As the preceding chapter suggests, vocational educators are likely to find more than one assessment strategy that will serve their purpose. Two important criteria for deciding which assessment to use in a particular situation are the quality of the information provided and the feasibility of the assessment process. This chapter describes these criteria and compares selected- and constructed-response alternatives in terms of quality and feasibility.

Unfortunately, it is usually not possible to maximize both quality and feasibility, so vocational educators must strike a balance between them. As assessment becomes more authentic, it also becomes more expensive to develop, to administer, and to score. In addition, greater quality usually involves greater cost and greater commitment of time. There is no simple formula for balancing these factors. Ideally, educators would establish standards for quality based on the uses to which the information was to be put, and then allocate resources appropriate for meeting those standards. In addition, they would only impose practical constraints that did not limit quality. In reality, this balancing act is more an art than a science, but we believe an understanding of the factors will lead to better decisions.

QUALITY OF ASSESSMENTS

The relative quality of the available alternatives should be a factor in selecting an assessment strategy. Concerns about quality are particularly important when assessments are used to make critical decisions, such as certifying individual skill mastery or rewarding successful training programs. Vocational educators face such decisions
regularly, so it is important that they understand something about the technical quality of assessments.

The quality of an assessment can be judged in terms of three questions:

- How accurate is the information?
- How confident can one be in the conclusions drawn about students or programs?
- Is the assessment fair to all students who take it?

These questions respectively correspond to the psychometric concepts of reliability, validity, and fairness. Given the present state of the art in alternative assessment, all approaches do not provide equally accurate information, support desired interpretations equally well, or provide all students with equivalent and fair challenges. Table 7 summarizes some of the quality differences between selected- and constructed-response measures, which are elaborated below.

**Table 7**

**Quality of Information of Selected- and Constructed-Response Measures**

<table>
<thead>
<tr>
<th>Dimension of Quality</th>
<th>Selected Response</th>
<th>Constructed Response</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reliability</strong></td>
<td>Automatic scoring is essentially error free</td>
<td>Rating process can increase error</td>
</tr>
<tr>
<td></td>
<td>Many responses per topic increases consistency of score</td>
<td>Fewer responses per topic reduces consistency of score</td>
</tr>
<tr>
<td></td>
<td>Strong theoretical basis for measuring reliability</td>
<td>Greater between-task variability in student performance</td>
</tr>
<tr>
<td><strong>Validity</strong></td>
<td>Large inferences from item to occupational behavior</td>
<td>Greater match between assessment task and real-world demands</td>
</tr>
<tr>
<td></td>
<td>Variation in administration conditions can complicate interpretation of results</td>
<td></td>
</tr>
<tr>
<td><strong>Fairness</strong></td>
<td>Quantitative techniques help identify potential unfairness</td>
<td>May have greater fairness because tasks are more authentic</td>
</tr>
</tbody>
</table>
Reliability

There are no perfect measuring tools in science, in the kitchen, or in education, so people who use tools to measure things need to know how much error there is likely to be in the information they receive. Reliability is a numerical index of the degree to which an individual measurement (such as blood pressure, volume of liquid, or a test score) is free from error. There are a number of ways to determine whether a household or scientific measurement is accurate: repeat it with the same tool or a comparable tool, have someone else make the measurement, etc. Similar techniques are used to assess the reliability of educational measurements. Researchers try to answer the following questions:

- If the assessment were given again, would the same results be produced (test-retest reliability)?
- If students responded to a highly similar assessment, would the same results be produced (parallel-forms reliability)?
- If student responses were read by a different group of experts, would the same scores be assigned (rater reliability)?

Each of these three factors may contribute error to the student’s score. Moreover, the errors are additive, i.e., the overall reliability of the assessment is reduced by each one. A reliable assessment is one in which scores are consistent across time, specific questions, readers, and other circumstances.

Reliability can be estimated mathematically on a scale from zero to one, with one representing the highest possible reliability. One form of selected-response assessment, commercial multiple-choice tests, usually produces score reliabilities of 0.80 and above using methods such as test-retest or parallel forms. The acceptable standard may be higher (0.90 or more) when tests are used for important decisions. Commercial tests employ a number of techniques to achieve these high levels of accuracy—e.g., they use selected-response options that control the range of possible answers, they have the results scored by machines, they use test development principles drawn from years of experience, and they include many questions about a given topic. This last technique increases the amount of information provided by the assessment in a given amount of time and reduces inconsistenc-
cies related to the specific questions asked. The Oklahoma assessments follow this model, and their reliability is quite high.

It is difficult to achieve equally high levels of reliability with constructed-response assessments. All of the sources of error mentioned above reduce the consistency of scores from performance assessments. For example, scoring introduces errors not present with selected-response measures. Rather than having an answer sheet being scored with almost perfect accuracy by a machine, human raters review essays, science projects, or pieces in a portfolio and assign scores on one or more dimensions. Both the C-TAP and Kentucky portfolios require expert readers to review the material and assign scores using a general rubric. The same is true for the NBPTS assessment activities. In all these cases, readers must make subjective judgments about the quality of complex student work, introducing inconsistencies that lower the accuracy of the scores.

There are methods to improve reader reliability. As educators become more familiar with alternative assessments, they are developing more accurate scoring procedures. For example, when scoring rubrics are aligned to specific tasks and carefully selected examples of student work are used as “anchors” for each score point, readers can score constructed-response questions with a high degree of consistency. Reports of high interreader consistency (above 0.80) are becoming more common. However, reader reliability is only one source of error in scores.

A second reason constructed-response assessments are less reliable than selected-response tests is that students do not perform as consistently on them. Research in a number of fields has found that a student’s responses vary more from task to task on constructed-response measures than on selected-response measures. As the demands of the task increase (in terms of complexity, breadth, integratedness, or any number of factors), consistency of performance (and therefore reliability) declines. Thus, for example, there may be differences between two pieces of cabinetry produced by the same student, which means that his or her scores on joinery might differ depending on which piece is rated.

Furthermore, because constructed responses are more complex and time-consuming than selected responses, less information can be
gathered in a given amount of time. For example, Kentucky performance events take a full class period and produce only one or two pieces of information about each participant. Similarly, students receive only a handful of scores on the Kentucky portfolios, which reflect many hours of work. With fewer pieces of information and greater variation between pieces, the judgment of skill or ability will be less accurate.

One consequence of the variation in performance is that more tasks are needed to produce a reliable score (Linn, 1993). Shavelson, Gao, and Baxter (1993) analyzed three different performance assessments (in science and mathematics) and reported that students would need to complete eight tasks on one of them, fifteen on another, and twenty-three on the remaining one to produce scores with acceptable reliability. Baker (1992) found that students would have to complete six or more history tasks (in which students read primary source documents and write essays explaining the important issues) to produce a score with reliability greater than 0.70. These results, and similar ones summarized by Linn (1993), mean that alternative assessments require considerably more time than selected-response tests to produce equally reliable scores. This translates into more development time, more classroom time, and greater cost (a problem addressed later in this chapter, in the feasibility discussion).

**Validity**

Scores can be accurate in the sense of being reliable and yet be used to draw wrong inferences. This is a problem of validity. For example, a student who knows how to solve applied mathematical problems may perform poorly on a written test of word problems because of reading difficulties. The test score may be reliable (i.e., the student consistently makes mistakes on written word problems), but to say the score means that the student does not know how to solve word problems in general would be incorrect.

An inference from a score is said to be valid if it is justified. Whereas reliability is a feature of the measure, validity is a feature of the way scores are interpreted by users. Consequently, assessments that are valid for one purpose may not be valid for another. For example, one might give a person studying to be a medical records clerk a multiple-choice test of spelling, grammar, and syntax to determine his or
her ability to identify errors in textual material. However, the same test might not provide a good measure of the student’s ability to write grammatically correct information on a record.

One of the primary motivations for adopting alternative assessments is to increase the validity of the inferences by making the assessment tasks more like the real-world activities the tests are supposed to reflect (Linn, Baker, and Dunbar, 1992). Selected-response measures constrain the assessment to a rigid format, which can narrow the types of skills that are measured. Constructed-response assessments present students with tasks that are more “authentic,” i.e., that match more closely the activities performed in practice. The hope is that the resulting scores will provide a better measure of the domain of interest than will those on a multiple-choice test. The Laborers-AGC environmental performance assessment duplicates conditions of the job, and success on this assessment is thought to be highly predictive of success on the job.

There are several approaches that can be used to help establish the validity of an assessment for a particular purpose:

- Experts can review the content of the assessment and confirm that it is measuring the desired skills or behaviors (content validity).
- Results can be compared with performance in a work setting, either at the same time or in the future (concurrent validity or predictive validity). If people who do well on the assessment outperform those who do poorly on it, greater confidence can be placed in the assessment’s ability to indicate job preparation.
- Results can be compared with those from other high-quality assessments of similar and dissimilar skills (construct validity). If the pattern of responses makes sense (i.e., assessments of similar skills produce similar results while assessments of different skills do not), greater confidence can be placed in the new assessment’s ability to measure the skills it is supposed to be measuring.
- Consequences from using the assessment can be examined to shed light on the meaning of the results (consequential validity).
The first approach, content validity, is quite commonly used as the initial step in building a case about the interpretation of assessment results. If the tasks to be performed are identical to tasks on the job, as they are in the case of Laborers-AGC, content validity may be adequate to satisfy the users of the information. The government is satisfied that mastery of the AGC performance tasks produces competent workers. Similarly, the standards that underlie the NBPTS assessments were developed by committees of experts who reached consensus on the critical features of accomplished teaching. Extensive professional review forms the basis for the NBPTS standards, the appropriateness of the specific tasks, and the passing scores.

The second approach, concurrent validity (same-time comparison) or predictive validity (future comparison), is based on the idea that a meaningful score will be positively related to real performance. For example, after vocational educators in Oklahoma determined that scores on multiple-choice tests were as good at predicting future job performance as scores on scenarios, which are lengthier, they deleted the scenarios from the assessment program. This saved time and expense without reducing the validity of the scores for their intended purpose.

The third approach, construct validity, may be the appropriate technique to use when the constructs being measured are complex and hard to define and when successful performance is a matter of judgment. Researchers evaluating KIRIS compared the trends in accountability scores with the trends in scores on the National Assessment of Educational Progress (NAEP) state assessment. They were looking for evidence that KIRIS scores reflect real changes in student achievement rather than unrelated factors such as test familiarization.

The fourth approach, consequential validity, examines the effects of the assessment on practice and uses this information in interpreting scores (Messick, 1989). For example, Shepard and Dougherty (1991) found that the use of multiple-choice tests in high-stakes situations led to an undesirable narrowing of the curriculum. Students began spending more time on isolated facts and procedures and less time on conceptual learning. The instructional domain was reduced, which affected the proper interpretation of the test scores. The po-
tential for distortion of classroom activities is just as great for alternative assessment as for selected-response tests (Linn, Baker, and Dunbar, 1992). It is incumbent on the users of assessment results to investigate carefully the broader effects of the assessment system before interpreting the results.

Alternative assessments present an additional validity challenge because they often have unstandardized components. More traditional, selected-response assessments dictate both the form of the test and the procedures for administration. These controls are designed to ensure that everyone has the same opportunity to perform and no one has access to special assistance. Similar standardization is possible for some constructed-response measures, such as performance tasks, but other alternatives are inherently unstandardized. Variations in the content of the assessment or the conditions under which the assessment is administered make it more difficult to interpret the results. For example, senior projects and portfolios have built-in flexibility with respect to the conditions of performance and the content of the assessment. One student’s C-TAP portfolio may contain different work artifacts and experiences than another’s. Furthermore, students may not perform their work under the same conditions. In particular, teachers may offer different levels of support, making it hard to know how much of the work that went into the final product was actually the student’s (Gearhart, Herman, Baker, and Whittaker, 1993). Even though it is possible to score individual portfolios using a common rubric, there may be questions about the meaning of the scores if they are based on different products done under different conditions.

**Fairness**

Users of assessments must take into account the fact that irrelevant factors, such as family background and experience, may affect the scores of certain students. Assessments are unfair, or biased, if students who are otherwise equal in the skill being measured perform differently on a particular question because of experience or knowledge not related to the underlying skill. It is not easy to detect possible bias. The most commonly used techniques involve careful review of measures by committees trained to be sensitive to factors that might affect particular groups of students. Expert reviews were
used by NBPTS to ensure that its certification system was fair to teachers regardless of their population group or the socioeconomic status of their students. Complicated statistical procedures can be used to determine whether test items are biased, but the results of these procedures are often confusing. Researchers have often found it difficult to understand what features of items lead to the differences they detect.

Many advocates of alternative assessments believe that these techniques, compared to selected-response measures, are more equitable to all groups because they involve more complete tasks and permit students to address the tasks in ways that are meaningful to them. However, there has been very little rigorous research on the fairness of alternative assessments. At least one study found evidence of unfairness of an unanticipated sort: minority students made poorer selections of pieces to include in their writing portfolios than did nonminority students (LeMahieu, 1993). If vocational educators are going to assess students from diverse backgrounds, they need to be sensitive to potential unfairness in the measures they select.

**FEASIBILITY OF ASSESSMENTS**

Practical considerations also play an important role in deciding what form of assessment to use. Selected-response tests are a model of efficiency, whereas alternative assessments can be more difficult to develop, more time-consuming to administer, and more troublesome to score, and can yield results that are more difficult to explain. That is not to say that alternative assessments lack positive features, but potential users need to be concerned about feasibility issues such as cost, time commitments, complexity, and credibility to key stakeholders. These features are summarized in Table 8 and discussed below.

**Cost**

In general, alternative assessments are more expensive to develop, administer, and score than are selected-response tests. The U.S. Congress, Office of Technology Assessment (1992) estimates that performance assessments are three to ten times as expensive as
Using Alternative Assessments in Vocational Education

Table 8
Feasibility of Selected- and Constructed-Response Measures

<table>
<thead>
<tr>
<th>Dimension of Feasibility</th>
<th>Selected Response</th>
<th>Constructed Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>Relatively inexpensive to develop, administer, and score</td>
<td>More expensive to develop, administer, and score</td>
</tr>
<tr>
<td></td>
<td>Teachers benefit from participation in scoring</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Efficient use of class time</td>
<td>Additional class time consumed</td>
</tr>
<tr>
<td></td>
<td>Few demands on teacher preparation time</td>
<td>More teacher preparation time needed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If embedded, may not consume class time</td>
</tr>
<tr>
<td>Complexity</td>
<td>Relatively easy for developers and users</td>
<td>May require special skills to develop</td>
</tr>
<tr>
<td></td>
<td></td>
<td>May need special materials to administer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficult judgments make scoring difficult</td>
</tr>
<tr>
<td>Credibility</td>
<td>Familiar and well known</td>
<td>Growing popularity among educators</td>
</tr>
<tr>
<td></td>
<td>Higher reliability leads to greater confidence</td>
<td>Unfamiliar to community members</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Credibility with employers</td>
</tr>
</tbody>
</table>

multiple-choice assessments on a per-student basis, and more recent evidence suggests that in some cases these estimates may be low (Stecher and Klein, in press). For example, the prorated cost of one class period of multiple-choice science assessment is less than $1 per student. California spent about twice this amount to conduct its 1993 statewide assessment in science, which involved about one class period of hands-on activity (Comfort, 1995). Doolittle (1995) reported total costs for the State Collaborative on Assessment and Student Standards (SCASS) science tests of $11 to $14 per student for a similar amount of testing time. Stecher and Klein (in press) found the cost of one period of hands-on science assessment involving equipment and materials to be over $30 per pupil.

A number of factors contribute to the added cost of performance assessments. The tasks themselves often are more complex than selected-response items, so they require more time to draft, pilot, and
revise before final versions are ready. Performance assessments can also be more difficult to administer, particularly if they involve the use of equipment, the collection of work products over time, or repeated review and revision of student responses. The additional costs occur in most cases because teachers have to be trained to follow more complicated administrative procedures. Even portfolios, which would appear to entail very little in the way of added administrative requirements, are not without added burden. Teachers in Vermont reported spending three hours per month of class time just organizing and managing portfolios (Koretz, Stecher, Klein, and McCaffrey, 1994). Some assessments use external administrators, rather than teachers, as a way to guarantee comparability of administration, which also increases costs. (This is the model NAEP uses.)

In addition, student responses must be scored by hand rather than by machine, which is probably the most expensive part of the process. Scoring alternative assessments can be many times more expensive than scoring selected-response tests. For example, it costs pennies per student per class period to score multiple-choice tests and produce detailed score reports. By contrast, it costs dollars per class period per student to score essays, performance tasks, and portfolios and to produce one or two scores per student. Vermont teachers reported spending five hours per month of their own time scoring student portfolios during the year (Koretz et al., 1993). The Vermont Department of Education spent about $13 per student to score final student portfolios at the end of the year (Koretz, Stecher, Klein, and McCaffrey, 1994). Hardy (1995) reports scoring costs ranging from about $1.50 to $6 per student for performance assessments (depending on the length of the answer and whether it was scored once or twice). Commercial publishers charge about $5 per student for scoring writing assessments and reporting a single score, either holistic or analytic (Hoover and Bray, 1995). Stecher (1995) reports similar costs for scoring hands-on science assessments. In the cases we studied, the overall cost of NBPTS certification was extremely high, in part because of the complexity of judging candidate performance.

Although the amount of research on assessment cost is quite limited, there appears to be considerable variation in cost from one form of assessment to another. For example, it appears to be far easier to develop a writing prompt than a hands-on science task. Hoover and
Bray (1995) report the total cost of developing forty writing prompts for the Iowa Writing Assessment to be $138,000 (or about $3,500 per prompt), with an additional $175,000 (or $4,000 per prompt) for scoring during the field-test phase. Stecher and Klein (in press) report spending about $70,000 on average to develop one class period of hands-on science assessment. These differing development costs are due primarily to differences in the amount of time required; the cost of the science equipment was relatively small compared to the cost of the professional time. Similar differences in scoring costs among types of assessments are reported above.

Because of the wide variation in reported costs, it is not possible to develop handy rules of thumb for the time or cost of developing alternative assessments. It appears to be the case that costs increase as the assessment becomes less constrained (i.e., more “authentic”) and more complex. Beyond that, there seems to be little agreement about how much time or money alternative assessments require. Computer-based assessments, discussed briefly in Chapter Two, may prove to be less expensive if they can be used on a large scale.

When thinking about differences in cost between alternative assessments and traditional, selected-response tests, it is important to consider differences in reliability as well. Most of the comparisons of assessment costs presented above are based on similar amounts of testing time, but not on equally reliable scores. For example, hands-on science tests were thirty times as expensive as multiple-choice science tests for one class period of assessment (Stecher and Klein, in press). However, three periods of hands-on science assessment were required to produce a student score as reliable as that from one period of multiple-choice testing. As a result, the cost ratio for equally reliable scores was ninety to one, rather than thirty to one.

Finally, it would be incorrect to consider costs without also mentioning benefits. The use of alternative assessments may have positive effects in the form of staff development that offset some of the costs. Teachers report that scoring performance assessments is an effective training activity. The process of reviewing student work and evaluating it against well-defined rubrics helps teachers develop a better appreciation for the range of student performance, weaknesses in student presentations, common misconceptions and problems encountered by students, the alignment between curriculum and as-
assessment, and other features that relate to instructional planning. In this way, the scoring experience may improve teaching and learning. If this is the case, then it becomes important to know how valuable these benefits are. For example, would they be more efficiently achieved directly through targeted staff development than indirectly through assessment development and scoring activities? To date, such questions remain unexplored in the research on alternative assessment.

**Time**

In addition to costs that must be borne directly, alternative assessments place greater time demands on administrators, teachers, and students. For example, alternative assessments usually require more class time to administer than do selected-response tests. The use of class time for assessment can have negative consequences on instruction. Scoring also commands a great deal of time. There are advantages to having teachers score their own students’ work. For example, they learn more about student performance, and there is no added cost for hiring outside scorers. However, scoring is an extremely time-consuming task, and teachers should be aware of the demands it may place on their preparation time.

When assessments are embedded in classroom instruction, such as is the case for senior projects and portfolios, the distinction between assessment time and learning time is blurred and the time problem is less troublesome. This is true of C-TAP and the KIRIS portfolios. These assessments do not place the significant additional demands on classroom time that stand-alone performance assessments do.

**Complexity**

Alternative assessments are more complex than traditional tests in a number of ways, including the situations that prompt student responses, the kinds of materials involved, the scope of the tasks, the cognitive demands placed on students, the procedures for collecting responses, and the procedures for scoring. As noted above, it is partly this complexity that makes alternative assessments more difficult to develop, administer, and score, which increases their cost. The complexity also demands more sophistication on the part of
users. For example, it can be more complicated to administer performance assessments that involve equipment and materials, than to administer pencil-and-paper tests. In the case of Laborers-AGC, the tasks can include the use of heavy machinery, hazardous materials, and dangerous working conditions. The equipment makes administration more complex and places greater demands on task administrators, who need to be specially trained to work under these circumstances. Similarly, it may take greater expertise to, say, develop good portfolio tasks or devise scoring rubrics for senior projects than to administer and score selected-response tests. The additional complexity inherent in alternative assessments may create practical problems for some educators and some educational settings. Additional training may be required, as well as additional equipment and materials, storage space, and facilities for assessment.

**Credibility**

To have any practical value, alternative assessments, like all assessments, must provide information that is credible to the people who use the results. In the case of vocational assessment, those who use the results include not only the usual educational audiences (students, teachers, and program directors), but potential employers, labor leaders, and other community members as well. If an assessment fails to meet reasonable technical standards, its credibility may decline in the eyes of some audiences. For example, Kentucky teachers still have doubts about the appropriateness of KIRIS as an accountability tool. Moreover, people who are familiar only with traditional tests may not place much trust in scores generated by such items as performance tasks, senior projects, or portfolios, even if these forms of assessment are found to be reliable and valid. Part of this discomfort may stem from unfamiliarity, a problem that should be resolvable with training. On the other hand, one of the advantages of alternative assessments is that employers and other stakeholders may give greater credibility to scores based on authentic performances than to scores from traditional tests. It appears that this has been the case for the Laborers-AGC environmental program and for the NBPTS certification program. It is true for VICA as well.