ED 389 778 UD 030 697

AUTHOR Seyfarth, John T.
TITLE Performance-Based Assessment: Questions and Answers.
INSTITUTION Metropolitan Educational Research Consortium, Richmond, VA.
PUB DATE Mar 93
NOTE 72p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Accountability; Comparative Analysis; *Educational Assessment; Educational Change; Educational Practices; Scores; State Programs; *Teacher Education; Test Construction; *Testing Programs; Test Reliability; *Test Use; Thinking Skills
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

Performance based assessment refers to tasks that require students to construct responses or take actions to demonstrate specific knowledge or skills. Performance assessment tasks appear in a variety of formats, but they focus on higher order skills and are nonroutine, and sometimes loosely structured, in nature. A number of concerns have been expressed about performance assessment, usually noting the difficulties of producing such assessments and implementing them on a large scale. Kentucky and California are among the states taking an early lead in using performance based evaluation. Advocates believe that performance assessment has the potential to change instructional practices in schools, and that as tests change, teacher practices will change. However, changes in teacher education will be necessary in order to adopt the successful use of performance assessments. Standardization will be necessary if the tests are to be used for accountability and comparisons. Special attention must be paid to the reliability of aggregated scores when such scores are used to represent the performance of classrooms, schools, and districts, as well as subgroups of students classified by race, ability, or sex. A four-item annotated bibliography contains some interesting resources for further reading. (Contains 74 references.) (SLD)

PERFORMANCE-BASED ASSESSMENT: QUESTIONS AND ANSWERS

METROPOLITAN EDUCATIONAL RESEARCH CONSORTIUM

MERC

CHESTERFIELD COUNTY PUBLIC SCHOOLS • COLONIAL HEIGHTS CITY SCHOOLS • HANOVER COUNTY PUBLIC SCHOOLS • HENRICO COUNTY PUBLIC SCHOOLS • HOPEWELL CITY PUBLIC SCHOOLS • POWHATAN COUNTY PUBLIC SCHOOLS • RICHMOND CITY PUBLIC SCHOOLS • VIRGINIA COMMONWEALTH UNIVERSITY

Metropolitan Educational Research Consortium

PO BOX 2020 • RICHMOND VA 23284-2020 • Phone: (804) 367-1332 •
FAX: (804) 367-1323 • E-MAIL: jpisapia@EDUNET.SOE.VCU.EDU

MERC MEMBERSHIP
John Pisapia, Director

CHESTERFIELD COUNTY PUBLIC SCHOOLS
Thomas R. Fulghum, Superintendent
Chairman, MERC Policy & Planning Council

COLONIAL HEIGHTS CITY SCHOOLS
Herman G. Bartlett, Jr., Superintendent

HANOVER COUNTY PUBLIC SCHOOLS
Stephen M. Baker, Superintendent

HENRICO COUNTY PUBLIC SCHOOLS
William C. Bosher, Jr., Superintendent

HOPEWELL CITY PUBLIC SCHOOLS
David C. Stuckwisch, Superintendent

POWHATAN COUNTY PUBLIC SCHOOLS
Margaret S. Meara, Superintendent

RICHMOND CITY PUBLIC SCHOOLS
Lucille M. Brown, Superintendent

VIRGINIA COMMONWEALTH UNIVERSITY
John S. Oehler, Dean
School of Education

Virginia Commonwealth University and the school divisions of Chesterfield, Colonial Heights, Hanover, Henrico, Hopewell and Richmond established the Metropolitan Educational Research Consortium (MERC) on August 29, 1991. The founding members created MERC to provide timely information to help resolve educational problems identified by practicing professional educators. MERC membership is open to all metropolitan-type school divisions. It currently provides services to 7,000 teachers and 120,000 students. MERC has base funding from its membership. Its study teams are composed of University investigators and practitioners from the membership.

MERC is organized to serve the interests of its members by providing tangible material support to enhance the practice of educational leadership and the improvement of teaching and learning in metropolitan educational settings. MERC's research and development agenda is built around four goals:

- To improve educational decision-making through joint development of practice-driven research questions, design and dissemination,
- To anticipate important educational issues and provide leadership in school improvement,
- To identify proven strategies for resolving instruction, management, policy and planning issues facing public education, and
- To enhance the dissemination of effective school practices.

In addition to conducting research as described above, MERC will conduct technical and issue seminars and publish reports and briefs on a variety of educational issues.

PERFORMANCE-BASED ASSESSMENT: QUESTIONS AND ANSWERS

Submitted by:

John T. Seyfarth
Virginia Commonwealth University
March 1993

*The views expressed in MERC publications are those of individual authors and not necessarily those of the Consortium or its members.

Executive Summary

A number of states, schools and school districts are engaged in efforts to develop and implement new forms of assessment. These efforts are known by several names. Performance assessment, the term used in this paper, refers to tasks that require students to construct responses or take actions to demonstrate specified knowledge or skills. Performance assessment tasks appear in a variety of formats ranging from open-ended questions to demonstrations of skill or comprehensive collections of bodies of work over time. The tasks focus on higher order skills, are nonroutine and are sometimes loosely structured. Students may be called upon to make, explain and defend assumptions, make predictions and estimates and explain connections and generalizations.

Performance assessment is usually proposed as a supplement to standardized multiple-choice tests, which have the advantages of low cost, objectivitity, ease in scoring, and requiring relatively little time to administer. Multiple-choice tests have been criticized for focusing teachers' attention on basic skills content at the expense of higher order outcomes and being based on outdated assumptions about learning. Performance assessment is considered to be superior to the multiple-choice format for measuring students' problem-solving strategies, ability to collect evidence and construct arguments, and skill at integrating previously-acquired knowledge to produce original insights or products. Proponents argue that teaching to the test, which is a concern when multiple-choice tests are used, is not an issue with performance assessment since it involves students' practicing essential knowledge and skills. Some experts disagree, however.

A number of concerns have been expressed about performance assessment. State testing directors have expressed reservations about the feasibility of implementing it quickly because of the high cost of development and the lack of necessary technical knowledge. On the bright side, the National Assessment of Educational Progress has recently developed scoring procedures for use with writing assessment that appear have acceptable levels of scoring reliability when scorers have had advance training.

One of the most ambitious efforts to implement performance assessment is taking place in Kentucky, where a statewide system is expected to be in operation by 1996. A state council is creating prototype tasks on which students may demonstrate achievement of specified higher-order outcomes, and teachers will later develop instructional strategies to help students successfully perform the prototype tasks. In effect, the assessment tasks become the new curriculum. The state will audit assessments to ensure uniformity in administration procedures and to safeguard accurate reporting.

Performance assessment is used most widely for writing. California pioneered the use of writing assessments by specifying nine genres of writing that students were expected to master. Teachers, aware that students might be tested on any of the nine, altered their teaching practices to place more stress on writing.

Efforts are now underway to develop assessment tasks for other subjects. The Mathematical Sciences Education Board and National Council of Teachers of Mathematics have developed model tasks in mathematics that test students' ability to estimate and to reason spatially and probabilistically. The Coalition of Essential Schools has developed model assessment tasks in economics and history. Assessments have also been developed in reading, science, art, foreign languages, music, geography, drafting, small engine repair, and employability skills.

Advocates believe that performance assessment has the potential to change instructional practices in schools. They suggest that the restrictive effects on instruction that result from use of multiple-choice tests come about because teachers attempt to prepare their students for the tests, and that when the tests are changed, teachers' instructional practices will also change. Another reason for optimism about the ability of performance assessment to change instruction is that many assessment measures model exemplary teaching practices. However, there is only limited evidence showing the effects of test results on teachers' instructional decisions. Additional research is needed to clarify the relationship between the two.

The Office of Technology Assessment of the U.S. Congress declared that the need for training for teachers is "critical" if performance assessment is to be successfully adopted in schools. Few teachers or administrators have received formal training in assessment. One area in which training is especially critical is in preparing teachers to score students' responses to performance assessment tasks. A recent Rand Corporation study of Vermont's statewide assessment program found that scorer reliability on writing portfolios was so low that the results were meaningless. However, some school systems in other states have attained satisfactory levels of scorer reliability on portfolios by employing a small number of highly trained raters.

Teachers and administrators who would like to use performance assessments in their schools are likely to find that none is available or that those that are available have not been standardized. Standardization is necessary if scores are to be used to compare schools or districts.

Aggregated scores are used to represent the performance of classrooms, schools, and districts, as well as of subgroups of students classified by ability, sex, race, and family income. Special attention must be paid to reliability in aggregating scores. A test that is highly reliable for individuals may be either low or high in reliability when scores are aggregated, depending upon the size of the aggregate and the variance in scores of the groups composing the aggregate. Higher aggregate reliability is achieved for tests with high individual reliability, for large groups, and for tests with high variance. However, a test with low individual reliability may have high reliability for an aggregate.Aggregating scores for the purpose of establishing "expected" scores for subgroups is not recommended since "expected" scores tend to become goals in teachers' minds. The focus of teachers' efforts should be instead on raising the performance levels of all students whose performance is low.

PERFORMANCE-BASED ASSESSMENT: QUESTIONS AND ANSWERS

Table of Contents

Preface

The Accountability/Assessment Study Group was charged by the Policy and Planning Council in May, 1992. The Study Group has explored a wide array of research agendas. These included program accreditation, outcome assessment programs, authentic assessment, the national standards movement, common core of learning and world class education assessment procedures. In addition the Study Group has prepared a concept paper to develop an authentic assessment research and development center to respond to requests for proposals from the Virginia Department of Education.

At its December meeting, the Policy and Planning Council requested that 1) a Review of the Literature, and 2) a brief survey of activities fostering authentic assessment in the seven school divisions be prepared. Professors Seyfarth and McMillan have collaborated on providing the information required to address the Council's request. They were aided by the Study Group's discussion and posing of the following focus questions to guide the Literature Review.

Focus Questions for the Literature Review

1. How do authentic assessment, performance assessment, and alternative assessment differ?
2. What are the pros and cons of performance assessment?
3. What strategies are being used by states, school divisions, schools, and teachers in the name of performance assessment?
4. In what subject areas are performance assessments being used?
5. How is performance assessment being implemented, and what are the most likely effects on teaching?
6. How do teachers use performance strategies to assess students?
7. How are teachers being trained in performance assessment?

8. When is it appropriate to use performance assessment, and when is it appropriate to use standardized multiple-choice tests?
9. What are the implications of performance assessment for norm-referenced tests?
10. How do we aggregate performance assessment data for policy and accountability purposes?
11. How do we insure that performance assessment data are reliable?
12. What are the implications for local assessment programs, such as end-of-course testing?
13. How does performance assessment relate to the Outcome Accountability Program in Virginia?
14. What are the implications of State Department of Education efforts in performance assessment for school divisions?

John Seyfarth is directing MERC's efforts in Performance-Based Assessment. He is advised by a study group consisting of Craig Organ and Lisa Kaufmann, Chesterfield County Public Schools; Mona Harrison, Colonial Heights City Public Schools; Carole Urbansok-Eads and Carol Stevens, Hanover County Public Schools; Sanford Snider, Henrico County Public Schools; Linda Hyslop, Hopewell City Public Schools; Doris Ownsby, Powhatan County Public Schools; and Joe Louis Simmons, Betty Williams and Richard Williams, Richmond City Public Schools. Susan Goins assisted the team and study group in meeting arrangements and document preparation.

John Pisapia
Director, MERC

QUESTION ONE

How do authentic assessment, performance assessment, and alternative assessment differ?

A number of states, schools and school districts have initiated efforts to develop and implement new forms of assessment for students (Darling-Hammond, 1991). These efforts are referred to by different names, the three most common being performance assessment, authentic assessment and alternative assessment. Although these terms are used loosely as synonyms, each carries unique connotations.

Performance assessment refers to testing that requires a student to perform tasks designed for the purpose of demonstrating specified knowledge or skills. These tasks may be activities in which the student solves a problem, identifies a malfunction, makes a decision, or implements a decision or solution (Tuckman, 1988). Performance assessment tasks require students to construct rather than select a response and requires them to focus on the process of problem solving rather than the solution itself. The criteria on which students are judged are clearly specified in advance (U.S. Congress, 1992).

Authentic assessment refers to evaluation activities in which students are expected to use previously-learned knowledge to produce integrated and original thoughts, objects, or performances that have meaning and significance independent of the assessment situation (Newmann & Archbald, 1992). Alternative assessment refers to evaluation activities that depart from prevailing methods of measuring achievement in favor of more direct measures of learning. For example, if students in school are expected to learn to write, speak, listen, produce artistic creations, find and cite evidence, and solve problems, then tests are designed to require students to do those things rather than answer multiple-choice questions about them (Wiggins, 1989). The term alternative assessment is used to emphasize the contrast between standardized multiple-choice tests and newer forms of assessment that emphasize demonstration of knowledge and skills in true-to-life situations or settings. The term performance assessment will be used in this paper except when reference is made specifically to some other form of assessment.

Several characteristics set performance assessment apart from other ways of measuring achievement. Performance assessment tasks appear in a variety of formats ranging from simple student-constructed responses to comprehensive demonstrations or collections of bodies of work over time (U.S. Congress, 1992). They assess behaviors of interest as directly as possible, and some are aimed at assessing cooperative rather than individual activities. Students understand the criteria on which their work is judged, and trained teachers or other qualified judges evaluate and score students' work. Results of performance assessments are usually treated as criterion-referenced data but may also be treated as norm-referenced if norms have been established (U.S. Congress, 1992).

Performance assessment yields a comprehensive account of an individual's or group's performance that goes beyond the testing of cognitive skills using pencil and paper techniques (Wood, 1984). It employs a variety of means to gather information about what students know and are able to do or about the quality of educational programs (Mitchell, 1992), frequently focusing on higher order skills ("Toward high standards," 1991). It is grounded in real-world contexts and presents nonroutine, open-ended and sometimes loosely structured tasks that require students to define a problem and determine a strategy for solving it, including decisions about what data are needed and how to collect, analyze and report them. Students may be called upon to make, explain and defend assumptions, predictions and estimates and to make connections and generalizations.

Among tasks that are considered examples of performance assessment are those that require analysis, investigation, experimentation, cooperation and written, oral, or graphic presentations of findings. Some activities call for group discussion and brainstorming and encourage students to monitor themselves ("Toward high standards," 1991; Darling-Hammond, 1991; J. Baron, quoted in Walker, 1992).

Examples of performance assessment activities are numerous in athletics and vocational subjects. Ratings of athletic performances such as skating, diving and gymnastics are examples of performance assessment, as is the evaluation of a student's diagnosis and repair of a malfunctioning automobile engine (Linn, Baker, & Dunbar, 1991). Other examples of performance assessment used in schools include exhibitions and demonstrations of speaking, debating, dance, and music; individual and team research projects; applied scientific experiments and projects; and field surveys and community involvement activities (Stefonek, 1991; Darling-Hammond, 1991).

QUESTION TWO

What are the pros and cons of performance assessment?

The principal advantage of performance assessment is its potential to bring about changes in curriculum and instruction that foster the development of important knowledge and skills in students. An instrument that is capable of inducing changes that facilitate students' learning the knowledge and skills the test is designed to measure possesses what has been called systemic validity (Frederiksen & Collins, 1989).

Standardized multiple-choice tests are criticized for "warping" the curriculum ("Toward high standards, 1991) and "corrupting teaching" (Darling-Hammond, 1991) by compelling teachers to teach to the test. Elaborate security precautions are in effect in most mandated testing programs to preclude teaching to the test, but those measures do not prevent teachers from narrowing coverage to focus on tested content, even though doing so means they must neglect more important outcomes ("Toward high standards," 1991; Rudman, 1987). It can be argued that teaching to the test is teachers' response to the way test results are used rather than a characteristic of the format, but many people believe that the multiple-choice format itself is inherently limiting.

Features of multiple-choice tests that are considered to be weaknesses are in most cases viewed as areas of strength for performance assessment, and vice-versa. Performance assessment measures student achievement of higher order outcomes, whereas multiple-choice tests drive instruction toward basic skills (Rothman, 1992c). Performance assessment tasks are not well-structured and do not have a single correct solution; they take more time to score, and the scoring is less objective and less reliable than multiple-choice items. On multiple-choice tests, the task is seldom ambiguous (although the response choices may be). These items usually have only one correct answer, and they can be scored quickly, reliably and objectively.

Teaching to the test is possible with performance assessment as well as with multiple-choice tests, but proponents of performance assessment assert that teaching to the test ceases to be an issue when it involves students' practicing essential knowledge and skills. Some of them argue that when tests measure intellectually challenging cognitive abilities that are the targets of instruction, teaching to the test should be encouraged (Wiggins, 1989; Nickerson, 1989; Shavelson, Carey & Webb, 1990). Other experts express reservations about that position, however (Mehrens, 1992).

Among the knowledge and skills that performance assessment measures that are not measured as well by multiple-choice tests are higher-order thinking skills and problem-solving strategies (Stefonek, 1991) and the ability to collect evidence, construct arguments, and take action (Wiggins, 1989). Another important characteristic of performance assessment tasks is that many of them require students to integrate previously-acquired knowledge to produce original insights or products (Archbald & Newmann, 1988).

Some features of performance assessment help increase student motivation. Motivating tasks involve activities that are meaningful, significant and worthwhile (Newmann & Archbald, 1992) or that require students to develop creative, novel and original approaches to ill-structured problems (Wiggins, 1989), all potential characteristics of performance assessment. Other features of performance assessment that help enhance student motivation are the nonroutine, open-ended nature of the problems they present and the fact that the problems are grounded in real-world contexts (Walker, 1992).

Performance assessment specifically challenges several assumptions that are implicitly made by many users of multiple-choice tests. One such assumption is that choosing a right answer from several alternatives is an adequate indicator of achievement (Wiggins, 1989). The belief that students' abilities are best demonstrated when they work alone is challenged by performance assessment tasks that require students to work in cooperation with peers (Linn, Baker & Dunbar, 1991).

The assumption that reasoning processes are transparent to multiple-choice responses is contradicted by some advocates of performance assessment, who claim that students may arrive at a correct answer on multiple-choice items by using incorrect reasoning processes or, conversely, may choose an incorrect answer using valid reasoning. Advocates also suggest that performance assessment provides information about students' ability to formulate problems, develop answers and think analytically, none of which is as likely to occur on multiple-choice tests (Nickerson, 1989; Shavelson, Carey, & Webb, 1990; Darling-Hammond, 1991).

Despite the many praiseworthy features of performance assessment just enumerated, a number of questions about it have been raised. It has been described as an untried methodology that may prove to be no better than current assessment methods (New York State Education Department, 1991). A survey of state assessment officers by Aschbacher (cited in Aschbacher & Herman, 1991) found that although the push for alternative assessments at the state level has been strong and about half of the states are involved at some level in alternative assessments, many testing directors expressed serious reservations about the feasibility of mounting successful new assessment programs. The directors' concerns were related to high costs of development, training, administration and scoring; the need for new and as-yet-undeveloped technical methods to assure reliability of scoring and validity of content; difficult and complicated logistical arrangements; and a lack of solid understanding and support for new assessment methods and the accompanying curricular and instructional innovations they require.

Performance measures that come closest to duplicating real life conditions are the most difficult and expensive to develop and implement on a large scale (Frechtling, 1991). It has been estimated that "truly authentic" assessment costs six to 20 times as much as current standardized testing programs (Sizer, 1992). U.S. News reports that performance assessment instruments being tested in Kentucky Schools cost four times as much as the standardized tests they are replacing and require much more teacher time ("The perfect school," 1993).

Much of the added expense of performance assessment is attributed to the cost of development and standardization, but some performance measures are also more costly and less convenient to administer. For subjects such as science, performance assessment usually requires the use of equipment that must be purchased, stored, set up, and dismantled, all of which increase costs. Administering performance assessment measures takes more time than is the rule for other forms of measurement, which increases opportunity costs and takes additional time away from instruction.

Teaching to the test is destined to become an issue for performance assessment, as it already is for multiple-choice tests, if students who receive advance knowledge of assessment tasks are enabled thereby to outperform peers who lack that knowledge. Since performance measures require more time to complete than multiple-choice items, fewer tasks are used, and advance knowledge will have a correspondingly greater effect on assessment results. Some authorities (Mehrens, 1992) suggest that many performance assessment tasks cannot be reused because of the possibility that advance knowledge will enable teachers to prepare students for the test. Developing all new performance measures each time assessment takes place adds considerably to cost of testing.

The time required to administer performance assessments is also a potential concern. Some performance assessment tasks involve activities that extend over several days or weeks (Frechtling, 1991; Darling-Hammond, 1991; "Toward high standards," 1991). For example, in completing a writing assessment, a student might be asked to prepare a rough draft one day and to edit and revise it the following day. Portfolios are collected over a period of weeks or even months. Many teachers believe that testing already takes too much time from instruction, and they are not likely to be pleased with the prospect of devoting even more time to it (Lockwood, 1991). A procedure that helps control costs and reduce the amount of time needed for testing is matrix sampling. However, matrix sampling has serious limitations, chief among them the fact that students do not receive individual scores (Bock & Zimowski, 1991).

Training costs add to the expense of performance assessment. Training is needed to prepare teachers to score students' responses and to prepare teachers and administrators to interpret and use data from performance assessments to improve instruction. Scoring responses to performance assessment measures is more complex than scoring multiple choice tests because the responses are usually longer and less uniform. Time must be allowed for graders to develop and maintain consistent scoring practices, which means they must reach consensus on standards (Gentile, 1992). Grading costs are further inflated by the need to use multiple graders to ensure reliability. Few teachers or administrators have had training in using data from performance assessments to make decisions about instruction; they must be trained to do that (Sherwood-Fabre, 1986).

The National Assessment of Educational Progress has recently developed scoring guides for use with three types of writing (informative, narrative and persuasive). Teachers were trained to use the guides and worked collaboratively in small groups. They scored a set of 10 papers and then compared and discussed their individual scores, repeating the process until group members felt confident that they could apply the guidelines consistently. About one-third of the scored papers were rescored as a reliability check. Reliability coefficients ranged from .76 to .89. Using a "percent adjacent agreement" score, the coefficients ranged from .96 to 1.00 (Gentile, 1992).

Computer technology is expected to help overcome some of the problems with performance assessment. Computers can be programmed to select performance tasks from item banks and produce a customized instrument for each student, thus simplifying the problem of test security and reducing the amount of student time required for assessment. It is also possible that software can be written that will emulate human scorers, thus reducing or eliminating the need for human scorers (U.S. Congress, 1992)

QUESTION THREE

What strategies are being used by states, school divisions, schools, and teachers in the name of performance assessment?

A variety of strategies are used to develop assessment tasks and procedures and to disseminate information about student performance. Groups involved in these efforts include individual schools, districts, states, universities, regional agencies, departments of the federal government and professional associations. Some of the efforts are described below.

States

The Connecticut State Department of Education has formed a consortium of teachers from six states (Vermont, Michigan, Minnesota, New York, Texas, and Wisconsin) and the Coalition of Essential Schools. Participating schools try out assessment tasks in order to determine whether they are ready for general use (Mitchell, 1992).

Kentucky is attempting to develop and install a statewide performance assessment system with the intent of influencing instruction. The newly-created Council on School Performance solicited opinions from Kentucky teachers and other residents and from curriculum experts to determine what students should know and be able to do after 12 years in school. Six outcomes were identified. Students will be expected to be able to (1) use basic communication and mathematical skills for purposes and situations they will encounter throughout their lives; (2) apply core concepts from mathematics, the sciences, the arts, the humanities, and social studies to situations they will encounter throughout their lives; (3) become self-sufficient individuals; (4) become responsible members of a family, work group, or community, and demonstrate effectiveness in community service; (5) think and solve problems in school situations and in a variety of situations they will encounter in life; (6) connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned, and build on past learning experiences to acquire new information from various sources (Foster, 1991).

The Council will create prototype tasks on which students will demonstrate achievement of the six outcomes. The tasks must have multiple objectives and require higher levels of thinking than is required on most paper and pencil tests. Teachers then will develop instructional strategies to help students successfully perform the prototype tasks. The current testing program will be replaced with an array of student performances that incorporate the six objectives. The intent of the program is to focus instruction on achieving specified outcomes rather than covering a certain amount of material. In effect, the assessment tasks become the new curriculum (Foster, 1991).

The state will audit assessments carried out in schools to ensure uniformity in assessment procedures and to ensure accuracy in reporting. Schools will be required to meet a threshold level of improvement established by the State Board of Education, which includes at a minimum an increase in the proportion of students who perform at benchmark levels, maintain a desired level of attendance, and remain in school. The State Board for Elementary and Secondary Education establishes the improvement threshold level for each school site every two years (Foster, 1991).

The Minnesota Statewide Assessment Program, which has been in use for 12 years, collects writing samples over a 3-day period to model exemplary writing process. Samples are scored at the state level, using local district personnel who have received trained in scoring. All districts are required to use this assessment for local program evaluation. A statewide performance assessment in science was conducted at three grade levels in Spring 1991. The assessment involved students in a variety of lab skills such as measuring, weighing, observing, and reporting. Additional statewide assessments, which will include a performance component, were planned for 1992 in visual arts and social studies. Math and music assessments may be conducted in 1993 ("Regional actions," 1991).

Minnesota maintains an electronic bank of traditional and non-traditional assessment and evaluation tools, techniques, and support materials called MIDEBANK. Videotapes on writing scoring and science performance are also available. An assessment and program evaluation manual is being completed ("Regional actions," 1991).

Maeroff (1991) described how a staff member from the State Department of Education in Rhode Island worked with students to evaluate and refine assessment tasks being developed for use in schools in that state. The representative talked with a group of third-grade students about a story they had read that told about a rooster who awakened people with his crowing. Among the factors the representative looked for were whether students possessed the background knowledge they needed in order to understand the story. In this case, a student who was not aware that roosters crow early in the morning would miss the point of the story. Although the purpose was not to evaluate individual students, the staff member took note of how long students took to read the story and whether individuals were able to cite information from the story to support their answers.

Alberta (Canada) is examining the possibility of using portfolios to assess student performance in math, science, social studies, and language arts. One pilot study involving four classrooms and 120 students has been completed. In that study, 21 third grade teachers read student portfolios and used holistic scoring guides to evaluate their work. The assessment showed that there was a need for teachers to devote more instructional time to the types of writing assignments used in the assessment and to helping students write self-evaluations of their work (Horvath, 1991).

The California curriculum frameworks effort has been cited as a model for a "systemic strategy" of curriculum revision. California used the National Council of Teachers of Mathematics standards as its base for creating a coherent curriculum in math and will use similar national reports as the basis for instruction in the sciences. Curricula will be organized around major ideas and themes, rather than unrelated bits of information, and new assessment methods will be aligned to the curriculum. Other efforts include pressuring publishers to produce better quality textbooks, making more extensive use of technology, and providing increased support for teachers ("Improving the math," 1991).

At least 40 states now require school districts to prepare and distribute information about student performance and school characteristics and resources to parents and the public. California was one of the first states to adopt these school report cards statewide. All districts in California report information about student progress toward meeting academic goals in reading, writing and mathematics; progress toward reducing dropout rates; estimated expenditures per student; progress toward reducing class size and teaching loads; quality and currency of textbooks and other instructional materials; availability of qualified personnel to provide student support services; adequacy of school facilities; adequacy of teacher evaluations and professional improvement opportunities; classroom discipline and climate for learning; curriculum improvement programs; and quality of school instruction and leadership (Chow & Matranga, 1990).

Other states which report on the schools' performance include Arizona, Connecticut, Maryland, Nevada, New York, and South Carolina (Brown, 1990). Nevada's report card includes information on educational goals and objectives; comparison of pupil achievement at each age and grade level for current and previous years; pupil-teacher ratios by grade level; teacher assignments and teacher qualifications and certification; expenditures per pupil for each source of funding; curriculum used by the district, including special programs; attendance, promotion and graduation rates; and information about efforts to communicate with parents (Chow & Matranga, 1990).

School Districts

The Littleton (CO) School District is a national leader in developing and implementing performance assessment in the schools. Fifth grade students at Mark Twain School in that district are required to write a paper on an approved topic and present it to a panel which includes the teacher and the principal. After the process was successfully implemented in fifth grade, lower grades also adopted it. The school has added a thinking-skills assessment which evaluates students' ability to determine probability and make predictions; a writing assessment, which measures student responses to a text; and a mathematics assessment which gauges students' measurement skills, use of fractions and creativity by asking them to design a playground. The school is pilot-testing a cultural literacy assessment for fifth graders and is redesigning a science assessment that will allow students to conduct experiments and analyze data. Students' assessments, along with other information about their school performance, are entered into a portfolio which parents can see (Rothman, 1992b).

By the end of 1992, Littleton High School was expected to have produced a set of demonstration assessment tasks, along with standards for proficiency and excellence, for the 19 competencies in which students are expected to be proficient. For example, a communications task asks students to write a letter to a public policy-maker about the official's position on a current political issue. A proficient letter is one in which the writer states an opinion and uses facts to support it; uses appropriate format and wording and makes few grammatical mistakes. An excellent letter is one in which the writer explains the writer's position and refutes an alternative position, uses interesting vocabulary and varied sentence structures, and makes essentially no errors (Rothman, 1992b).

The assessment program in Dallas schools uses both norm-referenced and criterion-referenced tests. Criterion-referenced tests tied to curriculum objectives in mathematics and communications are available for teachers to use in their own classrooms. Teachers may use part or all of any test, and a scoring and reporting service is provided (Grobe, Adkins, Arrasmith, & Sheehan, 1983).

Many school districts have made progress in recent years in aligning the curriculum with tests used to measure student performance. There is evidence to suggest that when the curriculum, instruction and assessment are all aligned, student performance improves (U.S. Congress, 1992). Alignment occurs when teachers and administrators take action to adjust policies and practices to achieve uniformity among the curriculum, instruction and testing. Even in districts with clearly stated policies in those areas, alignment may fail because of differences in beliefs about assessment among key administrators.

In a district cited by Williams and Bank (1981), the superintendent advocated using test results to identify topics which students had not mastered, so they could be retaught. The assistant superintendent in the same district believed test scores should be used to help identify areas in which staff development for teachers was needed. Principals in the district used information from the tests for planning, and most teachers reported they used the data for grouping students and diagnosing their progress.

Regional Agencies

The Council of Chief State School Officers (CCSSO) sponsors the Student Assessment Consortium (SAC). Six member groups of SAC are working on assessment in arts, social science, mathematics, writing, reading, and work readiness. SAC provides the means for people working on innovative assessment to discuss what is happening in their states. Among the activities planned for the near future is development of prototype assessment exercises. For more information about SAC, contact Edward Roeber, Director, Student Assessment Programs, CCSSO, One Massachusetts Ave NW Suite 700, Washington DC 20001 (202/408-5505; FAX 202/408-8072). CCSSO also sponsors the Student State Collaborative on Assessment and Student Standards (SCASS). States are invited to participate in SCASS on a project-by-project basis ("Survey of national," 1992).

The New Standards Project, sponsored by the National Center on Education and the Economy (NCEE) and the University of Pittsburgh Learning Research and Development Center, involves 17 states, including Virginia, and 6 school districts. Its purpose is to develop national standards and performance-based assessments in math, science, English, language arts, and history. The Project plans to pilot test performance tasks for fourth grade English and math. For more information contact Warren Simmons, NCEE, 1329 18th St. NW. Suite 401, Washington DC 20036 (202/783-3668; FAX: 202/783-3672)

("Survey of national," 1992). One project already carried out by the New Standards Project involved calibrating scoring of writing exeıcises from ten states (Roeber, 1991).

Federal Government

The National Assessment of Educational Progress (NAEP) sponsors a trial state assessment which conducts voluntary state-by-state assessments. National and state samples of students ages 9, 13 and 17 are assessed biannually in several subject areas, and teachers and principals are questioned about programs, policies and their own background and training. In 1994 NAEP will report on the 1992 assessments and will conduct assessments in reading, math, science, U.S. history and geography. For information, contact Stephen Koffler, director of test development and state services, NAEP, 30-E, Educational Testing Service, Rosedale Road, Princeton, NJ 08541 (609/734-1427; FAX 609/734-1878) ("Survey of national," 1992).

The Secretary's Commission on Achieving Necessary Skills (SCANS) was sponsored by U.S. Department of Labor and composed of representatives of schools, businesses, unions and government. Its purpose was to examine workplace demands and determine whether students were capable of meeting them. The Commission sought to define skills needed for employment; propose acceptable levels of proficiency; suggest ways to assess proficiency; and develop dissemination strategies. Six panels were established to examine jobs ranging from manufacturing to government employment. The Commission issued a report entitled "What Work Requires of Schools: A SCANS Report for America 2000" ("Survey of national," 1992).

QUESTION FOUR

In what subject areas are performance assessments being used?

Thirty-six states currently use mandated or voluntary assessments to measure student performance in writing, and nine others have plans to introduce writing assessments in the near future. Twenty-one states currently use performance assessments in other subjects, and 19 states are planning to introduce them (U.S. Congress, 1992).

Illinois administers statewide process assessments in reading and writing to students in grades 3, 6, 8 and 11. The reading assessment makes use of full-length passages from texts. Some questions on the texts are multiple choice items for which as many as three choices may be correct. Indiana plans to assess student performance in mathematics, language arts, social studies, and science ("Regional actions," 1991).

Minnesota's Statewide Assessment Program, assesses students' writing using an exemplary three-day writing process. All districts are required to use the assessment for local program evaluation. A statewide performance assessment in science was conducted at three grade levels in Minnesota schools in Spring 1991. The assessment involved students in a variety of lab skills such as measuring, weighing, observing, and reporting. Statewide assessments, which will include a performance component, were planned for 1992 in visual arts and social studies. Math and music assessments may be conducted in 1993. Minnesota maintains an electronic bank of traditional and non-traditional assessment and evaluation tools and support materials for use by schools in the state ("Regional actions," 1991).

New York administers a "hands-on" science assessment to fourth-grade students. Maryland has assessment initiatives planned or underway in reading, writing, language usage, mathematics, science and social studies (Stefonek, 1991).

An employability skills portfolio is planned for use as part of the Michigan Educational Assessment Program. Indiana also plans to assess life and employability skills ("Regional actions," 1991).

Connecticut has implemented a range of performance-based assessments in science, foreign languages, drafting, and small-engine repair (Wiggins, 1989). Rhode Island is preparing assessments in reading, writing, speaking, listening, and mathematics for third grade students (Maeroff, 1991).

Member groups of the Student Assessment Consortium (SAC), sponsored by the Council of Chief State School Officers, are working on assessment tasks in art, social science, mathematics, writing, reading, and work readiness ("Survey of national," 1992).

Stiggins (1991) identified four categories of achievement targets--subject-matter knowledge, thinking skills, behaviors, and products. Performance tasks currently used by various states for two of these outcomes (thinking skills and products) will be reviewed. All of the tasks described involve subject-matter knowledge, and several of them involve student behaviors.

Thinking Skills

Some tasks are designed to promote thought and provide feedback to teachers about students' processes of reasoning. Among these activities are tasks designed by the Mathematical Sciences Education Board (MSEB) which appeared in a recently released report. The Board pointed out that the tasks contained in the report were not to be viewed as an assessment instrument in themselves, since important areas of the math curriculum were not covered. It suggested that the tasks be viewed as models of activities that might be included as part of a mathematics assessment (Rothman, 1992d).

One of the math tasks recommended for fourth grade students and included in the MSEB report showed a diagram with names of six children. Pairs of names were connected by line segments, with an arrowhead pointing toward one of the names. The diagram represented results of games played in a checkers tournament, with the direction of the arrow indicating which member of the pair had won the game. For example, a segment connecting the names Jose and Alex with the arrowhead pointing toward Alex's name meant that Jose had defeated Alex. Students used the diagram to answer questions such as these:

* Who won the game between Pat and Robin?
* How many games have been played so far?
* Make a table showing current standings of the six children, putting the player who has won the most games in first place at the top.
* The tournament will be over when everybody has played everybody else exactly once; how many more games need to be played to finish the tournament? (Explain.)
* David and Lee have not played yet; who do you think will win when they play? (Explain.)

The arithmetic operations involved in answering these questions are simple, but fairly sophisticated reasoning is required. For example, making a table showing each child's record in the tournament requires students to aggregate information from the table in order to produce rankings, and determining the total number of games in the tournament involves the notion of combinations. In predicting whether David or Lee will win an upcoming game, students must refer to the diagram to determine probabilities and use that information to project a winner in the match (Rothman, 1992d)

Another sample task from the MSEB publication tests students' skill at spatial reasoning. A drawing shows the p^sitions of three stationary objects (merry-go-round, fort, and beach umbrella) on an island in the middle of a pond. A boy on a sailboat is pictured on the pond, and students are asked to decide how the three objects appear to the boy as he moves his boat about on the pond (Rothman, 1992d).

The National Council of Teachers of Mathematics (NCTM) suggested this prototype of an assessment exercise for grades K-4: Students are shown a large box of raisins and asked to estimate--not count--the number of raisins. They are provided with a balance, containers of different sizes, and a calculator. They use one method to make an initial estimate and a second to check it (Mitchell, 1992).

Another sample task proposed by NCTM asks students to find the average of a set of scores and then tell how much the average would increase if each student's score were increased by 1, 5, or 8 points. Students are also asked to write a statement telling how much the average would increase if X points were added to each score and write an argument to persuade another student that their reasoning is correct. A problem proposed for high school seniors asks whether a boy is correct in assuming that he is certain of being admitted to one of two colleges, provided he applies to both and knows that in the past one-half of the students from his school who applied at each college were accepted (Mitchell, 1992).

Fourth-grade students in New York State schools take a science test that involves hands-on activities. The students move from station to station, answering questions at each station that require manipulation of physical objects or logical thinking. At one station, students try to complete a circuit and cause a bulb to light. At another, they invent two categories and sort a variety of vegetables into the categories. The vegetables are lima beans, kidney beans, pinto beans, peas and corn (Mitchell, 1992).

A task that measures student knowledge of geography uses a map showing two cities along the shore of a large lake located in the middle of a continent at 40 degrees north latitude. Students are asked, "How will the presence of the lake affect the climate of each of the cities at different times of the year and at different times of day?" One of the cities is located on the east shore of the lake, and the other is situated on the southwest shore. A rivei flows near the city on the east shore of the lake. Students' responses are scored on the basis of recognition of two principles (winds in middle latitudes are prevailing westerlies, and large bodies of water warm and cool more slowly than land) and awareness of several seasonal effects (the city situated on the eastern shore will have cool springs and warm autumns and will be subject to lake-effect snow; under certain conditions both cities will experience land-to-lake breezes) (Bock, 1991).

Products

Many assessment tasks require students to construct products, including portfolios consisting of a collection of writing or artwork, utilitarian objects, or designs. One example of an assessment exercise that involves a product is the oral history project for ninth grade students designed by the Coalition of Essential Schools. Students create three workable hypotheses based on preliminary investigations and write four questions to test each hypothesis. They then obtain answers to their questions by interviewing and reading.

The finished product is evaluated using the following criteria: Student investigated three hypotheses; demonstrated that he/she had done background research; prepared at least four questions related to each hypothesis; asked questions that were not leading or biased; asked follow-up questions when appropriate; noted differences between fact and opinions in answers; used evidence to support his/her choice of the best hypothesis; and organized the writing and class presentation (Wiggins, 1989).

The Coalition also has developed a course-ending simulation exam in economics that has characteristics of a performance assessment. The instructions for this activity read: "You are CEO of an established firm. Your firm has always captured a major share of the market, because of good use of technology, understanding of market systems, and maintenance of a high standard for your product. In recent months new firms have entered the market and have captured part of your sales. Your board of directors has given you less than a month to prepare a report that solves the problem in the short run and in the long run. In preparing the report, you should define the problem, prepare data to illustrate the current situation, prepare data to illustrate conditions one year in the future, recommend action for today, recommend action over the next year, and discuss where your company will be in the market six months and one year from today" (Wiggins, 1989).

Tasks that must be completed in the course of completing the project include: deriving formulas for supply, demand, elasticity, and equilibrium; preparing schedules for supply, demand, costs and revenues; preparing graphs; developing a written evaluation of the current and future situation for the market in general and for the student's company in particular; preparing a written recommendation for the board; showing aggregate demand today and predicting what it will be one year hence; showing the demand for the firm's product today and predicting what it will be one year hence (Wiggins, 1989).

In Connecticut, foreign language students are asked to draft a letter to a pen pal in the target language. The letter is rated at one of four levels of proficiency (novice, intermediate, intermediate high, and advanced), based on such things as knowledge of vocabulary, creativity, mastery of idioms, and correct use of tense, syntax and sentence structure (Wiggins, 1989).

The Written Language Assessment, a commercial test prepared by Grill and Kerwin, is a standardized test designed to evaluate written language competence. Students write three essays over a period of 2-5 days using three types of writing (expressive, instructive, and creative). Essays are scored quantitatively and qualitatively. Qualitative scores are based on ratings of rhetorical skill, legibility, and overall quality. For each essay, three qualitative ratings are assigned (rhetoric, legibility, overall quality). These are combined into a single quality score, termed general writing ability (GWA). Three quantitative measures are also calculated. A simple word count is used as an indicator of productivity. Word complexity is calculated by subtracting total number of syllables from total number of words. Readability level is calculated using a formula adapted from Fry's principles. All of the scores (quantitative and qualitative) are combined into the Written Language Quotient (WLQ) score (Partridge, 1990).

The Illinois Writing Program was designed to integrate instruction and assessment. To accomplish that objective, students' writing is evaluated on focus, support/elaboration, organization, and conventions. The assessment also produces a holistic score (Integration) that reflects how well the composition as a whole accomplishes the objective of the assignment. Assessment results are presented as a score profile which teachers use to identify areas of instructional need (Chapman, 1990).

The Illinois program provides an information network through which teachers and other personnel can obtain help with using test results to improve instruction. Five years after the program began, more than 1000 teachers had been trained to assess students' writing. Elementary teachers were especially positive about the program, because most of them had had no previous training in teaching writing (Chapman, 1990).

At Mark Twain School, Littleton (CO), fifth grade students conduct a research project that involves composing two questions and collecting information to answer one of them. Teachers help the students to revise their questions so they will be clear and researchable. The teacher then chooses one of the questions, and the student must obtain information to answer it. The work is carried out over four days, and students spend about one-half their time during that period completing the assignment. They produce both a written paper and an oral presentation that must include a visual display. Among the topics previously researched by fifth-graders at the school are causes of World War II, training seeing-eye dogs, global warming, cancer treatments, and Albert Einstein (Lockwood, 1991; Rothman, 1992b).

A task force consisting of English teachers in California schools and headed by a professor from the University of California at San Diego began in 1984 to identify types of writing students should be able to produce after receiving instruction in the new English curriculum. Genres of writing included narrative or story, report of information, evaluation or judgment, autobiographical incident, solution of a problem, speculation about cause-effect, report of observations, interpretation, and discussion of a controversial issue. These were selected because they were the kinds of writing required in post-secondary education or in the world of work. The group prepared prompts for grades 8 and 12 writing assessments. The prompts specified a scenario, an audience and a purpose. Since students might be asked to write in any genre, teachers had to teach all genres in order to prepare students for the assessment (Mitchell, 1992).

In the Arizona Student Assessment Program students spend the first day writing a rough draft, and on Day 2 they revise and edit the first draft using a checklist that is provided to them. Reading and writing assessments in Arizona are matched in order to promote integrated learning. For example, on the reading assessment eighth grade students read a narrative about Helen Keller and write answers to three questions. One asks students whether what they have read helped them to understand disability. The second question asks students to suggest what more they might learn from reading Keller's autobiography. The third question asks students to tell why they think the author of "The Miracle Worker" chose that title for his play (Mitchell, 1992).

A task used in the Maryland School Performance Assessment Program requires students to prepare plans to open a restaurant. Students work in groups to construct a questionnaire and to plan and carry out a market survey to obtain information such as the type of restaurant that might do well. The next step in the activity involves selection of a site for the restaurant. Students use geometry skills to select one of three lots that best fits their needs, and they then produce a scale drawing showing how a 6000 square foot building and a 30-car parking lot will be situated on the lot. They also calculate the cost of the building, parking lot, and restaurant equipment. Finally, the students prepare a written summary of their decisions; the summary accompanies the drawing and other display materials in a presentation to a zoning board (Mitchell, 1992).

The most common use of performance assessment in education is for preparation of portfolios. Portfolios serve many purposes. Their major advantage is that they can be designed to function simultaneously both as teaching tools and as vehicles for assessment. As a teaching tool, they are used to provide student ownership and a sense of accomplishment, to involve students in self-evaluation, to help students and teachers set goals, to individualize writing instruction, to connect reading, writing, and thinking, and to aid in parent conferences. As assessment instruments, they are used as an alternative to standardized testing, for program evaluation purposes, or serve as an end-of-year culminating activity (Mitchell, 1992).

It has been recommended that portfolios convey explicitly or implicitly students' activities, including the rationale or purpose for creating the portfolio, goals or outcomes, the content or displays to be included, standards (how to judge the quality of the contents), and judgments or what the contents reveal about a student's performance (Paulson, Paulson & Meyer, 1991).

Portfolios are used most often for writing, but they are adaptable for use in almost any subject. They have been used in English/language arts, creative writing, art, and mathematics, and science portfolios are now beginning to appear. Based on the stated purpose, pieces to be included in a portfolio are selected by the student and teacher working together (Mitchell, 1992).

QUESTION FIVE

How is performance assessment being implemented, and what are the most likely effects on teaching?

Performance assessment contributes to change in instructional practice by modeling for teachers exercises and activities that exemplify the kinds of outcomes schools seek to accomplish. Change in teaching strategies been most apparent in the teaching of writing. Since the California Assessment Program (CAP) was introduced, students have been expected to be familiar with nine genres of writing (narrative, information, evaluation or judgment, autobiographical, solution of a problem, cause-effect, observation, interpretation, and discussion of a controversial issue). These genres were selected because they are representative of the types of writing required in higher education and the world of work. Since teachers do not know in advance which genre will be included on the assessment, they must prepare students to write all nine. An evaluation of the CAP found that 78 percent of teachers had increased the number of writing assignments for their students and 94 percent had assigned a greater variety of writing tasks after the assessment program went into effect, as compared to their practices prior to that time (Mitchell, 1992).

In Arizona, the Student Assessment Program (ASAP) models the process approach to teaching writing. The ASAP is a two-day activity. On the first day, students produce a rough draft, and on the second day they revise and edit the draft using a checklist. Reading is integrated with writing. Students are asked to read a lengthy excerpt and write one-paragraph answers to questions about it. Although the purpose of the exercise is to test students' reading ability, they are encouraged to review their written responses for sentence structure and use of correct spelling, punctuation, and capitalization. No data are available showing whether the assessment has had an effect on teaching practices (Mitchell, 1992).

New York State has introduced a performance assessment tasks that requires elementary school students to demonstrate their understanding of scientific concepts using hands-on activities that have the potential to change the way science is taught in elementary schools. On one of the assessment tasks, students are required to complete a batte y-powered circuit, and another requires them to sort several types of vegetables into two categories. The New York State Elementary Science Program Evaluation Test (ESPET) was developed by teachers and university faculty members without involvement from outside vendors or contractors (Mitchell, 1992).

Reading assessments now in use in schools in Illinois, Michigan and Wisconsin make use of longer and more complex passages and pose questions that require students to exercise inferential reasoning processes (Stefonek, 1991).

It is difficult to determine with accuracy what other areas of teaching practice are likely to be affected by performance assessment until more is known about the relationship between teachers' decisions regarding allocation of time and organization of the classroom and assessment information obtained by the teachers (Stiggins, Conklin & Bridgeford, 1986). However, several writers have speculated about the eventual effects of performance assessment on teaching. Some of their thoughts are discussed below.

There is general agreement that, as long as teachers are held accountable for their students' performance on tests, they will emphasize in their teaching the knowledge and skills that are measured on those tests. This, of course, is the source of the complaint that existing testing programs corrupt teaching. However, if assessment instruments that measure higher order thinking and problem solving skills are mandated in schools, teachers can be expected to turn their attention to helping students acquire those capacities, with the effect of increasing the amount of time spent on objectives that are educationally valid (Nickerson, 1989).

Teachers can be expected to use information from portfolios and other forms of assessment to tailor instruction to students' needs (Simmons, 1990). One of the areas in which performance assessment promises to be especially useful is teaching reasoning skills. Students use a variety of approaches to solve problems, and their reasoning processes are seldom obvious to teachers with tests currently in use. Cognitive performance strategies make reasoning processes accessible to teachers and permit them to model alternative reasoning strategies for students' use (Siegler, 1989).

When performance assessment becomes more widely used in schools, teachers' roles may change; teachers are likely to become more like coaches, focusing their attention on improving student learning rather than simply emphasizing coverage of content (Lockwood, 1991). Teachers can be expected to begin to play a more active part in developing and scoring assessments and in guiding students in the selection of work to be included in portfolios (Darling-Hammond, 1991), and they will probably spend more time discussing with colleagues what students should know and be able to do (Rothman, 1992b). Performance assessment is also likely to lead to emergence of new standards of performance, based on the extent to which students' work matches some exemplary or expert model of competence (Aschbacher & Herman, 1991).

Performance assessment will cause teachers to reassess their views of students' ability to learn (Aschbacher & Herman, 1991) and to take a broader view of achievement by de-emphasizing "right" answers and by emphasizing skills (such as evaluation of one's work) that are now seldom taught (Aschbacher & Herman, 1991). Teachers will spend more time teaching students to think, and instructional methods that have been pushed aside by the emphasis on basic skills instruction will reappear in classrooms. Among these techniques are student-centered discussions, writing essays or themes, research projects, and laboratory work (Darling-Hammond, 1991). To prepare students for performance assessments, teachers will change the emphasis in classwork in order to give students an opportunity to practice skills that are required on the performance assessment or that are to be demonstrated in portfolios (Maeroff, 1991).

It is easier to implement performance assessment in skills subjects than in content-laden courses. Performance assessment is used widely for evaluating writing and promises to have wide applicability in reading and foreign language instruction. All three subjects are skill-intensive. Developing performance assessment tasks that adequately measure student performance in such subjects as history and science is more difficult (Wiggins, 1991a).

Performance assessment promises to remove the wall of separation between teaching and testing. In some areas of learning, including medicine, psychology and social work, teaching and assessment merge, but in elementary and secondary schools, teachers think of them as distinct and separate activities. The widespread introduction of performance assessment into schools can be expected to cause teachers to re-examine their beliefs about the nature and relationships of instruction and assessment (Maeroff, 1991).

Some educators believe that students with above average ability have been neglected in the schools in recent years as more attention has been given to low achievers. Results from the Scholastic Aptitude Test (SAT) show that the average SAT score nationally declined 10 points between 1976 and 1983. However, students who scored in the top 10 percent among those taking the test had an average decline of 26 points, as compared to an increase of 7 points in the average score of students whose scores fell in the lowest 40 percent of the test-taking population (Rudman, 1987). These results suggest that instruction for low-achieving students improved during the period while instruction for high achievers deteriorated.

The SAT results must be interpreted with caution, since the test-taking group changes each year, but the results are not inconsistent with evidence from other sources which shows that in recent years schools have increasingly focused on the needs of low-achievers. Some states are now beginning to re-examine their priorities in order to attain a better balance in targeting instructional resources to children at all ability levels. There is an expectation among some educators that the introduction of performance assessment into schools will encourage teachers to shift instruction to areas in which students with above average ability are expected to do well.

QUESTION SIX

How do teachers use performance strategies to assess students?

It has been estimated that teachers spend between 20 and 30 percent of their professional time directly involved in assessment-related activities. This includes time for designing, developing, selecting, administering, scoring, recording, reporting, evaluating, and revising daily assignments, tests, quizzes, observations and judgments about student performance, and oral question and answer sessions. By and large teachers prefer to develop and use their own assessments, and have less faith in information derived from assessments that they have not developed themselves or selected as directly applicable to their classes (Stiggins, 1988).

Teachers have a high degree of confidence in their own assessment tools, but their faith may be misplaced. The small number of studies that have been carried out on the topic revealed significant deficiencies in teacher-made tests. Researchers found that these tests tended to be brief and that they contained a preponderance of items at the recall or comprehension levels of cognition. Researchers also reported that few teachers analyze individual test items for reliability or clarity, that little time is spent improving tests before they are reused, and that most teachers are not adept at setting criterion levels for student performance (Stiggins, Conklin, & Bridgeford, 1986).

It was estimated in one study that about three-fourths of all teachers use performance assessment measures with their students. About half of the measurement activities they described involved evaluation of processes, and most of the remainder involved assessment of products created by students (Stiggins, Conklin, & Bridgeford, 1986).

A promising way of improving the quality of instruction in schools is to encourage teachers to make greater use of performance assessment tasks in their teaching. Well

-designed assessment measures are effective both as instructional tools and for tracking student achievement. In the paragraphs that follow, several examples of performance assessment tools that are recommend for use for instructional purposes are described.

Paragraph frames

Some teachers use paragraph frames to model ways by which students can organize and report information they have learned in their content reading. Paragraph frames help ease the transition from narrative texts used in primary grades to expository reading and writing which students face in intermediate grades. Paragraph frames provide an outline which the student completes by filling in the blanks. The following is an example of a sequentially organized frame: "Before a frog is grown, it goes through many changes. First, the mother frog ___. Next, ___. Then, ___. Finally, ___. Now they ___" (Partridge, 1990).

This is an example of an enumeration frame: "Bats are unusual animals for several reasons. First, ___. Second, ___. Third, ___. Finally, ___. As you can see, bats are unique in the animal world." The following reaction frame models for students a way of blending previously-learned knowledge with new information gained from reading: "Although I already knew that stars are hct, I learned some new things about stars by reading this story. I learned that ___. I also learned that ___. However, the most interesting thing I learned was ___" (Partridge, 1990).

A variation of the reaction frame illustrates a method by which a student can report how his/her ideas changed as a result new information acquired from reading: "Before I read this selection, I thought ___. After ___. Besides this, I learned some other new facts about bats. First, ___" (Partridge, 1990).

Scenarios

A teacher at Urbandale High School in Des Moines (IA) uses a scenario about the atomic bomb in a social studies class. She gives students information about dropping the atomic bomb on Hiroshima and Nagasaki from several points of view. Each student is assigned a role (scientist, pilot, peace activist, President, etc). A hearing is held, and students are expected to be prepared to answer questions as if appearing before a committee of the U.S. Senate. The teacher reports that the activity increases students' motivation to learn (Lockwood, 1991).

Judging readiness

First grade teachers use performance assessment tasks to assess students' for reading and to record students' progress in recognizing letter sounds. For example, a first grade teacher in South Brunswick (NJ) who is assessing a student's reading readiness says to the child: "I'm going to read you this story but I want you to help me. Where do I start to read?" She records a check if the child knows to begin reading at the top left of the page, another check if the child moves his/her finger from left to right, and a third for going page by page (U.S. Congress, 1992).

The first grade teacher asks another child to spell truck, dress, and feet. When a child is asked to do this early in the school year, he/she strings together a random series of letters whose sounds have no relationship to the sounds of the words. Later in the year, students begin to use primitive forms of phonetic spelling ("t-r-k" for truck, "j-r-s" for dress, and "f-e-t" for feet). The teacher spends from 2 to 10 minutes assessing each child's readiness and then files the results in the child's portfolio (U.S. Congress, 1992).

Interpreting graphs

Romberg (1992) gives an example of a performance task that is used by teachers in Great Britain to assess students' ability to interpret information presented in visual forms. The child sees a map with a highway connecting the cities of Nottingham, London and Crawley. No scale is given, but the distance from Nottingham to London appears to be 3 or 4 times as long as the section between London and Crawley.

Both Nottingham and Crawley are relatively small cities located near the highway. London is much larger, and the highway passes through the city. Also shown is a graph which depicts segments of the trip, represented by letter names AB, BC, CD, and so on. (The letters appear on the graph but not on the map.) For example, AB is a segment that take one hour travel time and covers 60 miles; the segment BC takes 30 minutes and covers 0 miles; and CD takes almost 1 hour and covers 50 miles. All of the segments except DE are straight lines, indicating steady speeds. DE is curvy, indicating frequent changes in speed.

The student is instructed to describe each stage of the journey by synthesizing information from both the graph and the map and to describe and explain what is happening along each leg of the trip. On the first segment (AB) the driver is driving from Nottingham toward London, as indicated by the amount of time required and the constant rate of speed. For segment BC, 30 minutes elapse but no distance is covered, indicating a stop. Segment DE is recognized as the London leg because of the frequent variations in speed (Swan, cited in Romberg, 1992).

Teaching discourse

In social studies, teachers expect students to learn to use previously acquired information to develop a narrative, make a judgment, prepare an argument or explanation or analyze data or an argument. These products are referred to as discourse, and to qualify as discourse, a student's answer must go beyond the reproducing statements previously given by a teacher or authors of textbooks. Discourse requires that the student be able to integrate relevant knowledge of a field into his/her own language in response to novel problems (Newmann, 1992).

This passage is a hypothetical example of student discourse in social studies: "The Boston Tea Party before the American Revolution was a violent protest against the many restrictions on the colonists' trade imposed by the English Parliament. Although it was 'violent,' it really didn't hurt people--physically. It did do violence to property rights, because all that tea was destroyed by people who didn't own it. Still, I think the protest was justified, because the colonists had taken just about all the peaceful means they could to try to get Parliament to revise some very unjust laws" (Newmann, 1992).

Newmann (1992) suggests several sample questions that can be used by teachers to assess student discourse in social studies:

(1) The eighteenth century is considered one of the most important periods in the history of the U.S. Discuss some reasons why that century is considered so important. In expressing your ideas, describe key events, people, general trends, and ideas that seemed to take hold of people, and explain why these would be considered important.

(2) Joe is 25 years old and out of work. He graduated from high school and has worked most of his life in construction, but he lives in a region where construction workers have been laid off by the thousands. He asks why there aren't more jobs. He is told only that 'We're in a recession,' or 'Interests rates are too high,' or 'Demand is depressed.' Joe never studied economics and finds these statements hard to understand. Explain to Joe, in language he is likely to understand, several possible reasons why the economy may not be providing more jobs for construction workers.

(3) A friend is coming to town by car to visit you. He phones from the highway on the outskirts of the city, asking for directions on how to find your home. Use the map below to give verbal directions on the shortest way to travel from his arrival point to your home. [The map would include a variety of streets and landmarks, with several possible routes to the house.]

(4) The table below gives information on population and age distribution trends in different parts of the United States, economic production, unemployment, employment in different vocations, and educational attainment. Using this information, describe major changes that occurred in the U.S. between 1900 and 1960 and also describe the variables, if any, that have remained relatively stable for long periods of time. Wherever possible, explain the reasons for change and stability.

QUESTION SEVEN

How are teachers being trained in performance assessment?

The Office of Technology Assessment of the U.S. Congress declared that the need for professional development for teachers is "critical" if performance assessment is to be successfully adopted in schools (U.S. Congress, 1992). Few teachers receive formal education in assessment, and inservice training on the topic is rare (Stiggins, 1988). Almost no administrators have had coursework on the topic (Stiggins, 1991).

What should teachers know about assessment? One expert suggested that all teachers should possess skills to enable them to carry out three types of assessment. The first is the ability to construct paper and pencil tests to measure students' knowledge and thinking. The second is the ability to design assessments focusing on important behaviors and products, and the third is proficiency in interacting with individual students in ways that give insight into student learning (Stiggins, 1991).

One area in which training is especially critical before performance assessment programs are adopted is preparing teachers to score students' work. Scoring is a concern because of its potential effect on the reliability of test results. One of the problems encountered in scoring assessment tasks is accommodating diversity in students' work while ensuring rigor in scoring. Two methods that have been proposed as models for evaluating student portfolios are the Environmental Beauty Estimation method used by the U.S. Forest Service to make environmental management decisions and the comparative method used by political scientists to study diverse political systems (Paulson & Paulson, 1991).

Administrators must be trained to implement new procedures and utilize data collected from alternative assessment methods. Colleges and universities must develop new, more appropriate courses in understanding and using assessment information, and a variety of other educational agencies must provide ongoing assistance and inservice training for teachers on alternative forms of assessment and their uses. Administrators need training to implement new procedures and utilize data collected from alternative assessment methods. Not only must administrators know the correct procedures for conducting performance assessments, they also must understand the proper use and reporting of varied types of assessments. As a school leader, the administrator must set the tone for his/her staff to utilize the information appropriately (Fulford, 1991).

Vermont has provided national leadership in assessment by piloting the use of portfolio assessments in math and writing and has developed 17 networks to train teachers in the use of portfolios for assessment ("Improving the math," 1991). However, a recent Rand Corporation study of Vermont's assessment program identified serious problems of reliability related to scoring. That finding led the author of the study to warn that people hold unrealistically high expectations for performance assessment and that increased awareness of the difficulties involved in developing and implementing these instruments is needed (Rothman, 1992e).

Vermont's was the first statewide assessment program to measure student achievement using portfolios. The author of the Rand study pointed out that the low reliability figures, which ranged from .28 to .57 on the writing portfolios, indicated that the scores are meaningless. (Score reliability for a uniform test of writing ranged from .67 to .75). The report suggested that improved training for teachers who rate portfolios might be effective in raising reliability to acceptable levels. Some school systems that use portfolios have attained much higher levels of reliability by employing a relatively small number of highly trained raters.

A statewide inservice effort is underway in Indiana to improve teachers' skills in conducting classroom assessment, including performance assessment. This program is adapted from Richard Stiggins' work at the Northwest Regional Educational Laboratory. (For information concerning Classroom Assessment Workshops, contact Robert Fallon at 317/232-9144) ("Regional actions," 1991).

QUESTION EIGHT

When is it appropriate to use performance assessment and when is it appropriate to use standardized multiple-choice tests?

When no performance assessment measure is available, the user must choose between a standardized multiple-choice test or a teacher-made test. If both a performance assessment instrument and a standardized multiple-choice test are available, the decision about which to use is determined by the content and quality of the instrument and the use to be made of the results. Perfcrmance assessment tasks are more likely to measure complex, higher order knowledge and skills and should be used for those types of objectives when available.

If comparisons are to be made across schools or districts, however, a standardized measure is required. If no standardized performance assessment measure is available (as will often be the case), then one must rely on a standardized multiple-choice test. If no standardized multiple-choice test is available that faithfully reflects the curriculum and the only performance assessment measure is not standardized, one is faced with choosing between two unsatisfactory options. This is the situation many school districts find themselves in today, and it helps account for the surge of interest in developing and validating performance assessment measures.

The popularity of performance assessment is a reaction against the standardized multiple-choice tests that are the mainstays of mandated testing programs in schools. Many educators believe that these tests are narrow and restrictive (Stiggins, Conklin, & Bridgeford, 1986) and that they "warp" the curriculum ("Toward high standards, 1991) and "corrupt teaching" (Darling-Hammond, 1991) by compelling teachers to teach to the test.

Among specific features of these tests that draw criticism are the emphasis on low-level recall and comprehension skills (Archbald & Newmann, 1988), the underlying assumption that choosing a right answer from several alternatives is an adequate indicator of ability

(Wiggins, 1989), and the reliance on outdated assumptions about how people learn ("Assessment and reform," 1992). Also, the tests are criticized for providing little evidence of the quality of students' reasoning or of their ability to formulate problems, develop answers or think analytically (Nickerson, 1989; Shavelson, Carey, & Webb, 1990; Darling-Hammond, 1991). Many educators advocate the adoption of methods of assessment that will provide more information about the quality of students' specific accomplishments, including the specific tasks on which they succeed or fail (Archbald & Newmann, 1988).

Despite the criticism, however, standardized multiple-choice tests have defenders. Gregory Anrig of Educational Testing Service says that they "remain a valid, efficient and inexpensive means to measure certain important aspects of student achievement" ("Groups call for," 1990). The multiple-choice format can and is being used to measure higher-order outcomes. Pennsylvania's new honors test in science and Louisiana's new graduation test in social studies both use multiple-choice formats (Popham, 1987).

Regardless of the type of assessment chosen, it is important to interpret test results with care. Average or median scores for a school or district can be misleading (Meyer, 1992). Scores at a given grade level reflect learning over several years. For example, a test given to sixth grade students measures what was learned in grades K through 5 as well as learning that has occurred in the current grade. Low scores may occur because of deficiencies in instruction from previous years.

Test scores may also be distorted by student mobility. Mean scores for a test given in the Spring are out of date by Fall if in the meantime many children have transferred out of the school and others have moved in. It is necessary to observe the performance of the same set of students over a period of years if instructional decisions are based on test results (Frederick, 1987).

Scores of standardized tests may be misleading for other reasons as well. Outdated test norms can result in group means at or above the 50th percentile on a standardized norm-referenced test for students whose performance on their daily work is unsatisfactory. Since many educators and parents assume that a score at or near the 50th percentile represents average performance, there is little inducement to take action to improve instruction in such situations. Careful examination of appropriate aggregated data and examination of student performance on individual items or subtests may show that substantial improvement is in fact needed (Ewell, 1988).

QUESTION NINE

What are the implications of performance assessment for norm-referenced tests?

There appears to be a reasonably good chance that performance assessment will eventually take the place of some standardized norm-referenced tests now used in schools. However, it is not likely that norm-referenced tests will disappear from the schools altogether. Because of the amount of time and money involved in developing new tests, it will probably be a good while before performance assessment will be ready for widespread use in schools.

Cooley (1991) suggests that if districts are to be held accountable for improving student achievement, then comparisons of student achievement over time are necessary. Standardized tests are designed to permit comparisons of school districts and schools. At the present time few if any performance assessment measures have been standardized, so states which plan to use test results to compare the performance of districts or schools must continue to rely on standardized norm-referenced tests for that purpose.

An assessment that is used for accountability has to be administratively feasible, professionally credible, publicly acceptable, legally defensible, and economically affordable (Mehrens, 1992). Norm-referenced standardized tests meet most of those criteria, whereas performance assessments have trouble meeting several of them. The greatest weakness of norm-referenced tests is in the area of credibility, which is a recognized strength of performance assessment, whereas affordability is a strength of norm-referenced tests and a weakness of performance assessment measures. Administrative feasibility is also a strength of norm-referenced tests but poses a problem for performance assessment because many of performance assessment tasks are complex and require prior training for teachers who administer and score them. Standardized multiple-choice tests are suitable for providing information about trends in achievement, but for assessing individual students' knowledge and skills, performance assessment is recommended ("Assessment and reform," 1992).

QUESTION TEN

How do we aggregate performance assessment data for policy and accountability purposes?

Aggregation refers to the process by which scores of individual students on an assessment measure are combined to represent the performance of groups. The mean reading score of fifth grade students in a school or the district mean for eighth grade students on a mathematics achievement test are examples of aggregated scores. Aggregated scores are also used to represent the performance of subgroups of students classified by ability, sex, race, family income, or other factors.

Three major functions of assessment measures are to provide diagnostic information for teachers' use, to monitor how well students in a school or school district are performing, and to provide data about individual students' performance for use in decisions about selection, placement and credentialing ("Assessment and reform," 1992). If the results of an assessment are used as the basis for decisions about individuals, it is important that the reliability of individual scores be high. If the results are used to make decisions about groups, then the aggregated measures should have high reliability. (This assumes that the assessment measures are valid. If not, then the validity problem must be solved before questions of reliability are taken up.)

The reliability of aggregated scores depends on the reliability of the scores at the individual level, the size of the samples comprising the aggregates, and variability of the aggregated scores. Reliability for aggregated scores refers to the extent to which group performance is consistent over time. If differences between groups on the attribute being measured are small, the aggregated scores can have low reliability even though the individual scores from which the aggregated scores are calculated are highly reliable.

Conversely, if groups are relatively heterogeneous and sample sizes are sufficiently large, aggregated scores can be quite reliable although individual scores are low in reliability (Willms, 1992).

The objective of monitoring student achievement is to reduce differences in performance among subgroups over time. Comparing a group's performance at one point in time to its performance at an earlier time involves an internal standard. An external standard is used when a group's performance is compared to that of a reference group outside the school (Frederick, 1987). An exemplary school is one in which the gap between the best and the worst student performances is continually being narrowed (Wiggins, 1991b), and in which the average performance of all student groups in the school approaches or exceeds the average performance of the appropriate external reference group.

In quality organizations, there is an attitude that the organization's performance is only as good as that of the weakest members, but in schools, tracking often institutionalizes low expectations and exaggerates differences (Wiggins, 1991b). Aggregating scores by socioeconomic level or other factors can have the same effect. This practice raises ethical questions unless information from the scores is used to reduce differences in achievement between groups. Aggregating scores in order to establish levels of expected performance has the effect of lowering performance expectations and locking in differences in achievement. Districts should be held accountable for bringing all students to high, not just expected, levels of performance (Cooley, 1991).

In addition to the ethical issues involved, the use of aggregated scores in schools also raises important technical questions. When the scores are used as the basis for instructional decisions, teachers and administrators should be aware of their limitations. School means are often not relevant for informing instructional decisions, since instructional effects are more likely to be mediated by individual teachers or groups of teachers (grade level or department). For that reason, it makes more sense to aggregate scores by classroom and grade level or department than to aggregate them by school. The scores can then be used to judge whether instructional practices of grade levels, departments or teachers need to be re-examined and, perhaps, changed. The only practices that might be influenced by schoolwide scores are the few that are implemented uniformly across the school. Two examples might be homework and discipline policies.

Student mobility should be taken into account in comparing student performance over time if test results are to be used as the basis for decisions about programs and policies, since, under conditions of high mobility, test results are out of date soon after the tests are administered (Meyer, 1992). If a test is administered to a different group of students the second time it is given, there is no way to determine whether the results should be attributed to instructional practices or to pre-existing differences in student ability and motivation (Frederick, 1987).

It makes sense to link assessment measures to factors that are known or are believed to influence student learning and over which teachers and administrators have control (Brown, 1990). Teachers and administrators can control how much homework students are assigned and the types of inservice classes that teachers take, but they have no control over whether a student is male or female or comes from a poor or wealthy family. By linking assessment to factors that are under the school's control, there is a greater likelihood that assessment information can be used to lead to improvements in student performance.

To assess the performance of a school or district over time, it must be possible to trace the performance of individual students, but many school districts do not presently have data bases that are adequate for that purpose. Evaluation of school programs can help locate deficiencies but are not very helpful in identifying and choosing among potential solutions. In Dallas, an advocacy team technique was developed to overcome that problem. The advocacy team locates and assesses competing solutions, once a problem or need has been identified. Subsequently a convergence team is formed to prepare a plan for correcting the identified weakness that combines the best features of all the proposed solutions (Stufflebeam & Webster, 1988).

QUESTION ELEVEN

How do we insure that performance assessment data are reliable?

Proponents believe that performance assessment should be adopted for use in schools because it will yield more valid evidence about student learning than the standardized multiple-choice tests currently in use. Most performance assessment tasks involve activities and products that are directly related to the learning outcomes schools seek to achieve and for that reason have the potential for high content validity. They also involve tasks with aesthetic, utilitarian, or personal value apart from their use for documenting learner performance (Newmann & Archbald, 1992), which increases student motivation and thereby helps ensure reliable results.

However, several features of performance assessment pose potential threats to reliability. Administration of performance tasks is more problematical because they are less easily standardized than paper-and-pencil tests. Performance tasks are more complex than the items used on standardized multiple-choice tests, and that complexity can lead to differences in the way tasks are administered. For example, a performance assessment may require that a student observe and classify objects, collect data, manipulate equipment, or construct a product. Equipment can malfunction, pieces of a demonstration unit can be lost or damaged, and supplies can deteriorate or be consumed. When any of these things happens, reliability declines.

Since performance assessment tasks call for construction of products or the production of complex thinking, scoring requires judgment and therefore poses a threat to reliability. There are several ways to minimize the threat. Training graders and requiring them to calibrate their scoring procedures periodically help to maintain acceptable reliability (Gentile, 1992).

The use of multiple graders also helps reduce the threat to reliability, but even with extensive training, inter-rater reliability varies depending upon the complexity of the product being evaluated (Mehrens, 1992). Diving, skating and gymnastics are examples of activities that successfully use multiple judges. In the case of diving, five judges rate each dive on a scale from 1 to 10. The highest and lowest marks are discarded, and the remaining scores are weighted according to the difficulty of the dive, summed, and averaged. Judging bias is not eliminated altogether by this procedure, but discarding the most extreme scores helps to minimize its effects. Biased judging is more often observed in sports such as skating, where artistic elements and personal qualities are prominent elements of the performance.

Checklists and other standardized grading schemes may be used to help increase reliability in scoring. One such procedure is called Graded Mark Point Scoring; it was developed by Pinchas Tamir for scoring student responses to open-ended items and involves analyzing specific features of the response (Bock, 1991).

Scoring guides developed by the National Assessment of Educational Progress to evaluate students' writing include descriptions of several different types of writing, with progressively higher ratings for more sophisticated or well-developed features (Gentile, 1992). The opposite approach is used in holistic scoring, which involves reading a writing sample quickly and making an overall judgment about its quality by evaluating the work as an entity without analyzing specific features (Mitchell, 1992; Partridge, 1990).

Another possible threat to reliability arises because of the amount of time required to complete some assessment tasks. The closer an assessment task is to a real life problem, the more time is likely to be required (Frechtling, 1991). Writing is an example. Some writing assessments require students to prepare a draft one day and complete the finished product the following day. Students who choose to use their free time during the evening to locate additional information about a topic or to study models of writing may improve their performance on the second day over those who spend the time watching television or listening to music.

Time also affects reliability of performance assessment in another way. Since the tasks takes longer to complete than those on standardized multiple-choice tests, the number of different tasks that can be included on an assessment is relatively small, and that results in attenuation of reliability.

Another possible threat to reliability occurs when a response is produced by a team of individuals. In Connecticut, high school chemistry students collaborate to analyze the drinking water in their schools and to prepare a report for the principal with recommendations on how the water might be made safer (Mitchell, 1992). All members of the team contribute to the final report, but individual efforts may vary. The threat to reliability in collaborative activities can be reduced by using evaluators to observe and rate individual contributions while members of the group are working together.

Scorers who have a stake in the results of an assessment are subject to bias if they expect to be evaluated or rewarded on the basis of the results of an assessment. Teachers should therefore not be assigned to score the work of their own students in such situations (Mehrens, 1992).

QUESTION TWELVE

What are the implications for local assessment programs, such as end-of-course testing?

One of the objectives of local assessment programs should be to change students' mindset about tests. A recent study of 600 high school students (Haertel et al, cited in Stiggins, Conklin & Bridgeford, 1986) found that students think of tests as limited to formal, paper and pencil assessments that usually ask objective questions and are distinct from ongoing instruction.

The study reported that students think the purpose of testing is to assign marks and grades. And although students believe tests are important, and they report that they are willing to work to earn high scores, they see tests as requiring mostly memorization. Students acknowledged that many important ideas they study are not tested at all, and they said they know more than their test scores show. Students indicated they are most comfortable with objective types of items and dislike testing formats that require more extensive responses (Stiggins, Conklin & Bridgeford, 1986).

QUESTION THIRTEEN

How does performance assessment relate to the Outcome Accountability Program in Virginia?

The National Council on Education Standards and Testing (NCEST) identified five purposes for assessment. They are: (1) to exemplify for students, parents and teachers the kinds and levels of achievement expected; (2) to assist in policy-makers to make decisions about educational programs; (3) to improve classroom instruction and improve learning outcomes of all students; (4) to inform students, parents and teachers about student progress toward achieving the standards; and (5) to measure and hold students, schools, districts, states and the nation accountable for educational performance ("Toward high standards," 1991).

Purposes of the Outcome Accountability Program (OAP) in Virginia relate to three of the five objectives identified by the NCEST (Objectives 2, 3 and 4). As identified by the Virginia Department of Education (1992), a major objective of OAP is to provide information to diverse audiences, from teachers and parents to state level policy-makers, and to inform the public about the performance of students in Virginia schools. OAP is also intended to provide information to enable local educators to initiate changes that will lead to increases in student performance and learning, to identify effective instructional practices, to use available resources more effectively and to provide assistance to schools or school divisions that are in need of improvement (Virginia Department of Education, 1992).

The two objectives identified by NCEST that are not among the objectives of OAP are Objectives 1 (to exemplify the kinds and levels of achievement expected) and 5 (to hold schools, districts and others accountable). The Virginia Department of Education (1992) disclosed that no criteria or standards have been established for Virginia schools, so the

OAP would not be expected to exemplify levels of achievement expected, nor would it be likely to fulfill an accountability function. However, one outcome envisioned for OAP that does not appear on the list publicized by the NCEST is to recognize schools for their progress and achievements (Virginia Department of Education, 1992). Recognition of certain schools may have the effect of creating de facto standards for both the types and levels of performance that will then be emulated by other schools in the Commonwealth.

The need for clearly-stated, challenging standards of performance was emphasized in the NCEST report. The Council observed that lack of high standards and quality assessments to measure progress toward them has led in the past to accountability being equated with observation of rules and regulations or with student achievement on tests of minimal skills. The Council advocated redefining accountability to mean attainment of meaningful outcomes as evaluated by fair and reliable means. It emphasized that the need for better quality assessment measures is especially great in areas in which consequences for students and teachers are attached to the results and cautioned that no assessment instrument should be used as the basis for assigning rewards or sanctions until the qualities of validity, reliability, and fairness have been addressed ("Toward high standards," 1991).

Standards are important because they help to increase congruity between what is expected of students, what is taught, and what is assessed in the schools. A number of professional organizations are working to develop standards for various school subjects. The National Council of Teachers of Mathematics was perhaps the first to define common standards. The history group is now funded, and work is underway in civics, English, science, geography, and, most recently, the arts (Anrig, 1992).

QUESTION FOURTEEN

What are the implications of State Department of Education efforts in performance assessment for school divisions?

Historically, states have regulated schools by setting standards for inputs into the education system, including expenditures per pupil and staff qualifications, but they have not tried directly to regulate educational practices, teaching methods or curriculum content. In recent years, however, states have begun to collect and use information on student performance for accountability purposes. Minimum competency tests were an early example of the states' attempt to use performance data to upgrade instruction. Minimum competency tests emphasized basic skills and so offered no incentive for schools to improve the teaching of higher order thinking. Now states are being urged to link accountability with assessment by establishing performance standards and supporting the development of assessment instruments to measure student attainment of higher order outcomes (Cohen, 1990).

The development and validation of assessment instruments is expensive and time-consuming. A number of collaborative activities involving state departments of education, professional organizations and government agencies are underway to develop and implement performance assessment measures for schools. Virginia can benefit from the experiences of these groups by participating as a cooperating member or observer of selected ventures. The state is already taking part in at least one of these programs (the New Standards project) and has committed itself to participate in others that hold promise for helping the State achieve its objectives ("Virginia assessment system," 1992).

The Virginia Department of Education should be commended for its plan to develop and operate regional assessment centers collaboratively with school divisions in the State for the purpose of increasing the skills of teachers and administrators in both traditional and alternative form of assessment ("Virginia assessment system," 1992). The soundness of this plan is reflected in findings from recent research showing that paper-and-pencil multiple-choice assessments measure aspects of achievement that are not captured by performance assessments (Shavelson & Baxter, 1992).

In view of the significant technical problems yet to be solved with respect to performance assessment, Virginia would be wise to move at a deliberate pace in developing and implementing new assessment measures for Virginia schools. The recent RAND study of Vermont's assessment program, which identified serious problems of reliability related to scoring of portfolios, has reinforced the expressions of concern by testing experts that people have unrealistically high expectations for performance assessment (Rothman, 1992e). Testing experts are advising schools to proceed slowly, starting with small pilot projects and to continue to rely on traditional assessment measures for the immediate future (O'Neil, 1992). Given the problems with reliability faced by performance assessment measures, that seems to be sound advice. Considerable research is needed to answer the many unanswered questions about performance assessment before large-scale implementation takes place.

The plan to introduce new assessment measures into Virginia schools beginning in Fall 1993 ("Virginia assessment system," 1992) is ambitious and can lead to counterproductive results if, in trying to meet the deadline, developers rush the development process. When assessments are introduced into the schools, it is important that they be worth teaching to (O'Neil, 1992). Introducing new assessments prematurely may mean that teachers will change the way they teach without necessarily producing better results (Shavelson & Baxter, 1992) and that they will lose faith in the potential of performance assessment.

Kentucky is attempting to implement performance assessment statewide by 1996 (O'Neil, 1992), and its experiences should inform decisions about the speed with which Virginia moves to implement performance assessment in the schools of the Commonwealth. One element of Kentucky's assessment program that is worth copying is the development of a "tool kit" of exemplary tasks, task templates, and design criteria for assessment tasks. Kentucky is doing this at the statewide level, providing dozens of tasks and task ideas to teachers as part of the new state performance-based assessment system (Wiggins, 1992).

Training of teachers and administrators is an essential component of a successful performance assessment program. The classroom has been called the "prime locus" and teachers the "prime resource" for improved learning in schools (Anrig, 1992). Teachers need to learn how to design and score performance assessments for their own classrooms, and both teachers and administrators must acquire increased skill in using performance information for making decisions related to instructional practice. If Virginia fails to provide training in performance assessment prior to the introduction of new ways of assessing student performance, teaching practices will be forced to try to catch up with assessment practices (O'Neil, 1992). The regional assessment centers can be an effective mechanism for providing training.

REFERENCES

Anrig, G. (1992, April 1). Can tests lead the way to excellence? Education Week, pp. 40.

Archbald, D., & Newmann, F. (1988). Beyond standardized testing: Assessing authentic academic achievement in the secondary school. Reston, VA: National Association of Secondary School Principals.

Aschbacher, P., & Herman, J. (1991). Alternative assessments in schools: Report on status and results of local projects. Los Angeles: UCLA Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 342 799)

Assessment and reform--are we headed in the right direction? (1992, Spring). State Education Leader, 11(1), 1-3. (Available from Education Commission of the States, Denver, CO).

Bock, R. D. (1991). The graded mark-point method of scoring performance exercises and open-ended items (Technical Report No. 323). Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 341 734)

Bock, R. D., & Zimowski, M. (1991). Individualized educational assessment: 12th grade science (CSE Report No. 324). Los Angeles: UCLA Center for Research on Evaluation, Standards and Student Testing. (ERIC Document Reproduction Service No. ED 338 672)

Brown, P. (1990). Accountability in public education. Policy Briefs. (Available from Far West Laboratory for Educational Research and Development, San Francisco, CA 94103). (ERIC Document Reproduction Service No. ED 326 949)

Chapman, C. (1990, December). Authentic writing assessment. ERIC Digest. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation, American Institutes for Research. (ERIC Document Reproduction Service No. ED 328 606)

Chow, S., & Matranga, M. (1990). Nevada's annual report ca. Policy Briefs. (Available from Far West Laboratory for Educational Research and Development, San Francisco, CA 94103). (ERIC Document Reproduction Service No. ED 326 949)

Cohen, M. (1990). Key issues confronting state policy-makers. In R. F. Elmore and associates (Eds.), Restructuring schools: The next generation of educational reform (pp. 251-288). San Francisco: Jossey-Bass.

Cooley, W. (1991, Winter). State-wide student assessment. Educational Measurement: Issues and Practice, 10, 3-6.

Darling-Hammond, L. (1991). The implications of testing policy for quality and equality. Phi Delta Kappan, 73. 220-225.

Foster, J. (1991, February). The role of accountability in Kentucky's Education Reform Act of 1990. Educational Leadership, 48, 34-36.

Frechtling, J. (1991, Winter). Performance assessment: Moonstruck or the real thing? Educational Measurement: Issues and Practice, 10, 23-25.

Frederick, J. (1987). Measuring school effectiveness: Guidelines for educational practitioners. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC Document Reproduction Service No. ED 282 891)

Frederiksen, J., & Collins, A. (1989, December). A systems approach to educational testing. Educational Researcher, 18, 27-32.

Fulford, N. (1991). Commentary. Policy Briefs, Nos. 15 & 16, 7. (Available from North Central Regional Educational Laboratory, Oak Brook, IL 60521). (ERIC Document Reproduction Service No. ED 326 949)

Gentile, C. (1992). Exploring new methods for collecting students' school-based writing: NAEP's 1990 portfolio study. Washington DC: U.S. Department of Education Office of Educational Research and Improvement.

Grobe, R., Adkins, D., Arrasmith, D., & Sheehan, D. (1983, June). Dallas: A large-city assessment program that works. Paper presented at the Large-Scale Assessment Conference, Boulder, CO. (ERIC Document Reproduction No. ED 241 553)

Groups call for phase-out of standardized tests. (1990, February 7). Report on Education Research, p. 5.

Horvath, F. (1991, April). Assessment in Alberta: Dimensions of authenticity. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago. (ERIC Document Reproduction Service No. ED 331 893)

Improving the math and science curriculum: Choices for state policy-makers. (1991). Washington DC: U.S. Department of Education Office of Educational Research and Improvement.

Linn, R., Baker, E., & Dunbar, S. (1991, November). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20, 15-21.

Lockwood, A. (1991, March). From telling to coaching. Focus in Change, 3(1), 3-7. (Available from National Center for Effective Schools Research and Development, University of Wisconsin, Madison, WI).

Lockwood, A. (1991, March). A leap of faith. Focus in Change, 3(1), 9-13. (Available from National Center for Effective Schools Research and Development, University of Wisconsin, Madison, WI).

Maeroff, G. (1991). Assessing alternative assessment. Phi Delta Kappan, 73, 273-281.

Mehrens, W. (1992, Spring). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practices, 11, 3-9, 20.

Meyer, R. (1992, Spring). Education reform: What constitutes valid indicators of educational performance? The LaFollette Policy Report, 4(1), 14-18. (Available from R. M. LaFollette Institute of Public Affairs, University of Wisconsin, Madison, WI).

Mitchell, R. (1992). Testing for learning: How new approaches to evaluation can improve American schools. New York: Free Press.

Newmann, F. (1992). The assessment of discourse in social studies. In Berlak, H. et al. (Ed.), Toward a new science of educational testing and assessment (pp. 53-69). Albany: State University of New York Press.

Newmann, F., & Archbald, D. (1992). The nature of authentic academic achievement. In H. Berlak, et al. (Ed.), Toward a new science of educational testing and assessment (pp. 71-83). Albany: State University of New York Press.

New York State Education Department. (1991). Student assessment: A review of current practices and trends in the United States and selected countries. Albany: Author.

Nickerson, R. (1989, December). New directions in educational assessment. Educational Researcher, 18, 3-7.

O'Neil, J. (1992, May). Putting performance assessment to the test. Educational Leadership, 49, 9-13.

Partridge, S. (1990) Assessing students' writing in the 1990s: A discussion. (ERIC Document Reproduction Service No. ED 322 512)

Paulson, F., & Paulson, P. (1991). The ins and outs of using portfolios to assess performance. Paper presented at the annual meeting of the National Council of Measurement in Education and the National Association of Test Directors, Chicago. (ERIC Document Reproduction Service No. ED 334 250)

Paulson, F. L., Paulson, P., & Meyer, C. (1991, February). What makes a portfolio a portfolio? Educational Leadership, 48, 60-63.

Popham, W. J. (1987). Muddle-minded emotionalism. Phi Delta Kappan, 68, 687-688.

Regional actions anu agendas. (1991). Policy Briefs, Nos. 15 & 16, 1-2. (Available from North Central Regional Educational Laboratory, Oak Brook, IL 60521).

Roeber, E. (1991). Guest commentary. Policy Briefs, Nos. 15 & 16, 6. (Available from North Central Regional Educational Laboratory, Oak Brook, IL 60521).

Romberg, T. (1992). Assessing mathematics competence and achievement. In H. Berlak, et al. (Ed.), Toward a new science of educational testing and assessment (pp. 23-52). Albany: State University of New York Press.

Rothman, R. (1992a, April 22). In a pilot study, student writing in class gauged. Education Week, pp. 1, 24.

Rothman, R. (1992b, April 22). Testing shifts from memorization to investigation in Littleton, CO. Education Week, pp. 1, 22-23.

Rothman, R. (1992c, November 4). Performance-based assessment gains prominent place on research docket. Education Week, pp. 1, 22, 24.

Rothman, R. (1992d, December 9). Report offers glimpse of mathematics assessment of the future. Education Week, p. 9.

Rothman, R. (1992e, December 16). RAND study finds serious problems in Vermont portfolio program. Education Week, p. 1, 20.

Rudman, H. (1987, February). Classroom instruction and tests: What do we really know about the link? NASSP Bulletin, 71, 3-21.

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

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

Sherwood-Fabre, L. (1986, October/November). An examination of the concept and role of program monitoring and evaluation. Paper presented at the American Evaluation Association, Kansas City. (ERIC Document Reproduction Service No. ED 286 927)

Siegler, R. (1989, December). Strategy diversity and cognitive assessment. Educational Researcher, 18, 15-19.

Simmons, J. (1990, March). Adapting portfolios for large-scale use. Educational Leadership, 47, 28.

Sizer, T. (1992, June 17). The roundtable: Eight questions: On cost, impact, the politics of who chooses. Education Week Special Report (By all measures: The debate over standards and assessments), pp. S4-S5.

Spagnolo, J. A., Jr. (1992, January 22). Memorandum to Division Superintendents and Regional Representatives.

Stefonek, T. (1991). Alternative assessment: A national perspective. Policy Briefs, Nos. 15 & 16, 1-2. (Available from North Central Regional Educational Laboratory, Oak Brook, IL)

Stiggins, R. (1988). Revitalizing classrooms assessment: The highest instructional priority. Phi Delta Kappan, 69, 363-368.

Stiggins, R. (1991). Assessment literacy. Phi Delta Kappan, 72, 534-539.

Stiggins, R., Conklin, N., & Bridgeford, N. (1986, Summer). Classroom assessment: A key to effective education. Educational Measurement: Issues and Practice, 5, 5-17.

Stufflebeam, D., & Webster, W. (1988). Evaluation as an administrative function. In N. Boyan (Ed.), Handbook of research on educational administration (pp. 569-601). New York: Longman.

Survey of national assessment initiatives. (1992, Spring). State Education Leader, 11, 8-10. (Available from Education Commission of the States, Denver).

The perfect school. (1993, January 11). U.S. News and World Report, pp. 46-50, 52-53, 56-61.

Toward high standards: A report to Congress, the Secretary of Education, and the National Education Goals Panel. (1991). Washington DC: National Council on Education Standards and Testing.

Tuckman, B. (1988). Testing for teachers (2nd ed.). Orlando, FL: Harcourt Brace Jovanovich.

U.S. Congress Office of Technology Assessment. (1992). Testing in American schools: Asking the right questions. Washington DC: U.S. Government Printing Office.

Virginia assessment system. (1992, March). [Richmond: Virginia Department of Education].

Virginia Department of Education. (1992). OAP Interpretive Guide to Reports. Richmond: Virginia Department of Education.

Walker, S. (1992, Spring). Assessing what students show to tell what they know. State Education Leader, 11(1), 12-13. (Available from Education Commission of the States, Denver, CO.).

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Wiggins, G. (1991a). Restructuring the curriculum through assessment. (Cassette Recording No. 61291100). Alexandria, VA: Association for Supervision and Curriculum Development.

Wiggins, G. (1991b, February). Standards, not standardization: Evoking quality student work. Educational Leadership, 48, 18-25.

Wiggins, G. (1992, May). Creating tests worth taking. Educational Leadership, 49, 26-33.

Williams, R., & Bank, A. (1981). Uses of data to improve instruction in local school districts: Problems and possibilities. In C. B. Aslanian (Ed.), *Improving educational evaluation methods: Impact on policy* (pp. 131-147). Beverly Hills, CA: Sage.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: Falmer.

Wood, R. (1984, Winter). Assessment has too many meanings and the one I think we want isn't clear enough yet. *Educational Measurement: Issues and Practice*, *3*, 5-7.

ANNOTATED BIBLIOGRAPHY

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

This guide should be of great value to teachers who wish to develop and use performance assessment tasks in their classrooms. It includes chapters on determining the purpose for assessment, selecting assessment tasks, and scoring. Examples of checklists, descriptive scales, scoring guides, and forms for rating portfolios are provided. Discussions of reliability in scoring and the use of performance data for instructional decisions are useful.

Roeber, E. (1991, April). *Guidelines for the development and administration of performance assessments in large-scale assessment programs*. Lansing, MI: Michigan Department of Education.

The author states that "the purpose of these Guidelines ... is to help the user of performance assessments consider in advance some of the issues to be faced and help plan the manner in which such assessments will occur." The paper presents a detailed step-by-step outline of the tasks involved in developing and validating assessment items, including pre-assessment activities, development steps, preparation for administration, administration, and post-administration administration activities.

Stiggins, R., & Conklin, N. (1992). In teachers' hands: Investigating the practices of classroom assessment. Ithaca, NY: State University of New York Press.

The authors review research on assessment practices and report on their own findings from studies of assessment practices used by teachers in elementary and high school classrooms. The framework used by the authors to describe assessment includes these dimensions: assessment purposes, methods, criteria used to select methods, quality of assessments, feedback, and policy environment. The authors report that teachers view instruction as separate from assessment and often have no strategy for integrating the two. They also found that teachers wish to base students' grades on achievement but are influenced in these decisions by affective factors.

U.S. Congress Office of Technology Assessment. (1992). Testing in American schools: Asking the right questions. Washington DC: U.S. Government Printing Office.

This publication presents a comprehensive review of the use of tests in schools. One of the eight chapters deals with the history of educational testing, and another reports on the uses of tests in other nations. One chapter is devoted to standardized tests, and another deals with performance assessment. The authors point out that the move toward alternative forms of testing has been motivated by new understandings about how children learn and by changing views of curriculum. They point out that "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" (p. 204).