3

Evaluating Student Learning

There are many ways to approach the evaluation of student learning. The characteristics of good evidence of student learning include considerations of direct and indirect methods for gathering evidence of student learning, the appropriate use of quantitative and qualitative evidence, and other methodological considerations. First, however, it is important to understand the fundamental assessment concepts of formative and summative assessment and benchmarking.

Formative and Summative Assessment

Formative assessment is ongoing assessment that is intended to improve an individual student's performance, student learning outcomes at the course or program level, or overall institutional effectiveness. By its nature, formative assessment is used internally, primarily by those responsible for teaching a course or developing a program.

Ideally, formative assessment allows a professor, professional staff member, or program director to act quickly to adjust the contents or approach of a course or program. For instance, a faculty member might revise his or her next unit after reviewing students' performance on an examination at the end of the first unit, rather than simply forging ahead with the pre-designated contents of the course. Many instructors also solicit repeated brief evaluations of their teaching, and the data gleaned from these can be used to make adjustments that may improve learning, such as the introduction of more discussion into a class.

In contrast, summative assessment occurs at the end of a unit, course, or program. The purposes of this type of assessment are to determine whether or not overall goals have been achieved and to provide information on performance for an individual student or statistics about a course or program for internal or external accountability purposes. Grades are the most common form of summative assessment.

Goals for student learning will be expressed summatively when faculty members are describing what they expect students to be able to do or what skills they expect students to achieve when they complete a course or a program or when they graduate from the institution.

Formative and summative assessment work together to improve learning. They should be central components of assessment at the course level, and where appropriate, at the program level.

Benchmarking

The term benchmarking is now common in assessment plans and conversations about assessment. Originally, benchmarking was a term used in the corporate environment to define a set of external standards against which an organization could measure itself. The organization identifies comparable, peer, or "reach" organizations and systematically compares its practices or achievements against those of the other organization.

In higher education settings, a university might use benchmarking techniques to define its comparison group—its peer institutions—and to compare its

own outcomes to theirs. This benchmarking could be based, for example, on retention rates, five-year graduation rates, admissions yield data (the number of enrollees as a function of the number of students accepted), employment and graduate school placement rates, and performance on national or professional examinations. Theoretically, any outcome for which there are data from peer institutions and programs can be compared in a benchmarking study.

Two other related forms of benchmarking are used in higher education settings. A college or university can compare itself to a national norm by reviewing the data from a published test or survey such as the National Survey of Student Engagement (NSSE). Alternatively or in addition, an institution can set for itself the goals or benchmarks that it hopes to achieve within a specified time period (e.g., to increase job placement rates from 70% to 90% in five years).

The benefit of inter-institutional comparison is that it can flag problem areas to investigate the causes of results that differ from the norm. For example, two comparable liberal arts colleges with similar selectivity, similar student preparedness, similar socioeconomic profiles for their students, and similar science curricula, may discover that proportionately more students are accepted to medical schools from one institution than from another. Further investigation may reveal that the excelling college requires a hospital internship for all of its pre-med students.

The discovery that an institution's students are below the norm on a national metric (e.g., amount of time devoted to school work outside the classroom) challenges the institution to determine the reason for this result. Similarly, an institution that sets its own internal benchmarks must design and implement a program to achieve its goals.

Before beginning to articulate goals for student learning, program faculty and leaders of institutional assessment should consider how the use of benchmarks could enhance their institution's ability to achieve its goals and whether useful measures from comparable peer institutions are available.

Direct and Indirect Methods for Assessing Student Learning

The concepts of direct and indirect methods of evaluating student learning are often confused with each other and with quantitative and qualitative forms of information. Each of these has its merits and drawbacks.

Direct and indirect methods of evaluating learning relate to whether or not the method provides evidence in the form of student products or performances. Such evidence demonstrates that *actual learning* has occurred relating to a specific content or skill. Indirect methods reveal characteristics associated with learning, but they only imply that learning has occurred. These characteristics may relate to the student, such as perceptions of student learning, or they may relate to the institution, such as graduation rates.

When a student completes a calculus problem correctly and shows her work, learning is demonstrated *directly*. When the same student describes her own calculus abilities as excellent, she is demonstrating *indirectly* that she has learned calculus. Both of these pieces of information about the student's performance are important. For excample, a student's perception that she is doing poorly in calculus when she is actually doing well would provide important information to both the student and the professor. However, indirect evidence—in this case, a perception—is less meaningful without the associated direct and tangible evidence of learning.

Figure 5 includes examples of direct and indirect measures of student learning at the course, program, and institutional levels. Many of the examples presented in Figure 5 can be used as measures of student learning at more than one level. For example, portfolios of student work and student satisfaction surveys can be used at the course, program, and institutional level, and internship performance ratings could be used at the course or program level.

Figure 5

Examples of Direct and Indirect Measures of Student Learning (Course, Program, and Institutional Levels)

	Direct Measures	Indirect Measures
Course	✶ Course and homework assignments ✶ Examinations and quizzes ✶ Standardized tests ✶ Term papers and reports ✶ Observations of field work, internship performance, service learning, or clinical experiences ✶ Research projects ✶ Class discussion participation ✶ Case study analysis ✶ Rubric (a criterion-based rating scale) scores for writing, oral presentations, and performances ✶ Artistic performances and products ✶ Grades that are based on explicit criteria related to clear learning goals	✶ Course evaluations ✶ Test blueprints (outlines of the concepts and skills covered on tests) ✶ Percent of class time spent in active learning ✶ Number of student hours spent on service learning ✶ Number of student hours spent on homework ✶ Number of student hours spent at intellectual or cultural activities related to the course ✶ Grades that are not based on explicit criteria related to clear learning goals
Program	✶ Capstone projects, senior theses, exhibits, or performances ✶ Pass rates or scores on licensure, certification, or subject area tests ✶ Student publications or conference presentations ✶ Employer and internship supervisor ratings of students' performance	✶ Focus group interviews with students, faculty members, or employers ✶ Registration or course enrollment information ✶ Department or program review data ✶ Job placement ✶ Employer or alumni surveys ✶ Student perception surveys ✶ Proportion of upper-level courses compared to the same program at other institutions ✶ Graduate school placement rates
Institutional	✶ Performance on tests of writing, critical thinking, or general knowledge ✶ Rubric (criterion-based rating scale) scores for class assignments in General Education, interdisciplinary core courses, or other courses required of all students ✶ Performance on achievement tests ✶ Explicit self-reflections on what students have learned related to institutional programs such as service learning (e.g., asking students to name the three most important things they have learned in a program).	✶ Locally-developed, commercial, or national surveys of student perceptions or self-report of activities (e.g., National Survey of Student Engagement) ✶ Transcript studies that examine patterns and trends of course selection and grading ✶ Annual reports including institutional benchmarks, such as graduation and retention rates, grade point averages of graduates, etc.

Direct Methods

Direct methods of evaluating student learning are those that provide evidence of whether or not a student has command of a specific subject or content area, can perform a certain task, exhibits a particular skill, demonstrates a certain quality in his or her work (e.g., creativity, analysis, synthesis, or objectivity), or holds a particular value. Direct methods can be used at the course level, the program level, and, theoretically, at the institutional level.

Course Level. Most familiar are direct evaluations of learning that are applied at the course level. Examinations,[3] regardless of format, are designed to be direct evaluations of student learning. Similarly, evaluations of writing samples, presentations, artistic performances, and exhibits provide direct evidence of student learning, as do evaluations of student performance in internships, research projects, field work, or service learning settings. As discussed later, grading linked to clear learning goals is a valid and useful form of direct measurement of student learning.

Program Level. At the program level, examinations also are used frequently as direct measures of student learning. Such examinations would be more comprehensive than those embedded within a course and would be designed to evaluate cumulative, aggregate, or holistic learning after the conclusion of a program or during the course of the program.

For example, a writing examination might be given after the first two years of a general education program, with the goal of determining whether students' writing was enhanced as a function of the program. Standardized tests of disciplinary content might be administered to students after they have completed all program requirements for the major (e.g., American Chemical Society examinations). Honors theses, senior theses, or senior projects are other sources of direct evidence of student learning within a program. Ratings by internship supervisors of how well interns are demonstrating key learning outcomes are important, direct program-level evidence of student learning.

Institutional Level. Direct evaluations at the institutional level are used less frequently and are much more likely to take the form of an examination. A college or university might use the Academic Profile or the Major Field Tests from the Educational Testing Service, the Collegiate Assessment of Academic Proficiency from the ACT (American College Testing) or other graduate-level examination scores to demonstrate that learning has occurred.

An institution may wish to demonstrate that certain goals expressed in its mission were achieved through exposure to the entirety of its curriculum and co-curricular experiences. For example, it may wish to show that regardless of program or major, which co-curricular activities students have participated in, and whether students were residents or commuters, they exhibit cultural sensitivity and global cultural and geographical awareness. It could design an evaluation process to determine the degree to which graduating students exhibited these qualities (e.g., a rubric for reviewing an examination or portfolio).

It may appear that such qualities are abstract and, therefore, that the measurement of learning was not direct, but in fact that is not the case. In this example, the goal was to have students learn, through curricular and co-curricular programs, to be good global citizens, broadly speaking, and the hypothetical examination was designed to measure the degree to which this goal was achieved.

General education knowledge, competencies, and skills gained across the curriculum might be evaluated over the entire student experience, whether before or after graduation.

---

3 For the purpose of clarity, the term "examination" is being used here to refer to what are commonly called quizzes, exams, or tests designed to measure whether or not a student has learned something that he or she was taught prior to its administration. The word "test" is a more generic term and can apply to any measure that may be direct or indirect, or qualitative or quantitative.

Fundamental Importance of Direct Forms of Evaluation. The power of direct assessments of student learning is that, if designed properly, they answer the most important questions:

- What did students learn as a result of an educational experience?
- To what degree did students learn?
- What did students *not* learn?

Institutional stakeholders and the public can understand easily data gleaned from direct evaluations of student learning. They can understand, for instance, that students at Institution A have higher scores on the American Chemical Society examination than students at Institution B, and those same data provide assurance that a certain level of knowledge has been acquired by students at both institutions.

Limitations and Considerations Related to Direct Forms of Evaluation. Direct assessments, however, do not tell the whole story of student learning. There are two potential problems with using only direct assessments of student learning. The first problem relates only to direct methods, and the second pertains to both direct and indirect methods.

Direct assessments of student learning, while providing evidence of *what* the student has learned, provide no evidence as to why the student has learned or *why* he or she has not learned. The "why" of student learning is especially important when students have not learned, because one of the primary goals of assessment is to make future learning experiences more effective.

If students perform poorly on a mathematics exam, for instance, it is important for the instructor to know whether the students' performance resulted from not having learned the material or from having learned the material but also experiencing anxiety during the examination. Other data are needed
to answer this question.

It is important to consider that even direct forms of evaluation do not necessarily indicate whether "value-added" learning has occurred. The Commission does not require that its member institutions demonstrate value-added learning, only that the institution's learning outcomes are consistent with its goals.

In and of themselves, direct forms of evaluation do not always provide evidence that the targeted learning goal was achieved within the context of a course, a program, or an entire college education, or whether the demonstration of the learning goal was influenced by or a product of prior learning or even the result of innate abilities. If an institution or faculty members in a program are concerned about demonstrating that the learning occurred in a particular context, then care should be taken to design aspects of the assessment program to tap "value-added" learning.

At the course level, the contrast between value-added demonstrations of student learning and absolute levels of student learning is rarely meaningful. One can assume, for instance, that knowledge of college-level organic chemistry, elementary school teaching techniques, or Spinoza's philosophy was acquired within the context of the course specifically designed to teach that knowledge. The same reasoning applies to the program level; students are likely to have acquired the skills and knowledge specific to their programs while taking courses within the program.

At the institutional level, the distinction between student knowledge that was acquired before the student arrived at the institution and what he or she learned while in attending the institution may be a more salient one. Some institutions may want to demonstrate that the education they provide has had a fundamental effect on students' lives—i.e., changed them in a way that would not have occurred if the student did not attend college or attended a different type of college.

One college, for instance, may want to demonstrate that a personal atmosphere that encourages faculty-student mentoring relationships results in better preparation for acceptance to graduate school than a student might otherwise receive at a different type of institution. Another may want to demonstrate that it prepares its students for the real world in a way that a different college experience cannot. Yet another might use assessment data to show that students have dramatically increased their job marketability or their chances of graduating by attending the college.

If institutions seek to demonstrate such accomplishments, it is important to consider whether the assessment design truly demonstrates value-added learning rather than some other phenomenon. For instance, students entering an institution with very high SAT writing scores are likely to write well after they have been exposed to the college's General Education program. In other words, to the extent that high scores of graduating students on tests of writing skills reflect pre-college expertise, those scores reflect the effect(s) of one or more "inputs" but are not necessarily value-added.

Value-added gains can be useful in assuring the college community and the public that higher education provides cognitive, affective, and social growth beyond the levels the students had attained when they entered college. However, devoting too much attention to creating an assessment design that rules out other causes for student learning can take the focus away from the most important question: Have students who graduate from this college or university learned what the institution hoped they would learn?

Indirect Methods

Indirect methods of evaluating student learning involve data that are *related to* the act of learning, such as factors that predict or mediate learning or perceptions about learning but do not reflect learning itself. Indirect evidence often is acquired through the use of self-report format surveys, questionnaires, and interviews. Indirect evidence also is provided in the form of "demographic" statistics about the student population of an institution, such as overall GPA, student retention rates, graduation rates, and job placement rates. Qualitative information about graduates, such as names of employers, graduate schools attended, and alumni achievements are also common forms of indirect evidence.

Course Level. The most familiar indirect assessment of student learning is the course and/or teaching evaluation given at the end of the semester. These instruments usually have a quantitative section in a Likert (numerically-scaled) format, in which the student rates the quality of teaching and of the course, as well as a narrative section in which the student offers additional qualitative comments.

An instructor who regularly reviews his or her teaching evaluations and who changes the course as a result of those evaluations is engaging in improvement based on hypotheses derived from the indirect assessment of student learning. The same instructor can use this indirect method in conjunction with direct methods to improve student learning in the course.

For example, students might use the narrative portion of the evaluation to request more time for class discussion and might give the professor only moderate ratings for "engagement with the course material." The instructor decides to introduce more discussion into his or her class and subsequently students praise the use of discussion and give high ratings for the instructor's "engagement with course material." Most importantly, the instructor notices that student grades on quizzes or exams and work on assignments are higher in the semester after he made the change. This simple illustration of how indirect evidence can be used in conjunction with direct evidence can be applied in more complicated situations.

Program Level. At the program level, student satisfaction surveys may reveal that students want more one-on-one contact with faculty members. Upon reflection, faculty members may decide to offer more independent study experiences; consequently, scores on Graduate Record Examination subject area exams improve (direct evidence of student learning), as do graduate school admission rates (indirect evidence of student learning).

Institutional Level. Indirect means of evaluating student learning are important at the institutional level as well. National surveys, such as the National Survey of Student Engagement (NSSE), provide benchmarking opportunities for the institutions to gauge the qualities of their student populations relative to other institutions so that they can determine whether changes in programming affect students' perceptions and behavior inside and outside the classroom. Ultimately, such assessments can affect performance in the classroom.

For instance, if an institution finds that its students spend less time studying than the national average for study time, it might introduce curricular changes that link student evaluation (i.e., grades)

more directly to the amount of time studied, perhaps by providing assignments that demand more out-of-class time and by using class examinations which test areas that are not learned simply by attending class. The greater engagement that these changes create might serve to improve student performance on direct measures of student learning.

Indirect evidence often focuses on the learning *process* and the learning *environment.* Alexander Astin's input-environment-output assessment model, based on research from the past several decades (e.g., Astin, 1991; Chickering & Gamson, 1991; Pascarella & Terenzini, 1991) indicates that students learn most effectively when, in general, they are engaged in the learning process and they can see a connection among course goals, course content, and evaluation.[4]

The extent to which these inputs and processes exist may support the inference that student learning is taking place. Each of these discoveries about student learning was gained through indirect methods of assessment, such as surveys of student perceptions and opinions. The results of these surveys then were correlated with actual student learning outcomes (measured directly), demonstrating that when the existence of specified inputs and processes correlates with student learning.

Limitations and Considerations Related to Indirect Methods of Evaluation. The most important limitation of indirect methods is that they do not evaluate student learning *per se*, and therefore should not be the only means of assessing outcomes.

As with direct measures of student learning, it is important to consider that indirect measures do not necessarily imply that value-added learning has occurred. Students who express indifference to co-curricular activities after their first year may be expressing an indifference that is the result of dissatisfaction with campus programs, or they may have arrived on campus disinclined to spend time on anything but course-related work.

As noted above, the Commission does not require proof of value-added student learning. Nevertheless, an institution should consider whether value-added data are necessary to demonstrate that it fulfills its own mission. If so, it should ensure that data collection procedures warrant conclusions about the effectiveness of programs in teaching students.

Quantitative vs. Qualitative Evidence

In every example of direct and indirect assessment cited above, the evaluation of student learning could provide either qualitative or quantitative information. Both qualitative and quantitative information are valuable forms of evidence about student outcomes.

Quantitative evidence consists of data that are represented numerically. For instance, performance on a test or responses to a questionnaire may be scored so that a number represents the degree to which an individual performed or agreed/disagreed with a certain concept. Because quantitative data are expressed in numbers, they can be compared directly or subjected to statistical analysis, and they can enable the researcher to make certain assumptions when comparing one data point to another. Quantitative data also may peremit one to express numerically meaningful changes in performance (given certain conditions). One may claim, for instance, that a change in a test score from 50 to 60 represents a 10-point or a 20 percent gain in an individual's performance, expressed as a percentage of his or her original score. Quantitative data, therefore, are valued for the ease with which calculations and comparisons can be made, and for the easily understandable representations of performance that they produce.

Qualitative evidence typically comes in two forms. The first form involves simple categorization of individuals into discrete groups (e.g., employed or unemployed; participates in athletics or does not participate in athletics). The second form of qualitative evidence is data expressed in prose or narrative. A question is asked of an individual and

4 See Chapter 5 for a further discussion of this topic. See also Figure 21 for a list of conditions under which students learn most effectively.

he or she responds in a free-form manner, expressing, for instance, an idea, opinion, or evaluation. Because of their non-numerical nature, qualitative data cannot be subjected directly to statistical analyses, nor can easy direct comparisons be made without engaging in an intervening process to categorize or interpret the data. Qualitative data, however, can be "richer" than quantitative data, because they provide a more extensive variety of information related to a particular learning goal. Many faculty members, for instance, use the numerical scores (quantitative data) from their teaching evaluations to make overall judgments of their own performance, but they value the qualitative, narrative comments from students as more useful in revealing students' personal perceptions of a course.

A common misconception is that qualitative assessments are not as reliable, valid, or objective as quantitative ones. This is not necessarily the case. There are well-designed and statistically reliable means of interpreting and analyzing qualitative data and numerous resources for learning to use qualitative methods (see Silverman, 2001; Maxwell, 1996). For example, an instructor might assess the same learning goals using a multiple-choice test or an essay test. Similarly, an instructor might grade a senior project presentation quantitatively with a standard set of evaluation criteria (i.e., a rubric). Alternatively, he or she might provide the student with a prose evaluation, in a non-scaled format, citing the strengths and weaknesses of the presentation. However, it is best if this evaluation is organized around a standard set of criteria that were shared with the student beforehand.

A student survey designed to gather information on student satisfaction may elicit data that are quantitative (i.e., "On a scale of 1 to 7, how satisfied are you with the quality of advising?") or qualitative ("How would you describe your experience with academic advising?"). Similarly, employers asked to assess the strengths and weaknesses of alumni may be asked to assign "scores" to, or to describe, alumni characteristics.

Most beginning assessment initiatives are likely to rely more heavily on quantitative, rather than qualitative, forms of assessment for several reasons. Quantitative data are easier to collect and are in the form of a readily-analyzable numeric score. In contrast, qualitative data must be sorted, categorized, and interpreted (most often by humans rather than by computer programs) before a final judgment can occur. Methods of ensuring the reliability of qualitative data are time-consuming. For instance, to ensure that portfolio assessment is reliable, at least two raters are used to review each portfolio, providing a form of "inter-rater" reliability. Focus groups, another commonly used form of qualitative data collection, require large investments of time to gather data from comparatively few students.

A good use of qualitative evaluation is to help develop quantitative evaluation criteria (rubrics). For instance, one might conduct focus groups for the purpose of designing questions for a satisfaction questionnaire or use a scoring rubric for portfolios to determine what characteristics of students' writing might be evaluated.

For assessing student learning, *Characteristics* encourages the use of multiple approaches—both quantitative and qualitative—but it does not *require* the use of both approaches (see Standard 14). Institutions and faculty members in different programs should be thoughtful about which approach, or combination of approaches, best suits the student outcomes that are being assessed in each unique situation.

Other Methodological Considerations

Some of the other methodological considerations often raised with regard to assessment include reliability and validity; pretests, posttests, and longitudinal design; the role of grades, self-report measures, and statistical versus practical significance.

Validity and Reliability

In general, the terms "validity" and "reliability" refer to the extent to which assessment tools and methods provide accurate, fair, and useful information. Both concepts are important factors in choosing standardized assessment instruments and should be considered seriously when developing locally-created instruments for summative assessment.

Validity refers to the integrity of the instrument. Does the instrument measure what it was designed to measure, or does it actually measure something else? An instrument designed to assess student sensitivity to the cultural norms of others, for instance, may actually be measuring a student's sensitivity to detecting those responses desired by the professor or the institution that values such sensitivity. Obviously, the instrument would not provide a valid assessment of cultural sensitivity.

Three forms of validity are especially relevant to assessing student outcomes. An instrument with "construct validity" adequately taps the "construct" or conceptual framework that it is designed to measure because its questions have been developed specifically for that purpose. The test of cultural sensitivity described above lacks construct validity because it assesses student perceptions of faculty beliefs, not cultural sensitivity.

Content validity, and in particular "face validity," refers to the content and structure of an evaluation instrument: On the face of it, does it appear to assess what it is designed to assess (Gall, Borg & Gall, 1998). The cultural sensitivity instrument described above may appear to have face validity—the questions appear to be about cultural sensitivity—even though it lacks construct validity. In general, however, the content and structure of an instrument should make sense to those who are using it. Several methods are employed by test designers to ensure that instruments have both content and face validity.

A third important form of validity is referred to as "concurrent" or "criterion validity." Criterion validity means that a test or assessment instrument will yield results that are similar to those of other instruments designed to assess the same outcome. Two tests of college mathematics ability, for instance, should yield similar results when administered to the same students; if one measure of student satisfaction demonstrates that students are very satisfied, another should as well. This result also could demonstrate "predictive validity" if the strength of the correlation between the two measures was great. A test or other evaluation instrument with good criterion validity also will predict performance on other measures of constructs that should be related. For instance, student satisfaction should predict retention, and high scores on a test of ethical decision-making should predict ethical behavior. Additional concepts and examples related to reliability and validity are discussed in the section below entitled, "Key questions for choosing assessment instruments."

Reliability refers to the consistency of results for a test or assessment instrument over repeated administrations to the same individuals. For instance, an aptitude test for mechanical engineering, given twice to the same person, should yield similar results each time. Otherwise, it fails in its purpose of providing an accurate prediction of future success in mechanical engineering.

Reliability is established during the development of the test, when special populations are recruited to take the test more than once, before the test is used for its intended purpose. Reliability information about standardized tests is presented in the form of statistical correlations (which should be very high) among repeated administrations of the test in the same population.

The concepts of validity and reliability apply primarily to summative assessment, and not as directly to formative assessment, because instructor-created examinations and measures usually only exhibit "face validity," not the other forms of validity discussed here, and they are not usually subjected to rigorous pre-administration tests of reliability.

Pretests, Posttests, and Longitudinal Designs

A common misconception is that, in order to make any claims about "value-added" changes in student learning, one must use a pretest-posttest format. For instance, in order to demonstrate that a general education program has improved the writing skills of students, it appears that it would be necessary to have data on the writing skills of the *same* students before they began the program. This notion could thwart attempts to assess writing skills, and in a large institutional setting, it could be so daunting that it could short-circuit any attempt to demonstrate that writing skills have improved.

Two conceptual alternatives to a pretest-posttest are discussed briefly below. Research methods experts on most campuses could further explain these and suggest additional alternatives.

The first option would be to identify which general education experiences were designed specifically to enhance writing skill. Perhaps these experiences include courses in introductory composition, rhetoric, and an initial writing-intensive course in the major. One then could compare two populations of first-year students or two populations of sophomores—those who had completed these courses with those who had not. The group that has not completed the courses can serve as the comparison or "control" against the group that competed the courses.

A second option is to compare students against a national norm on a standardized test or against a targeted "benchmark" population. Suppose the learning goal in question is that students have gained a certain level of mathematical proficiency as a consequence of taking a two-course mathematics sequence in a general education program. One can administer a standardized test of college-level mathematics after students have completed the sequence and compare students' scores to national norms. In this case, no pretest was necessary; the national norm serves as the comparison or "control" group. This method is problematic if students at the institution are not drawn from an average population, as would be the case if the institution is highly selective or open-access. However, it does produce meaningful comparisons if an institution's student population roughly approximates an average population. Scholastic Achievement Test scores, for instance, might be used as a measure of the level of selectiveness used in admitting students.

If the institution's population is not average, a benchmarking strategy would be a more appropriate alternative. Students' scores on a test of college mathematics could be compared to the scores of students at institutions with comparable populations. Scores higher than those of the benchmarked school would be convincing evidence that the math curriculum of the target institution is successful.

A common assertion is that longitudinal research designs (those that follow the same individuals over time) are necessary to draw meaningful conclusions about what students have learned. Sometimes a longitudinal perspective is warranted because other approaches are less valid. For example, if an institution is interested in demonstrating that its graduates are successful in their careers, a longitudinal survey administered to graduates repeatedly over several years would be appropriate for several reasons. Demographic data tell us, for instance, that people change careers multiple times during their lives, so examination of a single "window" of time may not be an accurate assessment. In addition, the population of graduates offers the benefit that its members will be around long enough to be surveyed repeatedly over time.

Most importantly, however, a longitudinal design guards against "cohort effects" that could intrude if graduates from one generation were compared with graduates from another generation. Career trajectories may change historically, and the character of the institution may have been markedly different in the past. Thus, 1950s female graduates may hold a lower percentage of professional degrees than 1980s female graduates. This finding tells us more about historical context than institutional outcomes. However, the same question, asked of the same individuals at several different points in time yields meaningful information. A finding that female students from the same cohort had a greater percentage of graduate degrees 20 years after college than they did 10 years after college could be used (in conjunction with other outcomes data) to demonstrate that the institution produces lifelong learners.

In most cases, when student outcomes during or at the end of a higher education experience are being evaluated, longitudinal data are not necessary and may not yield meaningful information. Pre-test and post-test assessments, as well as alternatives which are discussed above, are more practical alternatives and provide answers to the same general question: "Has meaningful learning occurred as a result of an educational experience?"

Where Do Grades Fit into the Picture?

Faculty members and others often ask whether grades are appropriate and sufficient for assessment of student learning after the learning goals are defined. The answer is that grades have been, and will continue to be, an excellent indicator of student learning *if they are appropriately linked to learning goals*. The Commission recognizes that grades are

an effective measure of student achievement if there is a demonstrable relationship between the goals and objectives for student learning and the particular bases (such as assignments and examinations) upon which student achievement is evaluated (Standard 14).

In and of themselves, however, grades are not direct evidence of student learning. That is, a numeric or a letter grade alone does not express the *content* of what students have learned; it reflects only the degree to which the student is perceived to have learned in a specific context.

One reason "grade inflation" is seen as a problem is that grades alone cannot be relied on to reflect student performance accurately. One could ask: "Does one grade of 'A' equal another?" If instructors were to match grades explicitly with goals, it would become easier to combat grade inflation, because high grades must reflect high performance in specified areas.

Grades, however, can provide an excellent means for improving teaching and learning both during a course (formatively) and at the conclusion of a course (summatively). When grades serve as the final judgment of performance in a course, they provide a summative evaluation of students' performance as individuals and as a class. If the grades of individual students can be traced directly to their respective competencies in a course, the learning achievements of those students are being assessed in a meaningful fashion. If, however, examinations or homework assignments are not designed to test the skills and competencies that the course was designed to teach, then grades for that course are measuring something other than student attainment of the course goals.

Suppose, for instance, an instructor presents the content of an anatomy and physiology course that focuses on identifying and labeling anatomical structures and physiological processes.
An appropriate evaluation of student mastery of the course content would be an objective final exam requiring students to label diagrams, answer multiple-choice definitional questions, and fill in the blanks. In contrast, an examination that required students to evaluate a physiology experiment on its methodological merits would not be an assessment of student learning of the course content. Some students would do well on the essay exam, but their performance probably would not be related to what they learned in the course. In this example, a bad grade could not be attributed to a student's failure to learn the material or to prepare for the examination. Thus, even the use of grades as a summative assessment warrants a careful approach.

Thoughtfully-constructed syllabi and "test blueprints," which are discussed later in this chapter, are two of several possible approaches to connecting grades directly to desired course goals.

Grades and grading practices can be a component of formative assessment as well. For example, many faculty members use drafting and revising processes to teach writing. Students mimic the "real world" by writing multiple drafts, submitting them to critiques by the professor or their peers, and revising them for resubmission. Each draft may be assigned a grade in association with critical comments. Depending on the instructor's preferred strategy, all or only some of the interim grades may be used to determine the final grade. In this case, the grade for each draft, in conjunction with critical comments, gives the student an indication of his or her performance, what might be done to improve the product, and how the quality of the product changes over time.

Grading can be formative when there are multiple means and formats for assessing student learning and when there are repeated opportunities to demonstrate improvement within the context of one class. For instance, a professor might assign two or three papers (with required drafts), two class presentations, two objective format exams, and a journal that must be reviewed repeatedly by the professor during the semester. Each of these "assessment events" could be graded, providing students with at least two distinct types of opportunity to learn more or learn better. A student can compare his or her performance on the various assessment formats, thereby learning which skills he or she has mastered and which should be improved. In addition, a grade on the first test administration or the first paper or presentation serves as feedback (a formative assessment) that provides information on how to improve. This learning experience can be applied toward adapting study skills or work habits before the next attempt.

Self-report Measures

Concerns are often expressed about the use of self-report measures for answering questions about student learning. Sometimes these concerns relate to the use of indirect methods of assessing student learning and the concerns about qualitative versus quantitative assessment discussed previously. Often, however, concerns are related most directly to the validity and reliability of self-report measures. Self-report measures can be designed to be valid and reliable and can be assessed by applying the characteristics of reliability and validity described above.

Both common sense and face validity should be used to determine the value of a specific self-report measure. For example, if the goal is to determine whether students are satisfied, it seems that a self-report measure is the only means of gathering such data. Satisfaction, by definition, is one's feeling of liking, comfort, and fulfillment resulting from a specific event or situation. Similarly, it is appropriate to gather data on affective states (emotions) and social perceptions with a self-report instrument (assuming that it meets the criteria for reliability and validity).

It is possible to collect direct evidence of student learning using self-report measures, but these must be designed carefully to elicit evidence of student learning. For example, students may be asked to reflect on the most important thing they learned in a specific course, or what else they would like to learn on the same subject. In doing so, they would reveal the actual content of what they had learned. However, self-report questions such as, "Did you learn a lot in this class?" would not elicit such information. Self-report measures are most frequently used to provide valuable indirect evidence of student learning.

Statistical versus Practical Significance

Data related to student outcomes are often described as being "statistically significant" or "not statistically significant." The concept of statistical significance relates to the probability of a given result occurring by chance. If the result is too unlikely to have occurred by chance, it is said to be statistically significant.

For example, imagine two groups of students, each of whom has completed an introductory calculus course. Assume that members of each group were randomly assigned to two different teaching formats—one problem-based and the other traditional lecture—and that the same professor taught each course. At the completion of the course, both students are given the same standardized calculus examination. The average grade for the students in the problem-based course was 10 points higher than the average grade for the students in the traditional lecture course. Is a 10-point difference enough to make the claim that the problem-based course is a better form of teaching? Would a 2-point difference have been enough? Would 20 points be enough to make the claim? A test of statistical significance would reveal whether the 10-point difference could have happened by accident in a normally distributed population of students (i.e., the difference could have been caused by other factors, unrelated to the course, of which we are unaware), or whether the 10-point difference was large enough that in all likelihood it was caused by differences in the courses.

Judgments of statistical significance only become reliable when there are sufficient numbers of student test or survey results from which to draw inferences about a population of students. In many cases, faculty members will be studying outcomes data from small groups of students or engaging in formative assessment for which ongoing improvement in a class or a program is the goal. In these situations, faculty and staff members should make judgments and introduce changes based on "practical significance." Do the students' scores, or their change in scores from one time to another, reveal a pattern or appear to be meaningful or informative enough to support changes in a course or program?

In general, when large-scale assessments are being used, or when standardized tests are administered program-wide or institution-wide, statistical tests should be used to analyze the data. Guidance may be found in social science, education, mathematics and statistics, and other departments on campus that use empirical methods.

Judgments of student outcomes based on practical significance are equally valid when the number of students being evaluated is small, when data are qualitative rather than quantitative, and when the purpose is to engage in formative assessment.

Key Questions When Choosing and Implementing Evaluation Instruments

One should ask several questions when choosing assessment instruments:

❑ Is the evidence provided by the evaluation method linked to important learning outcomes?

This is perhaps the single most important way to determine the quality of most evaluation tools and methods. Regardless of whether an evaluation instrument is standardized (previously published and tested for validity and reliability) or "home grown" (created locally for a specific purpose), it is important to ensure that the instrument is designed to provide evidence of the desired learning outcomes. In research design terms, this involves determining whether the operational definition (the aggregate instrument or items on the instrument) actually assesses the construct (the learning goal) that it is intended to assess (construct validity). For many standardized instruments, the intended purpose will be apparent immediately.
A disciplinary test, for example, such as the American Chemical Society (ACS) test, evaluates students' knowledge of facts, skills, and procedures that should have been acquired as a function of the undergraduate curriculum in an ACS-accredited program. Subject-area Graduate Record Examinations (e.g., the psychology GRE) evaluate content knowledge in the respective disciplines they represent. Publishers of other standardized tests with other less readily obvious content will explain, in the test information materials, what the test is designed to assess.

It is important, however, not to assume that the linkage between every item on a standardized assessment instrument and the construct it is designed to assess will be readily apparent. Many standardized instruments have built-in reliability checks and so-called "lie-scales." Measures that are designed to evaluate affective and social development are especially likely to incorporate a series of questions that seem irrelevant, but that actually enhance the instrument's validity.

❑ Is a standardized instrument appropriate for the learning goals of the institution?

It certainly is not necessary to use standardized assessment instruments. In fact, for most learning goals, none will be available. Although a test created locally may not have the same statistical validity and reliability as a standardized instrument, its relevance to the specific learning goals in question may make it a more appropriate and effective instrument. A "test blueprint" (an outline that matches test items with the learning outcomes they are intended to assess) can be used to construct a test or instrument or to evaluate how well an existing "home-grown" instrument assesses key learning outcomes.

❑ Is the evaluation method appropriately comprehensive?

No assessment tool or method can assess *every* important learning outcome, but the best ones assess a comprehensive and/or representative sample of key learning outcomes. It is not financially feasible to use several published instruments to assess multiple outcomes, nor is it feasible to subject students to multiple tests or surveys. (The latter has its own measurement problems.) Regardless of whether an assessment instrument is standardized or specially created, it should be as comprehensive as possible.

❑ Are important learning outcomes evaluated by multiple means?

Few evaluation methods are perfect. It is important to triangulate around important learning goals, assessing them through various means, and through tests of various formats. For instance, a standardized test of disciplinary knowledge may be an adequate form of assessment of students' content knowledge of a discipline, but it may provide no indication of his or her preparedness to be a good practitioner in that discipline.

❑ Are the questions clear and interpreted consistently?

In addition to examining the correspondence between learning goals and the assessment measures being used, it is important to assess whether its "non-content" properties are adequate. For example, a test should not be culture-specific, its vocabulary and sentence structure should be at an appropriate level, and it should not contain ambiguous, unclear, or double-barreled questions (i.e., questions that actually contain two questions).

Questions should be phrased carefully to ensure meaningful responses. For instance, imagine that a targeted learning goal is that students' desire to engage in community service increases after exposure to a service-learning program. Imagine also the following two questions asked of a graduate:

- ❑ "On a scale of 1 to 7, how likely are you to participate in community service activity?"
- ❑ "On a scale of 1 to 7, how much influence did your community service during college have on your desire to participate in community service in the future?"

Both of these questions are indirect measures of learning or development, but only the second provides information that would help the institution to improve the service-learning program.

A specially created instrument should be reviewed by several colleagues and students to ensure clarity, and it then should be pre-tested by some students who have diverse backgrounds and characteristics in order to clarify ambiguous items.

❑ Do questions elicit information that will be useful for making improvements?

Questions should be designed so that, when possible, they yield responses that both evaluate an aspect of the educational experience and suggest options for improvement. For instance, a survey designed to evaluate student experiences with the Career Services Office should ask about perceptions of its efficacy:

- ❑ "On a scale of 1 to 7, how important was the Career Services Office in helping you find employment upon graduation?"

The instrument also should ask how the office might be improved. For example, the respondent might be asked to name the three most useful activities of the Career Services Office for helping students find jobs and to name three ways in which the functions of that office could be improved.

The concept of creating questions that are useful for making improvements can be applied to direct assessments of student learning as well. For instance, a complicated problem in a physics class can be divided into subsections to help the professor determine which content or skill areas need additional reinforcement.

❑ Does everyone interpret the responses the same way?

When assessments of student outcomes are subjective—that is, if they do not require discrete or quantifiable or unambiguous answers—it is important to develop a rubric (criteria used to score or rate responses) to ensure comparability of review. There should be collegial agreement on what constitutes acceptable, inadequate, and exemplary responses or performance for each assessment instrument to be used, whether it is a paper, a project, a presentation, or an artistic offering. A rubric created to reflect the agreement should be pre-tested by having colleagues independently score the same work samples to see if their scores are consistent. The strategy of inter-rater reliability can be used as well, by enlisting two or more colleagues to "grade" each student's work or performance.

❑ Do the results make sense?

It is important to use common sense when developing assessment instruments, designing a scoring system or rubric, or interpreting data resulting from assessment instruments. One would expect honors students to outperform other students on their senior theses presentations. One also might expect those same students to fare better in applying to graduate school, but not necessarily in being hired to entry-level positions in corporations. Students who have completed a general education sequence should score better on tests of general knowledge and skills related to specified general education outcomes than students who have not

completed the sequence. Unexpected results should trigger further inquiry.

❑ Are the results corroborated by other evidence?

It is always important to use multiple means of assessment to determine if a particular learning goal has been met. It also is necessary to compare assessment results for related goals for student learning and even for goals that would be expected to be mutually exclusive. For instance, rubric scores for the writing quality of senior theses should be corroborated by students' grades in composition classes. Faculty ratings and students' self-ratings of performance should correlate with each other. Focus group results should support survey results on the same topic. Conversely, students who demonstrate an increased personal emphasis on wellness by their attendance at the gym and by participation in athletics should not be engaging in increased alcohol and drug consumption. The latter finding would warrant re-evaluation of the campus wellness program.

❑ Are efforts to use "perfect" research tools balanced with timeliness and practicality?

Although institutions will do their best to ensure that the research designs they use yield meaningful results, they should remember that assessment cannot wait for the perfect research strategy. Indeed, there probably is no perfect strategy. For the purpose of managing the quality and change of an academic curriculum, assessment is a form of systematic inquiry—i.e., "action research" or "applied research," based on the collection and analysis of data about student learning that is undertaken with the best knowledge and resources permissible and within the time available. The resulting information guides decision makers in choices related to the curriculum, faculty, the use of physical space, and other areas that may have an effect on learning.

❑ Is evidence gathered over time and across situations?

Assessment is not a once-and-done process. As students, faculty members, curricula, and teaching methods evolve over the years, even institutions with very positive assessment results should undertake repeated assessments to ensure that students are learning as effectively today as they were a few years ago. Because each evaluation technique has relative strengths and weaknesses, there is no single perfect assessment that will yield absolutely accurate information and that is relevant to every situation. In order to have support the findings that each evaluation yields, more than one assessment strategy should be used to corroborate findings.

❑ How much should be assessed?

Plunging immediately into assessing a large number of students on a full range of learning outcomes will overwhelm faculty members and institutional resources. It will produce an overwhelming amount of information that may be impossible to interpret or to use in enhancing a program. It makes more sense to begin with a more limited approach. For example, faculty members assessing student writing skills might gain more from a thorough analysis of a sample of 30 papers than from a more perfunctory review of 300, as well as by assessing only a few key goals.

Just as every possible outcome need not be measured, it is not necessary to collect data about each student's performance. The Commission is interested in the institution's ability to graduate students with appropriate knowledge, skills, and behavior, not in a demonstration that every student was tested. Meaningful and representative sub-populations (randomly chosen when appropriate) can provide the basis for demonstrating that students across the institution are achieving learning goals.

Evaluating Student Learning

- **Use indirect measures to explain or support findings from direct measures.**
- **Choose the most relevant level for evaluation of the learning goals: institution, program, or course.**
- **Select quantitative or qualitative measures based on type of student learning goals.**
- **Ensure that grades are related directly to goals.**
- **Choose appropriate research design.**
- **Use formative assessment "mid-course" to improve teaching and learning.**
- **Use common sense: Is the result logical?**

❑ Are faculty and staff members who are knowledgeable about measurement serving as resources for developing assessment instruments?

The work of assessing student learning is essentially systematic inquiry in the tradition of social science or evaluation research, with its attendant need for validity, reliability, control, analysis, and interpretation, to the extent that these are possible. Although everyone involved in the enterprise is an expert in the content base of what is being researched (i.e., teaching and interacting with students in a higher education setting), few are expected to be experts in conducting research. While much of the work of assessing student learning has a common-sense base, it is also true that meaningful analysis of student learning, especially beyond the course level, requires expertise. There are usually faculty members on campus who are trained as social science, education, or other researchers. They can conduct careful, meaningful research and can construct measures. These faculty members, who can be found in psychology, sociology, education, business, and other departments, may be enlisted to serve as internal consultants, reviewers, statisticians, and mentors in the assessment process.

Easy-to-Implement Tools and Techniques

The assessment tools and techniques presented below yield useful information and are relatively easy to implement. They are not meant to be an exhaustive selection of tools but, rather, an overview of available options.

Rubrics or Rating Scales

A rubric is an instrument based on a set of criteria for evaluating student work. Rubrics help a professor or other evaluator to make explicit, objective, and consistent the criteria for performance that otherwise would be implicit, subjective, and inconsistent if a single letter grade were used as an indicator of performance. Rubrics delineate what knowledge, content, skills, and behaviors are indicative of various levels of learning or mastery. Ideally, "grading" rubrics are shared with students before an exam, presentation, writing project, or other assessment activity. Conscious awareness of what he or she is expected to learn helps the student organize his or her work, encourages self-reflection about what is being learned and how it is being learned, and allows opportunities for self-assessment during the learning process. Huba and Freed (2000) suggest that instructors consider involving students in the development of rubrics as a class progresses as a way of helping students to develop their own conceptions of what constitutes good and poor work. Both Huba and Freed (2000) and Walvord and Anderson (1998) offer extensive information on developing rubrics.

Figure 6 includes a description of the characteristics and components of rubrics. Huba and Freed (2000) present a thorough description of the uses and purposes for rubrics, along with a comprehensive primer on how to construct them.

There are four basic types of rubrics: simple checklists, simple rating scales, detailed rating scales, and holistic rating scales.

Figure 6

Criterion-based Rating Scales (Rubrics)

What is a rubric? A rubric is a criterion-based rating scale that can be used to evaluate student performance in almost any area. A rubric establishes the "rules" for the assignment (Huba and Freed, 2000). It contains *a priori* criteria for various levels of mastery of an assignment.

How is a rubric used? The person evaluating student performance uses a rubric as the basis for judging performance. Ideally, rubrics are available to students prior to their completion of the assignment so that they have clear expectations about the components of the evaluation and what constitutes exemplary performance.

What are some of the criteria that may be used within a rubric to evaluate student work? Criteria can include sophistication, organization, grammar and style, competence, accuracy, synthesis, analysis, and expressiveness, among others.

What are the components of a rubric? Huba and Freed (2000) describe the following elements of rubrics:

- Levels of mastery (e.g., unacceptable through exemplary)
- Dimensions of quality (see criteria above)
- Organizational groupings (macro categories for criteria)
- Commentaries (the junctures between levels of mastery and dimensions of quality; e.g., a description of the characteristics of an exemplary organization)
- Descriptions of consequences (components of commentaries that relate to real-life settings and situations).

Where can I see examples of rubrics and learn more? Walvoord and Anderson (1998) and Huba and Freed (2000) are both excellent sources of information about the characteristics of rubrics and how to develop them. They also provide examples of various forms of rubrics.

Simple Checklists. This form of rubric can be used to record whether the relevant or important components of an assignment are addressed in a student's work. For instance, a rubric might be used to assess whether a laboratory report contained required sections or whether a writing sample contained all of the assigned parts. A checklist of this sort is categorical, that is, it records whether or not a required aspect of an assignment is present, but it does not record quantitative information about the level of competence a student has achieved or the relative skill level he or she has demonstrated.

Simple Rating Scales. This form of rubric records the level of student work or categorizes it hierarchically. It is used, for instance, to indicate whether student work is deficient, adequate, or exemplary, or to assign a numerical "code" to indicate the quality of student work.
In most cases in which a numerical scale is used, it should contain a clear neutral midpoint (i.e., the scale should contain an odd number of rating points). However, survey designers should determine when this might not be appropriate. Occasionally, such scales are intentionally designed without a midpoint in order to force a non-neutral response.

Figure 7, an excerpt from an employee rating scale, is an example of a simple rating scale that does not provide information about the "value" of different points on the scale.

Detailed Rating Scales. Detailed rating scales describe explicitly what constitutes deficient, adequate, or exemplary performance on each criterion. Detailed rating scales are especially

Figure 7

Excerpt from a Simple Rating Scale

Employer's Final Performance Evaluation of Knowledge, Skills, and Attitudes (KSAs) of: _____

Dear Employer:

The College of Business Economics (CBE) understands the need for its graduates to be broadly trained and ready to perform immediately upon entering the job market, both as individuals and in teams. Therefore, its curriculum contains concrete, measurable, and attainable objectives throughout. As a result, each CBE graduate is expected to perform successfully in the following areas of Knowledge, Skills, and Attitudes.

Please rate your intern or employee's performance only on the areas that apply to his/her job.
The rating scale is: 5=Excellent; 4=Good; 3=Satisfactory; 2=Fair; 1=Poor; N/A=Not Applicable.

Excerpt:

COMMUNICATION: WRITTEN, SPOKEN, GRAPHIC, AND ELECTRONIC	5	4	3	2	1	n/a
1. Write articulate, persuasive, and influential business reports, proposals, and letters						
2. Make articulate, persuasive, and influential individual and team presentations						
3. Develop graphic, spreadsheet, and financial analysis support for position taken						
4. Display presentation skills						
5. Generate appropriate visual aids						
6. Use correct written structure, spelling, grammar, and organization						
7. Articulate another's viewpoint through verbal and non-verbal cue interpretation						
8. Resolve interpersonal and team conflicts						
9. Negotiate effectively						
THINKING: CRITICAL, CREATIVE, AND INTEGRATED	**5**	**4**	**3**	**2**	**1**	**n/a**
10. Use problem-solving techniques						
11. Use adaptable, flexible thinking						
12. Use critical thinking to produce comprehensive, supported, integrated conclusions						
13. Use creative thinking methods to produce ideas						
14. Distinguish fact from opinion, and critical from non-critical information						
15. Develop several workable solutions to a problem						
16. Show common sense						
17. Demonstrate continuous learning (learning to learn)						

Source: College of Business and Economics, Towson University, November 2001. Adapted with permission.

Some of the other characteristics that could be evaluated in the manner shown in Figure 7 include:

- ✶Technology
- ✶Ethics and Values
- ✶Business Disciplinary Content
- ✶Leadership, Entrepreneurship,
- ✶ Diversity - International and Demographic
- ✶ Practical Excellence
- ✶ Job Experience and Career Development

useful when several faculty members are scoring student work, because they communicate common performance standards and therefore make the scores more consistent. Detailed rating scales are useful to present to students when an assignment is given or at the beginning of a semester or even a program. They provide students with a clear description of what they are expected to learn and the criteria upon which their learning will be judged.

Figure 8 is an example of a rubric designed as a detailed rating scale.

Holistic Rating Scales. Holistic rating scales define deficient, adequate, or exemplary student work as an aggregate, by assigning a single score to a constellation of characteristics that have been fulfilled to a substantial degree, rather than rating each criterion separately. Holistic rating scales often are used when evaluating student work that may vary so widely in form and content that the same criteria may not apply to all. Capstone projects in an art program, for example, might vary so that they cannot all be judged using the same specific criteria. However, a faculty member could create a generic description of what constitutes exemplary work, adequate work, and so on, regardless of the medium or focus of the work.

Figure 9 is an example of a holistic rating scale.

Self-reflection

Asking students to reflect on what and how they have learned—in other words, to engage in metacognition—has several benefits. Student self-assessments give faculty members useful insights into the learning process, help students integrate what they have learned, and provide students with an understanding of the skills and strategies they need to learn most effectively. Classroom assessment techniques suggested by Angelo and Cross (1993) and other similar self-reflection strategies have the added advantage of taking very little faculty or student time. The student often is asked to write simply a phrase or sentence. Examples of self-reflection questions that might be a useful part of an assessment program are provided in Figure 10.

Ratings/Comments from Internship or Research Supervisors

Programs that place students in practica, such as internships, cooperative education, and student teaching assignments, usually require that the on-site supervisor rate the student on essential knowledge, skills, and attitudes. Such scales are relatively simple to construct (see Figure 7.) Because these experiences require students to integrate and use much of what they have learned in a program, these rating scales are evidence of what students have learned during the program. Brief comments from supervisors also provide valuable insights into the overall strengths and weaknesses of a program.

Placement Rates

For professional programs whose goals include preparing students for a particular career, the proportion of graduates who find positions in that career is important indirect evidence of whether students are learning essential knowledge and skills. If a large proportion of graduates from a teacher education program is successful in finding teaching positions, for example, it is likely that those graduates have the knowledge and skills that school administrators consider important for successful teachers. Similarly, if a program aims to prepare students for graduate study or professional programs—pre-medicine and pre-law programs are examples—the proportion of graduates who are admitted into graduate or professional programs is important evidence that students have learned what graduate programs consider important for success in their programs. Note, however, that placement rates alone do not provide insights into exactly *what* students are learning. Therefore, they are usually insufficient evidence of student learning if used alone.

Figure 8

Example of a Detailed Rating Scale

This scale is adapted from one used to evaluate a "book journal and review" for a cognitive psychology class. For the assignment, students were expected to read one full-length book, chosen from a list provided by the instructor and related to the content of the courrse but not included on the required course reading list. The purpose of the assignment was to provide a basis for making connections between the course content, other professional or popular work in the field, and students' daily exposure to topics or situations related to cognitive psychology in their personal lives and in their other courses. A further purpose of the assignment was to enable students to develop skills in describing research in cognitive psychology to the lay public. The assignment involved reading the chosen book during the course of the semester and keeping a journal of reflections related to the purpose of the assignment. Students also were expected to write a professional style book review (of the type that might appear in the *New York Times* Review of Books). The rubric is abbreviated for inclusion here.

	Unacceptable	Fair	Proficient	Exemplary
Book Journal				
Use of grammar and style to communicate ideas effectively	Grammar and style that interfere with a reader's ability to understand the ideas presented	Grammar and style adequate for the reader to grasp the main concepts presented	Grammar and style allow the reader to understand easily the concepts presented	Grammar and style enhance the reader's ability to understand the concepts presented, including nuances of thought; May provide a pleasurable reading experience
Engagement with the author's ideas	Author's ideas are simply repeated, indicating that engagement was at or below a surface level	Occasional discussion of the author's ideas, suggesting ability to engage	Frequent discussion and analysis of the author's ideas, including expression of well-supported opinions about those ideas, suggesting almost constant engagement	Rich, mature grasp of the author's ideas, coupled with analysis and synthesis with own ideas and ideas of other writers and scholars, suggesting constant and sophisticated engagement
Connections between the course and the book	Very few connections with course material	Sporadic but meaningful connections with course material	Regular and meaningful connections to course material	Continual connections to course material and sophisticated discussion of those connections
Connections between other experiences and the book	Very few connections with other experiences	Sporadic but meaningful connections with other experiences	Regular and meaningful connections with other experiences	Continual connections to other experiences and sophisticated discussion of those connections

Book Review				
Grammar and form	Grammar and style impede understanding of the "plot" or thesis of the book; not consistent with the form of a professional book review	Grammar and style are adequate for the reader to grasp the "plot" or thesis of the book; the form is consistent with that of a professional book review	Grammar and style allow the reader to understand easily the "plot" or thesis of the book; closely adheres to the style and form of a professional book review	Grammar and style enhance the reader's ability to understand the "plot" and thesis of the book; indistinguishable from a professional book review
Communication of cognitive psychology concepts to the reader	Ignores the reader's perspective of the reader and/or communicates cognitive psychology concepts inaccurately or without scientific analysis	Sometimes considers the perspective of the reader and occasionally communicates cognitive psychology concepts well	Consistently addresses the perspective of the reader and commuinicates cognitive psychology concepts accurately and usefully	Engages the reader and "forces" the reader to be interested in the topic of the book; describes cognitive psychology concepts accurately and usefully

Test Blueprints

The creation of local examinations—"traditional" examinations at the course level, or comprehensive examinations at the program level—ideally begins by writing a test blueprint before developing the actual test questions. Often called a table of specifications, a test blueprint is a list of the key learning outcomes to be assessed on the test, with the number of points or test questions to be devoted to each goal.

An example of a test blueprint appears in Figure 11. Note that in a test blueprint, an essential learning outcome might be represented by questions worth a total of 20 points, while a lesser learning outcome might be represented by only 5 points.

The test blueprint itself is important evidence of the test's validity. When matched with test scores, it offers clear evidence of what students have learned because it covers all learning goals. One could say with confidence, for instance, that a student earning an "A" on the test has mastered all or most of the important learning outcomes for a course or a program.

Other Assessment Tools

Some other assessment tools may be valuable components of many successful assessment programs, but they are more difficult or time-consuming to implement than the tools suggested above, and they also may require significant financial resources to purchase or administer. Careful consideration is warranted to determine whether information yielded from these strategies justifies the time and effort they require.

Multidimensional or Comprehensive Tests

As most faculty members are already aware, valid and reliable tests can be very difficult to design, especially those meant to assess higher-order thinking skills, attributes, or values. Tests of this type should be administered, analyzed, and revised over several semesters to eliminate poorly written items and to ensure optimal quality. It is best to seek the advice of a colleague who is an expert in "tests and measurements" before embarking on the construction of a comprehensive test of multiple student learning goals. Several books are primers on test construction. At the very least, they will provide the reader with an overview of the best

Figure 9

Example of a Holistic Scoring Guide (For Critical Thinking)

by Facione and Facione

[Ed. Note: The criteria below are shown from the highest score to the lowest.]

4 Consistently does all or almost all of the following:

- ❑ Accurately interprets evidence, statements, graphics, questions, etc.
- ❑ Identifies the salient arguments (reasons and claims) pro and con
- ❑ Thoughtfully analyzes and evaluates major alternative points of view
- ❑ Draws warranted, judicious, non-fallacious conclusions
- ❑ Justifies key results and procedures, explains assumptions
- ❑ Fair-mindedly follows where evidence and reasons lead

3 Does most or many of the following:

- ❑ Accurately interprets evidence, statements, graphics, questions, etc.
- ❑ Identifies relevant arguments (reasons and claims) pro and con
- ❑ Offers analyses and evaluations of obvious alternative points of view
- ❑ Draws warranted, non-fallacious conclusions
- ❑ Justifies some results or procedures, explains reasons
- ❑ Fair-mindedly follows where evidence and reasons lead

2 Does most or many of the following:

- ❑ Misinterprets evidence, statements, graphics, questions, etc.
- ❑ Fails to identify strong, relevant counter-arguments
- ❑ Ignores or superficially evaluates obvious alternative points of view
- ❑ Draws unwarranted or fallacious conclusions
- ❑ Justifies few results or procedures, seldom explains reasons
- ❑ Regardless of the evidence or reasons, maintains or defends views based on self-interest or preconceptions

1 Consistently does all or almost all of the following:

- ❑ Offers biased interpretations of evidence, statements, graphics, questions, information, or the points of view of others
- ❑ Fails to identify or hastily dismisses strong, relevant counter-arguments
- ❑ Ignores or superficially evaluates obvious alternative points of view
- ❑ Argues using fallacious or irrelevant reasons, and unwarranted claims
- ❑ Does not justify results or procedures, nor explain reasons
- ❑ Regardless of the evidence or reasons, maintains or defends views based on self-interest or preconceptions
- ❑ Exhibits close-mindedness or hostility to reason

For further information, contact the authors at Insight Assessment (info@insightassessment.com; Phone: 650-692-5628) or visit the website at http://calpress.com/rubric.html for a reproducible version and instructions.

Figure 10

Student Self-reflection Questions for a Course or Program

1. How do you feel about writing/teaching/biology/sociology/etc.?
2. What will you say to your friends about this course/program?
3. What suggestions would you give other students on ways to get the most out this course/program?
4. How do you feel about yourself as a writer/teacher/biologist/sociologist/etc.?
5. What are your strengths as a writer/teacher/biologist/sociologist/etc.?
6. What makes a person a good writer/teacher/biologist/sociologist/etc.?
7. What was the one most useful or meaningful thing you learned in this course/program?
8. What was your biggest achievement in this course/program?
9. In what area did you improve the most? What improvement(s) did you make?
10. What one assignment for this course/program was your best work? What makes it your best work? What did you learn by creating it? What does it say about you as a writer/teacher/biologist/sociologist/etc.?
11. Describe something major that you have learned about yourself in this course/program.
12. List three ways you think you have grown or developed as a result of this course/program.
13. In what ways have you improved as a writer/teacher/biologist/sociologist/etc.?
14. What have you learned in this course/program that will help you continue to grow as a writer/teacher/biologist/sociologist/etc.?
15. What was your favorite aspect of this course/program? Why?
16. What goals did you set for yourself in this course/program? How well did you accomplish them?
17. If you were to start this course/program over, what would you do differently next time?
18. What strategies did you use to learn the material in this course/program? Which were most effective? Why?
19. What risks did you take in this course/program?
20. If you could change any one of the assignments you did for this course/program, which one would it be? What would you change about it?
21. What problems did you encounter in this course/program? How did you solve them?
22. What one question about this course/program is uppermost on your mind?
23. What would you like to learn further about this subject/discipline?
24. In what area would you like to continue to strengthen your knowledge or skills?
25. Write one goal for next semester/year and describe how you plan to reach it.

Figure 11

Example of a Test Blueprint

Educational Research Methods: Final Exam Outline

The final exam will consist of 25 multiple-choice items, each worth 2 to 4 points, and five short essay questions, each worth 3 to 5 points. The items will cover most of the concepts listed below.

Validity and Reliability (Up to 16 points)

- ❑ Demonstrate an understanding of reliability and validity.
- ❑ Correctly identify the type of reliability and validity evidence being provided by given information on an instrument.
- ❑ Recognize examples of measurement error in a given situation.
- ❑ Assess the meaning and implications of measurement error.
- ❑ Apply general principles for ensuring validity.

Inferential Statistics (Up to 16 points)

- ❑ Demonstrate an understanding of the concept of a null hypothesis.
- ❑ Select the most appropriate inferential statistics (t, F, or ^{2}) for a given research situation
- ❑ Know the most common "cut-off" point that statisticians use in deciding whether two means differ statistically significantly from one another.
- ❑ Correctly interpret the results of t, F, and ^{2} tests as presented in research articles.
- ❑ Interpret the effect of standard deviation and sample size on the results of a statistical test.

Experimental Research (Up to 12 points)

- ❑ Interpret correctly the symbolic representations of experimental designs.
- ❑ Describe the benefits and limitations of each experimental and quasi-experimental design covered in class.
- ❑ Identify the appropriate research design for a given research situation.

Correlational Research (Up to 12 points)

- ❑ Demonstrate an understanding of regression and the use of regression equations.
- ❑ Understand what r, R^2, and partial correlations are and what they tell us.
- ❑ Understand what multiple regression analysis is used for and what it tells us.

Qualitative Research: Observation, Interviews, and Ethnographic Research (Up to 16 points)

- ❑ Describe and discuss qualitative research and its key characteristics.
- ❑ Identify the pros and cons of qualitative research.
- ❑ Describe the concept of a focus groups.
- ❑ Identify the pros and cons of focus group research.
- ❑ Understand the key principles in conducting focus groups.
- ❑ Define ethnographic research is and identify or describe examples of it.

Historical Research (Up to 10 points)

- ❑ Describe the need for historical research.
- ❑ Identify kinds of historical research sources.
- ❑ Recognize examples of primary and secondary resources.
- ❑ Understand how to evaluate historical research.

Content Analysis (12 points)

- ❑ Demonstrate an understanding of content analysis.
- ❑ Understand the pros and cons of content analysis.
- ❑ Recognize examples of different kinds of content analysis.
- ❑ Explain how to analyze content analysis data.

Multiple Units (Up to 6 points)

- ❑ Identify the most appropriate research method for a given situation.

questions to ask when seeking expert advice (Anastasi and Urbina, 1996; Haladyna, 1999).

Adding a published test to an assessment program will require time to identify, evaluate, and experiment with potential tests. Unfortunately, many published tests aimed at the higher education market offer limited evidence of quality (i.e., validity and reliability) and have been normed with relatively small groups of students.

It is most important to compare the test blueprint against the key learning outcomes of the course or program in question to see how well they match. A biology test that focuses on ecological concepts, for example, probably would not be appropriate as a key assessment instrument for a biology program that aims to prepare students for careers in health professions.

Figure 12 contains a list of published tests designed to test critical thinking and general education goals. It is presented here as an example of the various test characteristics that should be considered when choosing an appropriate published assessment instrument.

Ad Hoc Surveys and Pre-graduation Surveys

Many people view surveys as a quick way to collect assessment information. Unfortunately, surveys that are designed and administered quickly often have low response rates and poorly-phrased questions that yield information of questionable value.

Indirect assessments of student perceptions and satisfaction that are administered at the institutional level and are not embedded in course and program requirements—such as voluntary graduating senior surveys—take extra time and effort for both students and faculty members, and they present sampling problems. It also can be difficult to motivate students to participate in such extraneous assessment efforts, or to give their best possible effort and thought, thus reducing the validity of the assessment itself. It is often simpler, more efficient, and more effective to use assessment strategies that are intrinsic parts of course and program requirements. Graduating senior surveys, for instance, could be administered as part of a capstone course offered in every major.

If an institution determines that a survey is a key element of an assessment strategy, it should help to ensure useful results by conducting a pilot test of the survey. A draft should be administered to a small group, the responses analyzed, and unclear questions identified. Strategies to maximize the response rate should be included in the plans to administer the actual survey.[5]

Focus Groups

A focus group interview often is viewed as another quick way to collect assessment information, but the relatively small number of participants and the free-flowing format can reduce the credibility and value of the results. Focus groups are usually most appropriate as tools to help illuminate other assessment results, rather than as stand-alone assessment strategies.

Successful focus groups require time for planning, testing, and analysis to ensure a balanced discussion among a sufficient number of participants and to assure that the results have credibility and value. One should learn how to plan and conduct focus groups, hire a consultant, or enlist the aid of an on-campus expert before using focus groups as an assessment strategy.

Several sources introduce the science of conducting focus groups and their use as a source of information. For example, see Morgan (1997); and Krueger and Casey (2000).

Portfolios

Portfolios are structured, focused, and purposeful collections of student work. They are increasingly popular assessment strategies, because they provide an exceptionally comprehensive, holistic picture of student learning.

Figure 13 offers some questions that may help in a decision on whether or not to use portfolios. If the decision is made to use portfolios, it is best to

5 For a discussion of effective survey use, see Suskie (1996).

Figure 12

Commonly-administered Measures of Critical Thinking

Measure	Critical Thinking Definition	Subscales	Design	Appropriate Participants
Watson-Glaser Critical Thinking Appraisal	Comprises attitudes, knowledge, skills	Inference, Recognition of assumptions, Deduction, Interpretation, Evaluation of arguments	Parallel forms A & B; 80 multiple-choice items, based on readings; 40 mins. to complete	9th grade and higher
California Critical Thinking Skills Test	Purposeful, self-regulatory judgment	Analysis, Evaluation, Inference, Inductive, Deductive	Parallel forms A and B; 34 items; 40 mins.to complete	College age
California Critical Thinking Dispositions Inventory	Attitudinal inclination to apply critical thinking skills	Truth seeking, Open mindedness, Analyticity, Systematicity, Critical thinking self-confidence, Inquisitiveness, Cognitive maturity	Likert-type scale; 75 items; Response ranges from Agree to Strongly Disagree; 40 mins.to complete	College age
Ennis-Weir Critical Thinking Essay Test	Reasonably deciding what to do or what to believe	Getting the point; Seeing reasons and assumptions; Stating one's point; Offering good reasons; Seeing other possibilities; Equivocation; Irrelevance; Circularity; Reversal of conditional relationships; Straw person fallacy; Overgeneralizations; Excessive skepticism; Credibility; Using emotive language to persuade	Essay format; Responses written to questions about scenarios; 40 minutes to complete	Grade 7 to College
Cornell Critical Thinking Test	Reasonably deciding what to do or believe	Level X: Induction; Deduction. Credibility, Assumptions, Value judgment; Meaning Level Z: All level X subscaled plus semantics, prediction, definition	Level X: 71 multiple-choice items based on scenarios; Level Z: 52 multiple-choice items based on scenarios; 50 mins. to complete	Level X: 4th grade-college sophomore Level Z: gifted high school and college-aged adults

References: Adams, M., Whitlow, J., Stover, L., and Johnson, K. (1996). Critical thinking as an educational outcome: An evaluation of current tools of measurement. *Nurse Educator, 21 (3),* 23-31.

Facione, N.C. (1997). *Critical thinking assessment in nursing education programs: An aggregate data analysis*. Millbrae: The California Academic Press.

Ennis, R. H. & Millman, J. (1985). *Cornell critical thinking test, Level 2*. Pacific Grove, California: Midwest Publications.

Ennis, R. H. & Millman, J. (1985). *Cornell critical thinking test, Level X*. Pacific Grove, California: Midwest Publications.

Rane-Szostak, & D. Robertson, J. F. (1996). Issues in measuring critical thinking: Meeting the challenge. *Journal of Nursing Education, 35(1),* 5-11.

Note: This table contains information on the most commonly used measures of critical thinking only. It is not meant to be exclusive. There are many more measures available, including several domain-specific measures. Table prepared in 2001 by D. A. Redding,Ph.D., Instructional Assistant Professor, Mennonite College of Nursing, Illinois State University. Reproduced with permission..

start on a small scale. Portfolios may be especially appropriate for programs that enroll only a handful of students. Such programs would be ideal for piloting portfolio projects for later use with larger programs.

Portfolios can present significant logistical problems related to sampling, storage, development of evaluation criteria, and the allotment of sufficient faculty and staff time for review. These issues can be resolved, but the solutions may take time to identify and implement. For example, a number of institutions use electronic portfolios to solve the storage problem. Huba and Freed (2000) provide an excellent discussion of the development and assessment of portfolios.

Retention/Graduation Rates

Retention and graduation rates that do not meet the institution's goals may be signs of problems with student learning. However, they do not necessarily reveal what students actually have learned. They can be useful to the extent that they correlate with and illuminate direct learning assessments, or that they assess directly such institutional outcomes as cost effectiveness, diversity, student achievement, and other evaluations of institutional effectiveness.

Figure 13

Considerations when Deciding to Use Portfolios

1. What are the goals of the portfolio?
   - ❑ What do you want your students to learn by the act of creating a portfolio?
   - ❑ What processes or outcomes are to be evaluated by the portfolio?
2. How will students choose what to include in the portfolio?
3. How and when will work be inicluded in the portfolio?
4. How will student and faculty reflection occur in the portfolio process?
5. How will the portfolios be reviewed and evaluated? What would a successful portfolio in your program look like? What are your criteria for deciding if a portfolio is a "success"?
6. Will the portfolios be graded? If so, how?
7. How and where will portfolios be stored?
8. Will the portfolios be passed one faculty member to another? Will students retain ownership of portfolios?
9. What are the benefits of moving toward portfolio assessment? What are the areas of concern?
10. Is the collection of student work a feasible practice in your program?