Quality and feasibility are important factors in assessment planning, but they do not always present themselves in the general ways discussed in Chapter Four. Our case studies produced examples of six other administrative considerations related to quality and feasibility that also play a significant role in assessment planning. Not all programs will need to address all of these issues, but as a set they illustrate the additional complexities that may arise. These six additional considerations are

1. Single versus multiple measures
2. High versus low stakes
3. Stand-alone versus embedded tasks
4. Standardization versus adaptability
5. Single versus multiple purposes
6. Voluntary versus mandatory participation

**SINGLE VERSUS MULTIPLE MEASURES**

There are obvious advantages to basing an assessment on a single measure, but there are reasons to prefer multiple measures as well. We saw both options in our case studies. Although the Oklahoma assessment program contains both performance assessments and standardized multiple-choice tests, Oklahoma relies on the multiple-choice test to determine whether individual students have mastered the curriculum in each vocational area. Each entrant in a particular
Using Alternative Assessments in Vocational Education

VICA competition completes just one occupational task. It might be argued that the C-TAP portfolio is also a single measure, but, in reality, it subsumes many measures. The C-TAP portfolios can contain other kinds of assessment results, such as test scores or competitive awards.

The principal advantage of single measures is efficiency (see Table 9). Oklahoma provides a good example of this. In the past, the Oklahoma state vocational testing program had two elements: multiple-choice items and realistic scenarios followed by sets of related questions. The scenarios were more complex to develop and score, and the Oklahoma Department of Education decided that the multiple-choice items did an adequate job of predicting job-related performance. As a result, the scenarios were dropped from the testing program, leading to a reduction in cost and resource demands. Similarly, VICA made a determination that a single performance event was adequate for a competition whose goals are primarily honorary. Even with this simplification, however, VICA finds it quite difficult to prepare the task specifications and scoring guides and to train the raters for a single activity per occupation. Multiple activities would be prohibitively time-consuming and expensive.

Educational researchers recommend the use of multiple measures primarily to provide the validity that comes from having alternative windows on behavior. NBPTS strongly believes in multiple measures, arguing that the job of teaching cannot be captured in a single type of assessment. Laborers-AGC uses both a multiple-choice test of knowledge and a performance test of a candidate’s ability to carry out essential job tasks. Particularly where health and safety issues

<table>
<thead>
<tr>
<th></th>
<th>Single Measure</th>
<th>Multiple Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficiency of planning,</td>
<td>Efficiency of planning, administration,</td>
<td>Includes different types of skills and</td>
</tr>
<tr>
<td>administration, and scoring</td>
<td>Reduced time and cost</td>
<td>abilities</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Greater acceptability for evaluation of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>student performance (i.e., greater</td>
</tr>
<tr>
<td></td>
<td></td>
<td>validity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Drives programs toward more comprehensive</td>
</tr>
<tr>
<td></td>
<td></td>
<td>curriculum and instruction</td>
</tr>
</tbody>
</table>

Table 9
Advantages of Single Versus Multiple Measures
are concerned, the government dictates that candidates must demonstrate both job-related knowledge and the ability to perform essential tasks. KIRIS combines three types of student achievement measures (open-ended individual tasks, group performance events and individual portfolios) and noncognitive measures (attendance, retention, etc.) into a single school accountability index. This approach is seen as providing a more complete picture of the multiple outcomes of schooling. The added quality comes at a price, however, because multiple measures are more time-consuming to develop, administer, and score.

A second advantage associated with multiple measures is that they foster the use of varied types of instruction and preparation. In high-stakes situations, single measures can lead to an undesirable narrowing of instructional content and strategies (Shepard and Dougherty, 1991; Koretz, Linn, Dunbar, and Shepard, 1991). Under the same conditions, it is better to send richer signals to teachers and to push them to prepare students to succeed in many assessment situations.

HIGH VERSUS LOW STAKES

Many aspects of an assessment are affected by the consequences attached to the use of its results. The stakes—i.e., the degree to which the outcome is associated with important rewards or penalties—can affect the character of the assessment, its credibility, the validity of scores, and the influence it has on instruction (see Table 10). Assessment may have high stakes for individuals, programs, or schools. For example, a person will be denied certain jobs in the en-

### Table 10

<table>
<thead>
<tr>
<th>Advantages of High Versus Low Stakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Stakes</td>
</tr>
<tr>
<td>Greater motivation to perform well on the assessment</td>
</tr>
<tr>
<td>Greater emphasis on teaching the skills being assessed</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
vironmental hazard industry if he or she fails to pass the relevant Laborers-AGC examination. NBPTS hopes that teachers who pass its certification assessment will earn respect, position, and eventually greater rewards because of their proven skills. KIRIS, in contrast, has no consequences for individual students but has serious consequences for schools. Continued high performance may lead to financial rewards for teachers, and continued low performance can lead to intervention by the Kentucky Department of Education.

High stakes have two major effects: they lead to greater scrutiny of results, and they influence people's behaviors in anticipation of the assessment. A licensing examination serves as a good example of the first situation, and NBPTS comes closest to that model in the cases we studied. Certification carries with it valued consequences—in at least one state, NBPTS-certified teachers receive a salary bonus. Because the certification results in a valued outcome, teachers must have confidence in the process. There have been many instances in which both licensing and employment assessments have been challenged in court by people who failed to pass and therefore were denied a benefit. For this reason, technical quality is an essential element of high-stakes assessments. When stakes are low, as in the case of most classroom testing, tests are rarely subjected to such careful review.

A consequence of the premium on technical quality is that more time must be devoted to development and more research put into measuring reliability and validity. As a result, high-stakes assessments can be far more costly than low-stakes assessments. KIRIS models some of these conditions. Since rewards are based on improvements in school accountability scores, the Kentucky Department of Education must ensure the quality of the scores. This necessitates additional research and development with their associated costs.

The second effect of high stakes, changes in people's behavior, may also affect the meaning of assessment results. In the case of VICA, individual levels of interest and anxiety affect contestants' performance, so scores may not truly reflect performance under real-world conditions. Stakes may also drive teachers to unusual behaviors, both desired and undesired. In the case of KIRIS, researchers have detected both positive changes in curriculum emphasis and negative

Neither of these negative effects is found to any substantial degree when tests have low stakes. However, low-stakes tests have their own drawbacks. Test-takers may put less effort into their performance, leading to scores that do not represent their true abilities. And teachers may not be as motivated to make sure students learn the concepts being tested as they would be if the scores “counted” for something.

**STAND-ALONE VERSUS EMBEDDED TASKS**

Traditionally, tests are distinct events that follow, but are not part of, an ongoing learning process. Assessments stand alone, administered at the culmination of a set of learning activities. For example, when Laborers-AGC environmental trainees complete a safety unit, each must demonstrate mastery of that unit by passing a performance test. Similarly, the VICA skills competitions occur as a culminating activity after classroom training.

Stand-alone assessments have both logical and practical advantages (see Table 11). They serve as markers for accumulated knowledge and skills, and they allow the assessment developer to design tasks without worrying about the specific instructional activities employed by each teacher. Focusing on content specifications independent of classroom implementation simplifies the design and administration of assessments.

**Table 11**

Advantages of Stand-Alone Versus Embedded Assessments

<table>
<thead>
<tr>
<th>Stand Alone</th>
<th>Curriculum Embedded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater flexibility in designing assessments</td>
<td>More efficient use of class time</td>
</tr>
<tr>
<td>Greater standardization across classrooms</td>
<td>Greater connection with classroom lessons</td>
</tr>
<tr>
<td>Greater simplicity of administration</td>
<td></td>
</tr>
</tbody>
</table>
An alternative approach is to embed the assessment, either by building assessment events into instructional activities as part of the curriculum or by gleaning products from meaningful learning activities to use in the assessment. (We use the term curriculum-embedded assessment to refer to both situations.) C-TAP and the portfolio component of KIRIS are the best examples of embedded assessments in the group we studied. As C-TAP students complete work internships, they capture evidence of their experience and include it in their portfolios. Similarly, Kentucky students select their best classroom products in writing and mathematics to include in their portfolios.

Embedded assessments have advantages as well. First, they may be more efficient in that they do not require teachers to set aside valuable class time for testing. However, the time efficiencies associated with this approach can be difficult to achieve in practice. For example, Kentucky students spend considerable class time preparing their portfolios—reviewing their work, selecting pieces, compiling their portfolios, and writing a reflective introduction to the work. Second, embedded assessments lead to judgments based on student products from less artificial conditions. Assessments are produced under regular school conditions and may be more typical of student work. However, since the conditions of performance are different for every classroom, it may be difficult to interpret class or school comparisons.

STANDARDIZATION VERSUS ADAPTABILITY

Most state testing programs are examples of standardized assessment systems, i.e., assessment conditions are the same in every location. Individual sites have little flexibility to change what is assessed or how it is measured. For example, the Oklahoma Department of Education develops and maintains the vocational testing program, and each institution implements the tests according to standardized procedures. Similarly, the Laborers-AGC assessment is planned centrally and administered in the same fashion at every site. NBPTS, KIRIS, and VICA are also centralized systems, but local teachers and programs have a degree of influence on selected aspects of these assessments. Kentucky teachers select the assignments that generate student work for the portfolio component of the assessments, NBPTS
applicants provide materials drawn from their teaching experience, and active local VICA chapters can contribute to the planning of the national competitions. The most familiar form of adaptable assessment is classroom testing. Teachers are responsible for designing their own tests, and they implement them to meet their own classroom needs. C-TAP is an example of an adaptable assessment. Teachers adapt the portfolio framework to reflect their local emphases.

The advantages of a standardized approach include consistency of implementation and comparability of scores (see Table 12). All teachers administer the assessment according to the same rules, so results can be compared from one site to another. For example, comparable tests are given in each Oklahoma vocational program, and students who pass the test in one school are demonstrating the same mastery of materials as those who pass the tests in another. Similarly, students take the same constructed-response tests and participate in the same performance events throughout the state of Kentucky. All VICA contestants perform the same job-related tasks, as do all candidates for Laborers-AGC certification. Standardization is particularly important when high stakes are attached to the assessment. The pressures to perform well that can lead individual teachers and students to engage in inappropriate test preparation or administration activities are lessened when procedures are clearly standardized.

As Table 12 shows, adaptability has advantages as well. Most notably, it permits assessment to be more responsive to local needs. For example, teachers can customize C-TAP portfolios to their

<table>
<thead>
<tr>
<th>Standardized</th>
<th>Adaptable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater consistency of implementation</td>
<td>Greater sense of ownership among teachers (and of motivation among students)</td>
</tr>
<tr>
<td>Greater comparability of results across sites</td>
<td>Increased relevancy to local curriculum and community</td>
</tr>
<tr>
<td></td>
<td>Increased meaningfulness to individual students</td>
</tr>
</tbody>
</table>
course and to the employer base in the surrounding neighborhood. Students who use the C-TAP portfolios in a health program thus include work samples related to health; those in a transportation program assemble work samples related to transportation. Permitting individual teachers to tailor the assessment to their local needs has other positive effects. For example, teachers may endorse the assessment more, because it can be made more relevant to their programs. This is an important rationale for using portfolios in the Vermont assessment program (Koretz, Linn, Dunbar, and Shepard, 1991).

It is possible to combine standardized and adaptable components, as is done in KIRIS. The on-demand components—open-ended questions and performance events—are the same everywhere, while the portfolios differ from class to class based on the tasks assigned by the local teachers. Similarly, the NBPTS certification process uses both flexible and standardized elements. Candidates for certification supply a videotape of their own lesson and do an analysis of their own instructional planning and decision making. These unique individual video elements are combined with common assessment center exercises, so NBPTS obtains a profile with both unique and shared elements.

**SINGLE VERSUS MULTIPLE PURPOSES**

Most of the assessment systems we reviewed were designed to serve primarily one of the three purposes discussed previously—i.e., to improve learning and instruction, to certify individual mastery, or to evaluate program success. For example, the NBPTS examinations are designed specifically to measure mastery of job-related knowledge and skills, rather than to diagnose skill deficiencies or evaluate the effectiveness of teacher education programs.

Our Oklahoma case, however, is one example of an assessment system designed to serve at least two purposes. Student scores are aggregated to the program level, where they are used by the state to monitor program effectiveness and contribute to funding decisions. In addition, scores are reported to teachers, who use them to identify weaknesses in their curriculum or instruction and to make adjustments. In some occupational areas, tests have a third purpose:
certifying students’ competencies to increase their employment prospects in an occupation.

Although it is easy to differentiate the three purposes of assessment in the abstract, in reality they are interrelated. For example, knowing something about individual student progress through a unit offers some insight into the effectiveness of the program. In this sense, assessment results do not necessarily support only one use. However, information gathered with one purpose in mind is likely to be better suited to that purpose than to another. Certainly, aggregated or sampled scores that would be adequate for measuring changes at the program level are insufficient for monitoring changes at the individual level.

Some of the advantages of single-purpose versus multiple-purpose assessments are summarized in Table 13. One major advantage of a single-purpose assessment is that it can be made highly relevant to the needs of the users, leading to efficiencies in design, administration, and reporting. For example, when designing an assessment for program evaluation, it is possible to sample students and tasks rather than having all students complete all exercises. Sampling reduces the burden on participants while still providing trustworthy aggregate information for judging overall program effectiveness. However, this approach would not be appropriate for determining individual mastery, because each student does not respond to enough items to provide a valid score.

It is possible to design multipurpose assessments in theory, but it is difficult to do so in practice. One problem is that the size of the as-

Table 13

<table>
<thead>
<tr>
<th>Advantages of Single Versus Multiple Purposes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Purpose</td>
</tr>
<tr>
<td>Greater clarity in design, administration, and reporting of information</td>
</tr>
<tr>
<td>Less conflict between competing demands</td>
</tr>
<tr>
<td>Shorter and more focused assessments</td>
</tr>
</tbody>
</table>
assessment increases as the number of purposes increases. Another problem is that different purposes can lead to conflicting demands. The Oklahoma assessment is used for both program accountability and student learning. These two uses are complementary, but there are some tensions between them that have to be resolved. For example, although Oklahoma provides common curriculum handbooks, not all teachers use them, with the result that some teachers and students do not view the test as complementary to their curriculum. They thus see the test as serving state reporting purposes but having little direct value to them.

Similarly, an assessment designed to provide individual diagnostic information needs to produce scores at a finer level of aggregation than a test that does not have to help students and teachers plan instruction. For example, one might need to know whether students have learned specific grammatical conventions as a basis for instructional planning—should the class review the use of apostrophes in the possessive form? However, in a program evaluation setting, it probably is adequate to sample a variety of grammatical conventions within a written communication task.

On the other hand, there are potential advantages to assessment systems that serve multiple purposes. If the technical difficulties can be overcome, a single assessment system is more efficient than multiple assessments in terms of resources and testing time. In the ideal case, a common broad assessment could produce a core database from which information could be extracted to address different questions. One potential advantage of a multipurpose assessment system is improved communication among stakeholders who would share common terms, references, and results. Another potential advantage is that it might generate greater support among policy makers responsible for funding the assessment. McDonnell (1994) describes differences among stakeholders’ views of assessment, and points out that coalitions can sometimes be built if policy makers believe an assessment can support multiple goals.

**VOLUNTARY VERSUS MANDATORY PARTICIPATION**

One important element in the assessments conducted by Laborers-AGC, VICA, and NBTPS is that test-takers participate voluntarily. The
alternative is to require that all eligible individuals sit for the tests, as is done in state testing programs such as KIRIS and Oklahoma.

Table 14 illustrates some of the advantages associated with voluntary versus mandatory participation. Students and teachers who participate in voluntary rather than compulsory testing programs are often more motivated to do well because they have made a commitment to the outcome. If, in their desire to be successful, they pay more attention to the tasks, focus their energies, and make more efficient use of time, the validity of the assessment results may even be increased. Voluntary participation may also increase the value of the assessment as a signaling tool because students and staff who choose to participate attend to it more. Teachers may adjust their curriculum based on student scores, and students may change their study habits. The assessment may have greater utility as a lever for reform because it is given greater credence by program participants. Increased attention may also enhance the educational value of the assessment experience itself. Finally, those who choose to participate often are more engaged in the learning experience than those whose participation is compelled.

But there are also advantages to required participation. Oklahoma, KIRIS, and C-TAP can be motivating because they are required. Teachers may attend to the content covered by the tests because all students are required to take them, although the degree of influence may be a result more of the consequences than of the level of participation (see discussion of high versus low stakes, above). Required assessments affect all participants, so whatever value is obtained accrues to everyone, not just the self-selected few. In

### Table 14

<table>
<thead>
<tr>
<th>Advantages of Voluntary Versus Mandatory Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voluntary</td>
</tr>
<tr>
<td>Greater commitment from users, increasing likelihood that</td>
</tr>
<tr>
<td>• optimum performance is elicited (validity)</td>
</tr>
<tr>
<td>• curriculum and instruction are influenced</td>
</tr>
<tr>
<td>• users learn from participation</td>
</tr>
</tbody>
</table>
addition, assessment results are more likely to be useful for comparison across program units when all students participate. This is essential if the assessment is to be used for accountability purposes.

In most instances, program developers have little control over this aspect of assessment. The program context dictates whether testing is mandatory or not. However, there are cases in which design decisions can affect the amount of testing required of individuals and the way scores are used, which indirectly affects the emotional and psychological aspects of participation. For example, some state testing programs report scores on every student, which necessitates that every student complete the full test. Other states report only aggregate scores (e.g., at the classroom, school, or district level), which permits them to use matrix or item sampling. While all students must participate, each takes far fewer items, lessening some of the negative associations that accompany extended testing programs. In some instances, e.g., in Kentucky, a sample of students is selected for participation in the performance events, reducing further the perceived burden and giving participation an aura of specialness.