Large-Scale Authentic Assessment

Marie Hoepfl
Appalachian State University

Student assessment is the most widely used approach being taken by state and federal policymakers in their attempts to leverage improvement of instruction in the nation’s schools. In particular, performance assessment is seen as a way to set targets for what students should know and be able to do, encourage curricular reform, and improve teaching methods. Among other claims, performance assessments are viewed as mechanisms for promoting greater educational equity (Roeber 1997). The reliance on standardized, norm-referenced multiple-choice tests for large-scale assessments has yielded recently to an increased emphasis on performance skills and “thinking abilities” needed in the workplace and in daily life. Proponents claim that performance assessments can better tap the skills and abilities that students need (Darling-Hammond, Ancess, and Falk 1995).

Since 1970, when standardized tests began to be more widely used, educational researchers have seen slight increases in basic skills test scores, but declines in measures of higher-order thinking skills. Officials within national organizations, ranging from the National Research Council to the National Councils of Teachers of English and Mathematics, among others, have attributed this decline to the emphasis on tests of basic skills, which have driven the curriculum (ibid.).

The structure behind performance assessment contrasts sharply with the discrete items found on multiple-choice assessments. Rather than artificially separating desired knowledge and skills into small pieces, performance assessment attempts to measure behavior as an intact whole (Yen 1993). “In assessment reform theory, all performance assessments must require students to structure the assessment task, apply information, and construct responses, and, in many cases, students must also be able to explain the processes by which they arrive at the answers” (Khattri, Reeve, and Kane 1998, p. 2).

The latest rounds of curriculum reform advocate the use of performance assessment as a lever for encouraging curricular and instructional change. This emphasis on performance assessment stems from three sources: (1) a backlash against the pressure for accountability through standardized testing, (2) the expansion of cognitive science (with its emphasis on constructivist teaching and learning), and (3) concern from the business community that schools are not adequately preparing youth for today’s workplace. In addition, several national, nongovernmental projects designed to address curricular, instructional and assessment reform have gained prominence in recent years. These include the New Standards Project, the Coalition of Essential Schools, and the College Board’s Pacesetter program, all of which have influenced a shift toward the use of performance assessments (Khattri, Reeve, and Kane 1998).
Harris and Kerby (1997) believe that the strongest argument in favor of performance assessments is the chance that they will balance the scores of students who perform relatively poorly on multiple-choice tests. For example, men tend to outperform women on multiple-choice tests, so essay tests could yield an inappropriate misclassification of women’s knowledge or abilities based on these types of test scores.

The national standards that have been developed in many curricular areas, including the technology education standards (currently being developed by the International Technology Education Association), emphasize the acquisition of higher-order thinking and process skills. Unfortunately, when these curricular reforms encounter high-stakes decisions about students, programs, or schools based on mandatory standardized tests of basic skills, teachers have little incentive to pursue alternative approaches to instruction, and “the tests win out” (Darling-Hammond, Ancess, and Falk 1995, p. 10).

One can also trace the emphasis on accountability testing in general to our enhanced ability to test. Increasingly sophisticated tools, from the first Iowa Test of Basic Skills in 1929 to our current capacity to process enormous quantities of data electronically, have contributed to a kind of technology-driven push for accountability testing. As Rothman (1995) notes, Americans, fascinated by technological solutions to social problems, find that tools such as electronic scoring and recordkeeping make testing almost irresistible. More positively, as computer software becomes more sophisticated, it is becoming more possible to analyze rich qualitative data on a large-scale basis.

Federal Mandates and Initiatives

The Goals 2000: Educate America Act passed in 1994 mandated that states detail how student performance will be measured against established standards. Goals 2000 provided federal funding for the development of standards-based education systems, which provide the base for authentic assessments. Other federal mandates include Title I legislation, which is designed to encourage a move away from norm-referenced testing by allowing districts the flexibility to develop their own standards and assessments, provided they are as rigorous as those of the state (Khattri, Reeve, and Kane 1998). The net result of federal legislation promoting assessment alternatives is that “substantially more assessment is likely to occur in our nation’s schools and to take place in areas traditionally not assessed (such as the arts) using assessment strategies (such as performance assessments and portfolios) not typically used” (Roeber 1997, p. 6).

The Carl D. Perkins Act requires states to develop performance standards for vocational programs. The states are also required to measure the effectiveness of vocational education programs related to the attainment of identified skills, school retention, program completion, job placement, and the progress of special populations. The 1990 Perkins legislation marked a “significant turning point in federal accountability by explicitly tying the process of state and local review to standards based on outcomes” (Office of Technology Assessment 1994, p. 5). As a
result, state assessment activity in vocational education increased during the 1990s to meet these accountability requirements. The Perkins Act does not, however, specify the types of assessment strategies that must be used.

In general, vocational skills assessment falls into four categories: academic skills, job-specific vocational skills, generic workplace skills, and broad technical skills. The diversity of assessment methods used to measure these skills is broad, ranging from student portfolios, to structured ratings of student capabilities demonstrated through classroom work, and organized competitive events (Office of Technology Assessment 1994). According to an OTA survey of state assessment directors, it is in the areas of vocational skills and generic workplace skills that the greatest expansion of assessment activities is likely to occur. In fact, vocational educators have used authentic assessment strategies and tools for many years.

Another federal initiative that has spurred the move toward performance-based assessment is the U.S. Department of Labor Secretary’s Commission on A chieving N ecessary Skills (SCANS) Report. “The SCANS commission envisioned setting proficiency levels for SCANS competencies and developing an associated assessment system based on demonstrating SCANS competencies through applied, contextualized problems” (Khattri, Reeve, and Kane 1998, p. 5).

The effectiveness of using testing to implement educational standards and ensure accountability for outcomes is yet to be determined. Although there continues to be considerable political and popular support for the concept of accountability through standards and assessment, significant technical, political, and logistical problems remain (Madaus and O’Dwyer 1999; Milne 1998; Wildavsky 1999).

**Statewide Efforts to Use Authentic Assessment**

Statewide systems of standards and measures of performance were mandated by the 1990 Perkins Act, which required an accountability system built around standards, outcomes, and performance measures. These systems were required to address mastery of academic and occupational skills, program completion, and employment. The framework of standards and measures adopted by each state should serve as a common tool for evaluating and improving vocational education programs (Milne 1998).

Data from the 1992 National Assessment of Vocational Education (NAVE) Omnibus Survey showed that virtually all states were in the process of developing performance standards, and over 75% of states were assessing (or were planning to assess) secondary student performance based on these standards. Prior to 1991-92, only 18 percent of the states were involved in this type of aligned standards-based process (Milne 1998). It should be noted that these data do not indicate what type of assessment is being used or planned.

There are a variety of performance measures that can be adopted to assess vocational programs. These can be grouped into the following categories: enrollment numbers, academic skills, occupational skills, school completion, job placement,
wages, and/or job retention. Of these, several can appropriately be considered forms of performance assessment. In the NAVE survey, more than 80 percent of the states reported plans to use at least five of these seven types of measures (Stecher, Farris, and Hamilton 1998). The most significant trend was toward a greater use of skill measures, particularly those involving advanced skill measures.

State vocational education officials, school administrators, and local vocational education administrators and staff have been instrumental in developing standards and measures for vocational education in more than 70 percent of the states surveyed in the NAVE study. Representatives from special populations were the only other group that has been heavily involved in the process. Employers, students and parents were also consulted in approximately 85 percent of the states (Stecher, Farris, and Hamilton 1998). Thus, the primary stakeholders in vocational education have been engaged in this developmental effort.

The Perkins Act has included provisions permitting states to adjust state performance standards to accommodate special populations, school resources, and local conditions. Fifty percent of the states report that they plan to make adjustments accordingly. In 64 percent of states, all students who take vocational courses were measured, whereas only 7 percent apply performance measures to the most narrowly defined population (e.g., vocational completers). Interestingly, those states that have done the most to promote academic-vocational integration were found to have also done significantly more with performance assessment measures (Stecher, Farris, and Hamilton 1998). This is good for the academic areas and it also reflects positively on vocational education.

A case study of 16 districts and states conducted by Khattri, Reeve, and Kane (1998) found that the characteristics of performance assessments varied from site to site. There is a wide range in the type and complexity of tasks required of students under the umbrella of performance assessment. Everything from open-ended, short-answer questions to completion of extended projects can make up the universe of “constructed responses,” which are a characteristic part of performance assessments. Significant differences also exist between state testing policies for general education and vocational education. What is most important—

no state vocational education agencies directly administer a program of mass testing or assessment of all students at a fixed point in time. In most states, the primary assessment responsibility of the agency is to set policies for local programs to follow. (Office of Technology Assessment 1994, p. 11)

According to the OTA, there is virtually no tradition in vocational education for the use of norm-referenced tests. Vocational education has long embraced the concepts of competency assessment, skill attainment, active student involvement, and assessment embedded within instruction. “In all of these respects, the traditions of testing and assessment in vocational education resemble what is [now] being advocated elsewhere in the rest of education” (ibid., p. 40).
California, Kentucky, Maryland, and Vermont are considered to be early leaders in the development and use of state-level performance assessments (Khattri, Reeve, and Kane 1998; Rothman 1995). California’s Learning Assessment System (CLAS) was canceled by Governor Pete Wilson in 1994 following strong public criticism, but the programs in the remaining three states appear to be fairly well established. The Kentucky Instructional Results Information System (KIRIS), in full-scale implementation since the mid-1990s, includes both multiple choice and performance event components in assessment of vocational studies and “practical living,” as well as art, math, reading, science, and social studies (Khattri, Reeve, and Kane 1998).

The OTA survey conducted in 1994 found that states increasingly appear to be expanding their use of written testing in vocational education “at the very time that questions are being raised in the rest of education about the effectiveness of standardized testing” (p. 59). This raises the question of what the long-term effects on instruction in vocational education are likely to be.

**Implementation Issues**

The increased development of performance measures, reported in the 1992 National Assessment of Vocational Education surveys, led to an increased burden on state officials. Vocational education staff in almost 80 percent of the states reported having more responsibilities related to these tasks than they had a decade earlier (Stecher, Farris, and Hamilton 1998). These researchers suggest that the expertise of educators at the local district level should be tapped when establishing new assessment systems, thus saving time and effort. This makes sense, not only from a resource standpoint, but also in terms of commitment and expertise.

Additionally, states can look to national projects or commercial entities for assistance in developing, delivering, and/or scoring alternative testing systems. Performance assessments have always been a part of some of the College Board’s Advanced Placement (AP) programs. For example, the Studio Art Portfolio Evaluation has no written or multiple-choice portions. The AP Art Portfolio is one prominent example of an established, national portfolio examination (Khattri, Reeve, and Kane 1998).

In Kentucky, students are expected to complete performance tasks and submit a portfolio of best work over the course of a school year, in addition to a more traditional component to its statewide testing system (KIRIS). These state assessments, with the exception of the student portfolios, are scored by a private firm called Advanced Systems in Measurement. Maryland, which has also incorporated traditional and performance components into its state-level assessments, contracts with CTB/McGraw-Hill to operate its program (Rothman 1995). This use of the private sector to develop and implement large-scale assessments is probably wise, given the demands on human and capital resources required to conduct large-scale performance assessments.
Obstacles and Challenges

The obstacles or challenges facing states that want to implement performance assessments have been grouped into two broad categories: practical challenges and technical challenges (Roeber 1997). Practical challenges are those that involve administering and scoring large-scale assessments, whereas technical challenges have to do with the validity and reliability of the assessments themselves. These challenges are present in any assessment situation. They are particularly problematic when assessment is conducted on a large-scale basis. Both types are discussed further in this section.

A primary problem with all large-scale assessment is accurately matching the performance being tested with the stated goals and objectives. Mager (1973) provides a humorous example to illustrate the problem. Suppose an instructor gave you the objective “On a level paved street, be able to ride a unicycle 100 yards without falling off.” You work hard to develop this psychomotor skill, only to find out on assessment day that the “test” consists of the following questions:

- Define unicycle.
- Write a short essay on the history of the unicycle.
- Name at least six parts of the unicycle. (p. 1)

The assessment items being used in this example suffer from a lack of validity. The obvious message is that educators must be careful to match the assessment to the desired behavior or condition, a task not always easy to accomplish. The technical terms for this are construct and content validity.

Sources of Invalidity

Messick (1996) describes two major threats to validity of performance assessments. Construct underrepresentation means that the assessment is too narrowly focused, failing to include important dimensions of the knowledge or skill it aims to assess. Construct-irrelevant variance refers to assessments that ask for responses that are not relevant to measuring the desired knowledge and skills. Thus, “a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance” (Messick 1996, p. 5). If the irrelevant tasks are overly difficult, assessment scores will likely be invalidly low. If the irrelevant tasks are overly easy, assessment scores may be invalidly high.

Crocker (1997) provides an excellent overview of the elements that must be considered when judging content representativeness. These include “(a) the relevance of the test item content to the knowledge domain of interest and (b) the balance of coverage of the items in relation to the breadth of the domain. Some experts also consider review of the technical quality of items and fairness to examinee subgroups” (p. 84). The problem of subjective decision making with regard to test item content is exacerbated on performance assessments because they have
fewer, and thus more heavily weighted, items. Subjectivity can also be introduced through scoring rubrics, which are influenced by the preferences of the rubric developers (Crocker 1997).

A concrete example helps to illustrate the issue of content representativeness. A student has mastered 90 out of 100 concepts from the material to be tested. Given time limitations, not all concepts will be included on the test. With a multiple-choice test, there could be 60 items, compared to 6 items on a constructed-response (essay) test. The likelihood of achieving good content representativeness is much higher on the test with 60 discrete items than the test with 6. Theoretically, all 6 essays could come from the 90 known concepts or from the 10 unknown concepts. So the students’ scores could range from 0 to 100 percent. In these circumstances, “the essay exam is much more likely to underestimate or overestimate the student’s true knowledge of the domain and result in an erroneous decision about the student’s competence.” In large-scale testing programs, there is no opportunity to mitigate this possibility with a variety of other classroom scores (Phillips 1993, p. 108).

Another problematic dimension of validity has to do with the consequences associated with interpreting scores. Specifically, the concern is that any negative impact that results from the use of an assessment should not stem from any source of test invalidity. Assessment must include efforts to discover the intended and unintended consequences, in the short- and long-term, of how scores are used (Messick 1996). In high stakes, large-scale assessment efforts, these concerns are even greater. Relatively minor technical challenges can escalate into major political issues (Barton 1999; Wildavsky 1999).

Because of the particular, and sometimes sweeping, claims made for the benefits of performance assessments, there are some specialized criteria by which they should be judged. To be worthwhile and motivational educational experiences in their own right (as the claims go), the tasks posed should be meaningful to students and clearly communicate what is expected. These traits have been referred to as “meaningfulness” and “transparency” (Dunbar, Koretz, and Hoover 1991; Messick 1996, p. 13).

Yet another validity issue has to do with the variability that can be introduced by using different methods or prompts to introduce the performance task. Student performance can be “extremely sensitive to subtle changes in format and presentation,” resulting in scores that do not truly reflect ability levels (Phillips 1993, p. 11).

**Fairness**

Bond, Moss, and Carr (1996) identify two aspects of fairness with regard to assessment. The first is test bias, which relates to the validity of an interpretation or action based on test scores. Bias exists when there is evidence of differential validity for any relevant subgroup of persons assessed. The second aspect of fairness relates to the soundness of the educational system upon which the assess-
ment is based, or what these authors call equity. In other words, did all students being assessed have access to the same quality of educational experiences? Evidence suggests that curricular changes caused by traditional standardized testing have affected nonwhite students disproportionately. They are more likely to spend time in direct test preparation (content drilling rather than more motivational types of activities) than their white counterparts (Bond, Moss, and Carr 1996). Given the differential access to high-quality education, there is no reason to expect that underserved minority groups will fare any better with performance assessments than they have traditionally done, and may actually do worse (Herman, Klein, Heath, and Wakai 1994). Studies of performance assessments in science found considerable variance in mean performance from one ethnic group to another (Dunbar, Koretz, and Hoover 1991).

**Reliability**

Reliability refers to the consistency of a measure over time, closely related to the concepts of generalizability and comparability described later. Among the many challenges to the appropriateness of performance assessment scores, interrater reliability is of least concern. Evidence from several studies indicates that high levels of agreement between scorers can be achieved, given sufficient training. The issues of where to set cut-off scores and how to deal with scores that fall near those cut-off points, however, have not yet been adequately addressed (Jaeger, Mullis, Bourque, and Shakrani 1996; Khattri, Reeve, and Kane 1998; Shavelson, Baxter, and Gao 1993).

**Generalizability**

Concerns about the content representativeness of performance assessments seek to ensure that interpretation of test scores need not be limited to the sample of assessed tasks, but rather be generalizable to the broader set of skills and abilities desired (Yen 1993).

[The] issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address. However, because of the extensive time required for the typical performance task, there is a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. (Messick 1996, p. 11)

One way this has been addressed has been through the use of “matrix-sampling,” where different samples of students perform different (but only a few) sets of tasks. The amount of time spent by any one student is minimized. Scores are evaluated in the aggregate, permitting comparisons between larger groups such as districts, states, or nations, rather than at the individual student level. This makes matrix-sampling useful for large-scale efforts like the National Assessment of Educational Progress (Calderone, King, and Horkay 1997).
In addition to matrix sampling, Brennan (1996) suggests combining performance assessments with traditional testing. Alternatively, a series of short-term performance tasks could be devised, so that a larger number of items could be included. In this way, the benefits of performance assessment (such as greater authenticity) could be realized without sacrificing generalizability.

**Comparability**

Another goal of large-scale testing is that scores should be comparable over time as well as from sample to sample. This comparability requires that the content assessed with each sample remain proximally the same. On multiple choice tests the number of items means that the significance of any one item is small and it is easy to create comparable tests over time. Because performance tasks are fewer in number and are more distinctive, they must be sampled with greater care than multiple-choice items, so that the knowledge and skills assessed remain stable over time (Haertel and Linn 1996).

**The Challenges of Setting Assessment Standards**

Closely related to comparability is the issue of setting assessment standards against which student work is to be judged. Jaeger et al. (1996) identify some factors that complicate the process of setting standards. One is that performance standards rarely occur naturally in ways that make it obvious where the boundary between acceptable and unacceptable work lies. This is less true for some skills tasks, such as might be found in vocational areas. Students can either perform the task correctly, or they can’t. Another factor is that the people who set performance standards are not always trustworthy judges of the quality of the standards they have set. If performance assessments are to be used in accountability-based strategies for promoting systemic educational reform, the issue of setting assessment standards must be addressed.

Developing comparable standards across performance assessments appears to be the most problematic venture of all. There is, first, the problem of designing appropriate performance tasks in terms of content standards, and identifying student work on the tasks that exemplifies success in meeting the standards. Implementing such assessments under consistent conditions and evaluating resulting performances reliably pose enormous operational challenges. (ibid., p. 87)

Curriculum standards specifying what students must know and be able to do as a result of instruction have been developed for mathematics, English, civics, geography, history, foreign languages, science, social studies, and the arts. In Spring 2000, standards will also be issued for technology education. Formal assessment standards for determining when standards have been met are not typically a part of these documents (Jaeger et al. 1996). However, some of these national standards projects are in the process of developing (or proposing to develop) assessment standards (e.g., the Technology for All Americans project).
The essential concern with setting cut-points for levels of performance is that the process imposes artificial dichotomies on what is in reality a continuum of proficiency. In the real world, proficiency does not occur at discrete, easily recognizable points. “The problem is how to treat the gray areas around the cut-points, since a certain proportion of examinees just above or just below a cut-point will almost inevitably be misclassified due to measurement error” (ibid., p. 104). One study found that nearly six times as many students would fail an assessment when using one standard-setting measure as opposed to another. As with other assessment considerations, political realities demand that issues surrounding interpretation of scores be done in a manner that accounts for the many “shades of gray” that are involved in making assessment-based decisions.

A related concern is the degree to which curriculum standards, once adopted by a state, are implemented at the local level. As a result of its Education Reform Act of 1993, Massachusetts developed a series of tests for grades 4, 8, and 10, based on its curriculum framework for technology education. Results from the first year of testing show that student performance at the higher grade levels is significantly lower than at grade 4. One reason suggested for the low first-year test scores at grades 8 and 10 is that, at the time the tests were administered, only 30 percent of the school districts in the state had aligned their curriculum to the new standards (Bouvier and Corley 1999).

Multiple Purposes

The primary purposes of assessment are to monitor student progress, establish accountability, certify student achievement, and align curriculum, instruction, and assessment. The ability to judge the effectiveness of performance assessment systems, particularly at the state level, is hampered by the fact that many systems are set up to achieve multiple purposes. Factors that facilitate the achievement of one purpose (e.g., standardization for purposes of accountability) may serve as barriers to the achievement of another purpose (e.g., informing instructional practice) (Khattri, Reeve, and Kane 1998). With large-scale assessment, there is an increased chance that the challenges associated with competing priorities and goals will occur.

Cost

Developing and implementing performance assessments is an expensive undertaking. Other aspects of the assessment reform process that require financial support include research on assessment methodology, delivering professional development, disseminating information, storage space for assessment materials, time spent by teachers preparing for assessments, and more (Khattri, Reeve, and Kane 1998). Additionally, some of the costs of performance assessment are not obvious or are not known, such as time spent with governmental and nongovernmental agencies, state departments of education, etc. (Hardy 1995).

Estimates of costs are also difficult to establish because there are few authentic assessments in place on a large scale (Rothman 1995). In 1992, the OTA reported
on testing in American schools. In this report, the costs associated with traditional testing in large districts (including both direct and indirect costs) were estimated to be approximately $37 per student. The General Accounting Office (GAO), in a similar study, estimated per student costs of traditional testing to be much lower, at $16. Estimates of the cost for performance assessments made by these same agencies ranged from twice as much as traditional (GAO) to 3-10 times as much as traditional (OTA) (Rothman 1995).

A limited number of private companies are developing performance assessments. Hardy (1995) examined several large-scale performance assessment programs to estimate the costs of development, implementation, and scoring. He notes that development costs can often be hard to assess, particularly when existing staff is used. Based on available data, however, development costs ranged from $5,000 to over $14,000 per task. Costs tend to be lower when the student outcomes are well defined, when smaller sample sizes are used to pilot assessment tasks, and when the size of the development teams is kept to a minimum. Development costs can also vary considerably depending on the content area. When local educators are used to develop items, the cost of their training must also be added.

The costs of performance assessment fall into three categories: development, administration, and scoring. All three can vary widely depending on the nature of the assessment task, the type of work produced, and the amount of information required from individual responses. Hardy (1995) examined several large-scale performance assessment programs to estimate the costs of development, implementation, and scoring.

Administration costs include materials and staffing. Performance assessment kits for science and mathematics tasks developed by the National Assessment of Educational Progress (NAEP), by the Educational Testing Service for the state of Georgia, and others ranged in cost from a low of $.70 to a high of $13.50 per kit. Ways to reduce the cost of materials include testing only a sample of students or using the same kit over a multiyear period and prorating its cost. Another possible approach is to require all classrooms to have a common set of equipment or materials that would be used in the classroom over the course of a school year for instruction, as well as on the performance test. Staffing costs for actual delivery of the assessments can also be difficult to calculate, particularly when local personnel are used. Kentucky uses external task administrators at a cost of approximately $5 per student (Hardy 1995).

The costs of scoring performance assessment tasks are considerably higher than those associated with scoring traditional multiple-choice tests. Most performance tasks require some form of human analysis, if not outright hand scoring. Estimates of scoring costs, largely based on writing tasks conducted in various states, range from $3-$6 per student (Hardy 1995).

The U.S. Office of Technology Assessment estimates the cost of using performance assessments will be from 3 to 10 times greater than the costs associated with traditional tests. Other estimates have suggested they could be up to 60
times more costly. However, as performance assessments are more widely used, their cost per assessment unit will likely decrease (Hardy 1995). The savings may be less significant because per student development costs per student drop with larger numbers, whereas the other major cost factors (e.g., materials, administration, scoring) do not (Stecher 1995).

Professional development in the use and scoring of authentic assessments is critical to their success (Khattri, Reeve, and Kane 1998). The cost for the level of training required for reliable scoring of these assessments is considerable, particularly compared to traditional forms of testing. Other aspects of the assessment reform process that require financial support include research on assessment methodology, disseminating information, storage space for assessment materials, time spent by teachers preparing for assessments, and more (ibid.). Some of the costs of performance assessment are not obvious or are not known, such as time spent with governmental and non-governmental agencies, state departments of education, etc. (Hardy 1995).

According to the NAVE 1992 Omnibus Surveys, less than half of school districts surveyed reported any increase in state assistance with accountability assessments, and fewer than 20 percent noted any state-sponsored training programs on student assessment or performance assessment for vocational educators (Stecher, Farris, and Hamilton 1998). In addition, although their use was mandated by the 1990 Perkins Act, few districts used Perkins Title II basic grant funds to develop or expand vocational performance assessment systems.

Stecher (1995) calculated the approximate cost of traditional (paper and pencil) testing in science using the California Test of Basic Skills at $.30 per student. By contrast, open-ended written-response items on the same test cost $4.80 per student per prompt. In his study, Stecher examined the costs of developing, implementing, and scoring performance tasks for science. The study suggests that the cost of hands-on science assessment can run as much as 100 times higher than standardized multiple-choice tests. Hands-on science performance tasks developed and implemented in this study were calculated at $30 per student per test period of 45-50 minutes, provided 100,000 students take the test and the economy of scale is realized. With fewer students taking the test, costs will go up significantly. This cost does not include teacher time for administering the performance test, which was considered “contributed time.”

Time

There are two dimensions of time that may be problematic when using large-scale performance assessments. One challenge is the amount of student testing time required to administer a sufficient number of items to satisfy validity and generalizability concerns. The second challenge is the turn-around time for test results, which can sometimes be as long as several months. “This time lag between assessment and reporting is so large that local educators may view the results (and the overall assessment program) as relatively useless” (Roeber 1997, p. 8).
Preliminary Reactions to the Large-Scale Use of Authentic Assessment

Some responses to large-scale adoption of performance assessments have been less than positive. Problems that face all large-scale assessments of any type, including inappropriate testing practices, breaches in test security, adverse impacts on historically disadvantaged groups, and others can plague performance assessments just as they do traditional multiple-choice tests. In fact, some preliminary data suggest that many of these issues may be even more pronounced with performance assessments (Phillips 1993). Implementation in some states has led to poor results, whereas other states continue to plan to implement performance assessments in the near future. Givens (1997) suggests that communication regarding the “myriad problems” associated with this form of testing has to date been limited.

In some cases, the issue has become highly political. For example, in the early 1990s, educators in Littleton, Colorado attempted a system of reforms that they hoped would help students develop better problem-solving and communication skills. They established standards, redesigned instructional practices, and created new performance assessments. Although many teachers, parents, and students felt positive about the reforms, a vocal and, as it turned out, powerful group of residents opposed them. They viewed the new assessments as being too new and untried, and too reliant on teachers’ judgments to be appropriate for high-stakes decisions such as determining whether students would graduate from high school. The critics also complained that the schools should not focus on problem-solving abilities, but rather on knowledge of a core body of information. In 1993, in a heated school board election, three community members who opposed the changes were voted into office. They subsequently scrapped the reform program and removed the superintendent of schools from office (Rothman 1995).

Rothman believes there are several reasons why reforms based on standards and new assessments have met with strong resistance in some communities. First, the establishment of explicit standards, while necessary from the standpoint of clearly communicating expectations, also invites challenges about what we really do want students to know and be like, and who should decide that. Second, many opponents object to the methods of teaching, constructivist in nature, that shift greater responsibility for acquiring knowledge onto the student. Critics of California’s CLAS reforms raised similar objections, saying the assessments chosen were designed to measure attitudes and beliefs, rather than academic knowledge and skills (Lewis 1996; Rothman 1995). As noted earlier, use of the CLAS system was halted.

How Are Data Being Used?

At a 1993 meeting of the International Congress on School Effectiveness, educators from the United States, the United Kingdom, Sweden, and Holland discussed the role assessment plays in school reform. A widely held belief that emerged from
these discussions was that “school improvement would not occur if schools were left to take action on their own.” In the absence of external evaluations, participants agreed, schools would continue to do what they had been doing (Riley and Nuttall 1994, p. 126). However, if alternative assessments are to drive education reform, as many would like, the data should at some point be fed back into the decision-making structure at the local level, where school improvement must be sustained.

The OTA (1994) found that states use occupational skill assessment data differently than they use data from academic skill assessments. Occupational data is most often used to evaluate student attainment for certification or program completion. The second most frequent use is for accountability (is the program doing what it is supposed to be doing?). The third most frequent use is for making decisions about the improvement of courses, programs, or schools. In other words, schools are least likely to use assessment information to improve programs. They are unlikely to link information about academic skills to instruction, but rather collect that information for Perkins accountability purposes only.

The practice of “teaching to the test” is a recognized outcome of high-stakes assessment. Teaching to the test can cover a range of interventions, not all of which are ethical. One problem with teaching to the test is that it can narrow the curricular focus to only what is on the test, or sacrifice material at the expense of covering tested material. Another problematic trend related to standardized tests is that disadvantaged students are less likely to receive instruction in science, art, and thinking skills, and more likely to receive drilling on the so-called basic skills (Rothman 1995). The reality of high-stakes testing is that it will have an effect on instructional practices. For this reason it is imperative that teachers have a clear understanding about the measures being taken, so that they can organize their instruction accordingly (Popham 1999).

Certain educators have questioned the trend toward large-scale performance assessment and, in fact, the whole foundation upon which large-scale assessments of any kind are based (Andrews 1997; Barton 1999; H Aertel 1999; Lewis 1996; Lissitz 1997; M adaus and O’Dwyer 1999). Lissitz (1997) believes there is little evidence that performance assessment will lead to better teaching, any more than traditional assessments have done. He and other critics maintain that, if we really hope to reform classroom teaching, we should advocate for change in the teaching environment, not for changes or additions to state-level accountability testing. According to Eisner, what really needs to change is the conception of schools in the minds of the public. “A shift needs to be made from a conception of schooling as a horse race or a kind of educational Olympics to a conception of schools as places that foster students’ distinctive talents” (Eisner 1999, p. 660).

At the bottom line, the question that must be addressed is: “What will performance testing do that cannot be accomplished more reliably, quickly, and cheaply with fixed-response (multiple-choice) instruments?” (Harris and Kerby 1997, p. 132). Implicit in this question is the recognition that large-scale accountability testing is a political imperative. Given this, the challenge is to identify those
situations for which performance testing represents the most valid and appropriate form of assessment.

Viewed from a different perspective, the question might be “what are the ramifications of not including performance components in large-scale testing in vocational education”? In the Massachusetts Technology Education assessment, the development team concluded that all but 1 percent of the standards could be suitably evaluated using a large-scale written assessment. Faced with the argument that authentic assessments might better reflect the hands-on nature of technology education (and thus result in better test scores), one state official responded that the “development committee will continue to explore, identify, and evaluate content that can be included in a written model” (Bouvier and Corley 1999, p. 29). This suggests, perhaps unintentionally, that the nature of the curriculum could change to accommodate traditional modes of testing.

Successful Models of Implementation

In spite of the problems, there have been some successful implementation models. The National Assessment of Educational Progress (NAEP), conducted by the U.S. Department of Education for over 29 years, provides one model for carrying out large-scale performance assessments. NAEP tests were first administered on a statewide basis in 1990, and since that time most states have voluntarily participated. In 1997, the NAEP included a small-scale operational assessment of performance in the visual and performing arts, in addition to the main assessment areas of reading, writing and civics. Recent NAEP tests reflect the trend toward authentic assessment that was begun in 1992. Reading and writing test items include a large proportion of constructed-response questions. The civics assessment items also reflect this trend. The operational arts assessments used in 1997 required the students to create, perform, and/or interpret works within the discipline (i.e., art, music, theater, or dance). Student “responses” were recorded via videotape, audiotape, or photograph (Calderone, King, and Horkay 1997).

The sheer numbers of test items and scorers needed to process student responses is daunting, and the process used provides valuable insights into how large-scale performance assessments should be carried out. For example, in the 1996 NAEP, nearly 9 million constructed responses in mathematics and science were scored by a total of 675 scorers, with an elapsed scoring time of only 12.5 weeks (ibid.). A high level of reliability in scoring was achieved through the following steps:

- The development of focused, explicit scoring guides that match the assessment frameworks;
- Recruitment and rigorous training of qualified scorers, including post-training qualifying tests;
- The use of a digital image processing and scoring system that allows all responses to a particular exercise to be scored continuously until done, thus enhancing validity and reliability of scorer judgments;
• Monitoring scorer consistency by “backreading” approximately 10 percent of each scorer’s ratings, and calibrating scores to be sure that scorer drift (the tendency to grade an item higher or lower over time) is minimized;
• Checking for interrater reliability to ensure consistent ratings; and
• Keeping careful documentation of the entire process.

Historically, vocational educators have relied on performance assessments at the classroom level. Four vendors have created assessment tools for vocational education on a national level. Although their use and influence remain relatively small, they provide information regarding the trends in assessment on the national level for vocational education.

**Work Keys.** Work Keys is a system developed by ACT for teaching employability skills and generic workplace skills. All of the Work Keys tests emphasize workplace application of skills rather than academic applications. Work Keys materials include tests suitable for large-scale, high-stakes testing, along with other reporting tools.

**V-TECS.** The Vocational-Technical Consortium of States (V-TECS), founded in 1973, has as its goal the promotion of competency-based vocational education. Beginning in 1986, V-TECS created banks of test items for members to use in constructing their own competency-based tests. The test banks include both written and performance-based items. The V-TECS materials are readily available and frequently modified to fit local needs, and thus do not represent secure tools for large-scale assessments (Office of Technology Assessment 1994).

**NOCTI.** The National Occupational Competency Testing Institute (NOCTI) began developing competency tests for vocational students in the late 1970s. Since that time the organization has created, with its member states, over 70 Student Occupational Competency Achievement Testing (SOCAT) exams. The SOCAT tests have both a written and a performance component, tied to the competencies required of entry-level workers in the respective fields for which tests have been developed. Performance tests are supposed to be judged by industry representatives, who examine both the process and the product. “Although NOCTI has traditionally discouraged the use of the written tests alone, in 1992 the organization began making the written test available for pretesting because of accelerated interest in using it to fulfill Perkins requirements” (ibid., p. 79).

**C-TAP.** A program known as the Career-Technical Assessment Project (C-TAP) was developed for the state of California by the Far West Laboratory for Education Research and Development. Within occupational clusters, students will be certified job-ready through a series of cumulative and administered assessments. These cumulative assessments include supervised practical experience, an assessment project, and a portfolio of work. The administered assessments consist of structured exercises given to students at a certain time, and include project presentations, written scenarios focusing on solving a technical problem within the vocational area, and an on-demand test (ibid.).
Suggestions for Implementation

“The adequacy of the amount of time allowed for development, introduction, and institutionalization of assessment reform can have a dramatic impact on a state’s ability to sustain its reform efforts” (Khattri, Reeve, and Kane 1998, p. 74). Unfortunately, when performance assessment measures are introduced into the political realm, the pressure to show quick results is greater. A low level of involvement on the part of teachers in the development and implementation processes does impede the acceptance of changes in teaching practice (ibid.).

Several policy implications for adoption of large-scale performance assessments have been identified by Khattri, Reeve, and Kane (1998):

- Clearly state the primary purpose of the assessment system.
- Match the format of the system with the purpose.
- Coordinate assessment reform with other elements of education reform and with other testing requirements.
- Articulate in clear and simple terms the content and performance standards the assessment system is intended to measure.
- Institute procedures to ensure the technical quality and fairness of the assessment system.
- Design a system that contains a mix of different types of performance assessment tasks and procedures, to obtain a comprehensive picture of student learning.
- Tap existing resources when developing performance-based assessments.
- Communicate to the public the purposes of, and the theory underlying, the assessment. (pp. 153-157)

Because validity issues are such a major concern with any large-scale assessment, and in particular authentic assessments, using existing resources such as the Educational Testing Service, NOCTI, and others may be the best approach for the states. A s Barton (1999) notes, the use of standardized tests for accountability purposes “without meeting standard and well-known methods of validation amounts to testing malpractice” (p. 9). Professional testing organizations, which specialize in the development of assessment tools, can serve as contractors to state-level and local education agencies. NOCTI, for example, provides customized assessments for local clients, which could be tailored to address vocational education standards adopted by a state (NOCTI, n.d.). In addition, states can look to national projects or commercial entities for assistance in delivering, and/or scoring alternative testing systems. For example, in Kentucky students are expected to complete performance tasks and submit a portfolio of best work over the course of a school year, in addition to a more traditional component to its statewide testing system (KIRIS). The state assessments, with the exception of the student portfolios, are scored by a private firm called Advanced Systems in Measurement. Maryland, which also incorporates traditional and performance components in its state-level assessments, contracts with CTB Mcmillan/ McGraw-Hill to operate its program (Rothman 1995).
Regardless of who is responsible for development of assessment tools, a diverse panel should be assembled to develop and score the assessment, to reduce internal bias. Perhaps most critically, steps must be taken to ensure that the content framework upon which the assessment is based is appropriate (Bond, Moss, and Carr 1996). This can be accomplished, in part, by linking content to national standards, where available. Up-to-date job and task analyses, which have traditionally formed the basis for content frameworks in vocational education, can provide straightforward standards upon which to base valid performance assessments.

In a climate of education reform, unfortunately, new assessment measures are sometimes introduced in an effort to bring about curricular change, and there is political pressure to show quick results. This pressure can impede the acceptance of desired changes in teaching practice, particularly when teachers have not been involved in curriculum reform efforts, or when they are given inadequate time and training to make the necessary changes in teaching practice (Khattri, Reeve, and Kane 1998). Some educators also worry that moving too rapidly toward adoption of state performance assessments might backfire. Determining where and when both traditional and performance assessments can most effectively be used is more important than advocating for one type versus the other (O’Neil 1992).

Finally, national, state, and local assessments should be coordinated so that together they present a coherent view of student performance. In a comprehensive system, for example, various assessment strategies can be implemented. At the local level, portfolio assessment could provide data to improve instruction. At the state level, matrix sampling could be used to strengthen the local data, and could provide information for reporting purposes (Roeber 1997). State assessments can be linked with national efforts like the NAEP to provide meaningful, comparable data (Barton 1999). More than one measure should “count” if assessment data are used to make high-stakes decisions related to grade-level promotion, graduation, or teacher income. In this way, the multiple purposes of assessment can be better addressed.

Summary

Performance assessments are viewed by many as a key component of assessment reform that will, in turn, drive curricular and teaching reforms. Performance assessments can provide more authentic measures of student capability than standardized, multiple-choice tests, while at the same time encourage instructional practices that emphasize acquisition of more sophisticated thinking and process skills.

Vocational education has a long history of using criterion-referenced, standards-based measures and performance assessments. With some exceptions, these have not been used at the state or national level. What has occurred in the current climate is that accountability measures imposed by federal Perkins funding have resulted in an expanded use of written, standardized tests in vocational education.
So far, measures of academic skills have remained largely separate from vocational assessment efforts. Studies show that use of large-scale assessment data to improve instruction is, in reality, a relatively low priority.

Implementation of performance assessment measures on a large scale carries with it a host of practical and technical challenges, including issues relating to validity, generalizability of data, cost, and equity. These obstacles have led some to suggest that performance assessment measures are best taken at the classroom level, where they can provide meaningful information for use in improving instruction. Others advocate for partnerships between schools and private test-development entities, which may be better able to solve the challenges inherent in large-scale authentic assessment.

Large-scale authentic assessment tools have the potential to be useful, particularly for vocational fields where occupational skill attainment standards can be clearly identified. Proponents of this approach must garner the political support needed for adoption of more costly authentic assessments. Decision makers must also address some significant challenges before implementing high-stakes performance measures for vocational education on a large scale. If the decision is made to adopt such measures, care must be taken at all steps to ensure that the outcome achieves the stated purposes.
Large-Scale Assessments (Hoepfl)