles as public meetings. But it is not easy. The press may be among the most difficult audiences to educate, since simple measures and statistics, ranking and ordering, and comparing and listing winners and losers makes news. Nevertheless, they may be the most important audience, since so much of the public's awareness of testing comes from press reports.

Policy Implications

Policymakers need to carefully consider the importance of keeping the public and press aware of the goals behind changing testing procedures and formats and the results that accrue from these tests. If not, there is a strong likelihood of misunderstanding and impatience that could affect the ability to proceed with long-term goals.

A Final Note

Writing assessment is up and ruining in many States. Although careful development is needed and issues of bias and fairness need attention, this technology is now workable for all three major testing functions.

Other methods of performance assessment (e.g., portfolios, exhibitions, experiments, and oral interviews) still represent relatively uncharted areas. Most educators who have worked with these techniques are optimistic about the potential they offer for at least two functions-testing in the classroom for monitoring and diagnosing student progress, and system monitoring through sampling. However, much research is needed before performance tasks can be used for high-stakes applications where students are selected for programs or opportunities, certified for competence, and placed in programs that may affect their educational or economic futures. Some of this research is now under way for tests used for professional certification (see ch. *8),* but much more research support is needed for understanding the implications in elementary and secondary schooling. Finally, even the most enthusiastic advocates of performance assessment recognize the importance of policies to guard against inappropriate uses. Without safeguards, any form of testing can be misused; if this were to happen with performance assessment, it could doom a promising educational innovation.

[83] Joe B. Hansen and Walter E. Hathaway, "A Survey of More Authentic Assessment Practices," paper presented at the National Council for Measurement in Education/National Association of Test Developers symposium, More Authentic Assessment: Theory and Practice, Chicago, IL, Apr. 4, 1991.