1 Understanding the Varieties of Assessment

In this chapter, we examine the purposes of assessment and describe assessment language. In the course of doing so we establish an assessment vocabulary for this book, so authors and readers will speak the same language. The definitions we present have been developed from our assessment research with districts, schools, and teachers, as well as from our work with the SERVE Regional Educational Laboratory.

Another, more essential, purpose of this chapter is to begin the process of increasing your assessment repertoire by looking at the diverse functions of various assessments. The more you know about what each kind of assessment seeks to assess, the better able you will be to select the assessment that tells you what you need to know about students' levels of knowledge and ability. In turn, you will be able to create assessments that will be most relevant to your students.

One can view assessment from many different perspectives: its purpose, it methods, its processes, its objects, its data results, its measurement accuracy, its relationship to activities outside of school. Moreover, because they consider different things, these perspectives are not mutually exclusive. There is much overlapping, and any given assessment activity may fall into several of the groupings we discuss here.

Moreover, from this discussion of assessment terms one should not infer that one type of assessment is "good" while another is "bad" or that one type of assessment stimulates higher-order thinking more than another. There is a place for *all types of assessment* in the classroom; the key is to use a *variety* of assessment types to assess student learning. This chapter will give you an overview of the kinds of things you will be considering as you design or select assessments to meet particular classroom goals.

Assessment Purposes

Assessment is the act of collecting information about individuals or groups of individuals in order to better understand them. The twin purposes of assessment are to provide feedback to students and to serve as a diagnostic and monitoring tool for instruction. The definition of *classroom assessment* expands on these purposes: "Classroom assessment is an ongoing process through which teachers and students interact to promote greater learning. The assessment process involves using a range of strategies to make decisions regarding instruction and gathering information about student performance or behavior in order to diagnose students' problems, monitor their progress, or give feedback for improvement. The classroom assessment process also involves using multiple methods of obtaining student information through a variety of assessment strategies such as written tests, interviews, observations, and performance tasks" (McMunn, 2000, p. 6).

Assessment is not a thing that is done to students but a process that can lead to improved learning. In essence, assessment raises or answers the following questions:

> Did the students achieve the intended standards?
>
> If the student did not achieve the intended standards, will the feedback she received help improve the student's performance?
>
> Was the instruction effective?
>
> If the instruction was *not* effective, how can the teacher improve instruction to meet the needs of all students?

The results of the assessment are shared with both the students and the teacher. If the assessment indicates a need for improvement, students can explore new study strategies, and teachers can search out and implement new instructional techniques that target the student's strengths and weaknesses.

Many texts use the terms *assessment* and *evaluation* interchangeably. However, in our view the two terms are not synonymous. *Evaluation* is a judgment regarding the quality or worth of the assessment results. This judgment is based on multiple sources of assessment information. Envision each classroom assessment as a snapshot of what students know and are able to do. A number of these snapshots can be collected into an album and used as evidence in an evaluation. This evaluation process goes beyond just collecting information, however; evaluation is concerned with making judgments about the collection. Evaluation thus involves placing a "value" on the collection. Assume for a moment that the album contains real photographs and belongs to a professional photographer. When she applies for a job, she brings her photo album (her portfolio of her best work [assessments]) along. She has performed a personal evaluation of each snapshot in the album (judging which individual pieces to include), and has made a decision about whether or not to include each one. Now, her future

employer (she hopes) can use the multiple examples of her work presented in the album to make an informed judgment of the overall proficiency of her photography work (judging all the pieces as a whole). The photographer *assesses and evaluates* her ongoing work, and the future employer *evaluates* her worth as a photographer based upon multiple examples of good evidence. Similarly, grades given to students are also based upon the *evaluation* of *assessment* information. (That is why, as we will discuss later, a student's final grade should also be a result of quality, best-work data from the assessment process in the classroom, not the compendium of grades on everything the student has done or attempted in the classroom.)

Evaluation, then, is mostly a *summative* process whereas assessment, if done correctly, is both *formative* and *summative*. Formative assessment sets targets for students and provides feedback on progress toward those targets in ways that foster *more* progress. In the classroom, teachers use formative assessment on a daily basis and then use summative assessments as a culminating experience, which give information on students' mastery of content, knowledge, or skills. Summative assessments would be scored events that are placed in a teacher's grade book. These grades are evaluated into final grades for the end of a marking period, course of study, or mastery of standards, and are reported for student achievement.

Such summative assessments may include teacher-made tests or large-scale assessments. Unfortunately, the final evaluation, the "grade," can only be as good as the assessment information collected. If a teacher is producing or collecting poor assessment snapshots, the grade given for the full photo album will be of little use in determining what the students really know or are able to do. Nancy's story about Len, in the Introduction, illustrates this possibility.

One other purpose for assessment in addition to formative and summative processes not mentioned previously is *diagnostic* assessment. The purpose for this assessment is designed to determine student's knowledge, skills, or misconceptions prior to planning instruction. An example of this type of assessment would be when a middle school social studies teacher gives students a map and asks students to locate places, interpret the legend, and calculate distances prior to a unit on mapping. This would help a teacher know what vocabulary or skills needed to be taught.

Assessment Language

Formative, diagnostic, and *summative* are terms that relate to the overall purposes for which assessment is being carried out. However, over the years, numerous other terms, such as "traditional," "non-traditional," "alternative," "authentic," "performance," and "sound assessment," have filled books and journals. The key to using assessment well is to understand the terminology. We describe below key terms and provide examples to help you understand the importance of this language of assessment.

Selected Versus Constructed Response

Assessments may be considered from the point of view of the methods or techniques they employ. Some assessments ask students to choose a response from a given list. Both classroom and larger-scale assessments have traditionally relied heavily on this assessment type. Such *selected-response* (more traditional, or paper-and-pencil tests) assessments include the standard true-false quiz and the multiple-choice test so familiar to students. However, matching exercises also fall under this category, as do fill-in-the-blank activities when students are given a "word bank" from which to choose answers. In these assessments, students are expected to recognize that one particular choice or best answer to the question asked is sought. A selected response example is the following:

An acid*

a. Turns red litmus paper to yellow
b. Releases hydroxide ions in solution
c. Tastes sour
d. Feels slippery to touch

* The correct answer is c.

Of course, this can have limiting effects on students with creative minds, those who can think of reasons that many of the choices would work. These assessments can also be detrimental when test questions are written that may unintentionally trick students with an answer choice such as "(e) I don't know." This choice is counted as a wrong answer if it is chosen, although it might be a true answer in that the student really does not know the answer. In addition, on selected-response items, students can guess at the answers and often do well on the assessment even without a true understanding of the concepts covered. Assessments seeking selected responses have a place, especially in assessing certain types of understanding, but they should not be the only measure of student achievement of learning targets.

In contrast, assessments may also be designed so that students must create, or *construct,* a response to a question or prompt. In the past we sometimes called these constructed responses *alternative* (nontraditional) assessments because they were alternative to the more traditional, selected-response assessments just described.

Assessments requiring a constructed response include stock classroom assessments such as short-answer and essay questions, in which students are called upon to respond to a question by using their own ideas and their own words. Thus, formats for assessments include either selected or constructed responses where the "information is presented in one form, and students are asked either to construct or to select the same information in a different form" (Anderson, Krathwohl, Airasian, Cruikshank, Mayer, Pintrich, Raths, & Wittrock, 2001, p. 71). A constructed response example follows:

Differentiate between an acid and a base.

Of course many other activities require student creativity in the classroom. Also included in this category are musical recitals, theme papers, drama performances, student-made posters, art projects, and models, among many others. It should be evident, then, that using constructed response forms of assessment does not necessarily require inventing new ways of assessing students because many assessments that ask students to construct responses are already in use in classrooms around the country. We simply encourage *more* teachers to use this type of assessment *more* often. However, teachers must be careful to use and design assessments that measure targets or skills that have been made clear to students. Without clear targets, these assessments can simply become activities that go nowhere. Teachers must think about the purpose for the assessment in terms of how it will be judged and what instructional strategies will help students achieve the assessed targets.

Performance and Product Assessment Methods

As mentioned previously, constructed responses can include both performances (musical recitals and dramas, for instance) and products (essays and posters). Both of these assessment methods require that students obtain mastery of learning targets outlined in the curriculum. Products are student creations and performances that show what students can do; however, both assessment methods must align to the learning targets. These assessment methods will be explored more in Chapters Four and Five. Here we try to differentiate between literal or true assessment methods and basic activities that teachers sometimes use that may not lead to assessing student learning.

The word *performance* often elicits a vision of a musical recital, a dance, a concert, or a play. However, an understanding of the influence performance assessment has on the learning process requires a broader view of this type of assessment. Using performance assessment methods, student expectations for learning may take a variety of forms and are not limited to the arts. Making a speech, performing a laboratory experiment, demonstrating the construction of a birdhouse to specifications, or driving a car in driver's education class may all be construed as types of performance assessments.

Teachers are sometimes confused about the difference between products or performances used as assessment methods and those that are simply classroom activities. Often teachers have students engage in very enjoyable activities that are, however, not aligned with the standards for the course and therefore do little to forward the curriculum for the course. A true performance or product assessment, conversely, *demonstrates student mastery of a portion of the curriculum.* Therefore, in using a true performance or product assessment method, the targeted curriculum is linked directly to the result because the curricular standards are used to define the student expectations for learning, and the instructional strategies are selected to aid students in achieving the targets. Thus, when teachers plan lessons and units, it is important that curriculum, assessment, and instruction be considered together in order to ensure a quality learning experience for students.

Differentiation between classroom activities and these assessment methods is one of the harder concepts to convey to teachers—perhaps because teachers may enjoy doing a particular activity with their students and believe the lesson has merit simply because it is so enjoyable. For example, a chemistry teacher made peanut brittle with her students at Christmas. The scientific-sounding title of this activity was "Partial Degradation of a Six Carbon Sugar, Utilizing Protein Inclusions." Although it sounded scientific, the activity was not effective in forwarding the curriculum because very little learning about science occurred. Therefore, the construction of the peanut brittle in this activity could not be classified as a true assessment method, since it did not forward the curriculum for the course.

Authentic Assessment

Some assessments elicit demonstrations of knowledge and skills in ways that prepare students for life, not just to take a test. These assessments may resemble "real life" as closely as possible. For example, being able to subtract $1.57 from $5.00 on paper does not mean that the student could make change in the real world. Making change is authentic; subtraction on paper may not be. *Authentic* types of assessment may be perceived as realistic and relevant to the student's needs and interests if these assessments are meaningful, challenging, performance driven, and if they integrate rather than fragment knowledge for students. An authentic response example follows:

> Your mother took a TUMS™ tablet last night for acid indigestion. Why? Trace the TUMS through her system, describing the correct chemical reactions. Why did she burp?

When students participate in politically oriented debates, write for the school newspaper, conduct student government, club, or research group meetings, or perform scientific research, they are engaging in tasks that are authentic. Students appear to learn best when they have a personal reason (see relevance) for learning and when the learning environment is familiar to them. Authentic assessments provide this environment and relevance for students. For example, one way to implement such assessments is to strive to assess students as they would be assessed in the workplace or when carrying out some task that is especially meaningful to them now. Speaking (not just reading) a foreign language, developing paintings for educational offices to use, seeking out information on why cast iron frying pans are good sources of iron, and determining what brand of bubble gum has the highest percentage of sugar are all engaging assessments for students.

Exhibit 1.1 illustrates a sampling of selected-response, constructed-response, and authentic assessments in an elementary classroom.

Quality Assessment

Another term, prevalent in recent literature, is *quality assessment.* When teachers are clear in their expectations for students regarding an assessment, consider bias and purposes of the assessment, and share those expectations in advance of the assessment, they are

EXHIBIT 1.1. SAMPLE ASSESSMENTS FROM AN ELEMENTARY CLASSROOM.

Selected response: Noelle wishes to buy three apples. If each apple costs 11 cents, how much money must she spend?

a) 31 cents
b) 22 cents
c) 33 cents
d) $1.33

Constructed response: Noelle has $1.00 to spend on candy. She wants to buy a lollipop for herself and one for each of the other ten players on her softball team. Will Noelle have enough money to buy these lollipops? Explain your answer.

Authentic: (Teacher's instructions) Jesse, take a $5.00 bill from your practice (play) money to the "classroom store." Choose one of the items in the store (nothing in the store costs over $3.00) and pay the storekeeper. Noelle, as the storekeeper, you are responsible for giving Jesse his correct change.

practicing quality assessment. Quality assessment also necessitates providing good feedback to students, using assessment data to improve instruction, and using a variety of assessment methods. One key to understanding quality classroom assessment is to view assessment as an ongoing, student-participatory activity, not just as something the teacher "does" to students. Teachers must strive to give students quality work to do if they want students to do quality work for them.

Tests

Testing involves using a method or instrument to measure skills, knowledge, performance, capacities, intelligence, or aptitude of an individual or group. Tests are generally only one piece of classroom assessment information. Tests are constructed to meet a specific need or purpose, such as individual diagnosis, summative assessment of individual achievement, or school accountability for teaching a curriculum.

Standardized and High-Stakes Tests

The various tests the states administer are sometimes referred to as "standardized" tests or "high-stakes" tests. These large-scale tests are used to collect information about student learning and are administered in the same way across many classrooms so that the data can be used for making comparisons.

The U.S. Congress, Office of Technology Assessment (1992), defines a standardized test as one that uses uniform procedures for administration and scoring. Therefore, any test can be standardized if the conditions under which it is given are controlled and if identical scoring mechanisms are used for each group who takes

the test. This means that standardized tests can include multiple choice tests, oral examinations, essay writing, and performance-based assessments. However, in general use, the term *standardized test* usually refers to a multiple-choice type of exam. Standardized tests are of "high stakes" when the results are used to mandate actions that affect stakeholders in education or simply when the public perceives the tests to be of high importance. Examples of actions that may follow from high-stakes tests include evaluation and rewarding of teachers or administrators, allocation of resources to the school or school districts, school or school system accreditation, and graduation, promotion, or placement of students. For example, a competency test given to students in high school can mandate who will graduate and who will not receive diplomas. This example is high stakes because the results mandate a particular action: graduation from high school. The SAT (Scholastic Aptitude Test) does not mandate any action, but is still perceived as high stakes because of the importance the public places on this test. The public believes that SAT scores are of the highest importance in gaining admission to universities. In fact some universities may view such results as secondary in importance to factors such as high school grade point average, admissions essays, and references from teachers. However, regardless of whether the SAT is the deciding factor for admission or not, the public perceives it to be. Therefore, the SAT is classified as a high-stakes test. (In Chapter Twelve we discuss a number of important issues in standardized, high-stakes testing.)

Conventional Classroom Tests

Conventional tests are typical tests used or created by teachers. For example, teacher-created tests can be quizzes, multiple choice, true or false, and writing prompts for essays or literature readings. Many ready-made tests can also be found online or in textbook resources. Problems with alignment to instructional learning targets may arise, however, if such ready-made tests are used.

Norm-Referenced Versus Criterion-Referenced Tests

All tests, like other assessments, may be further classified into two categories: norm-referenced and criterion-referenced. *Criterion-referenced* assessment tells the teacher how well students are performing in terms of specific goals or standards. *Norm-referenced* assessment compares student performance to the performance of a normal group of students, either national or local. In order to understand and make use of the information that tests reveal about student achievement, it is essential to understand the differences between these two test types.

Norm-referenced and criterion-referenced tests basically differ in the method by which content is chosen and how a score is determined. Norm-referenced tests are used primarily to classify students. Therefore, the content of a norm-referenced exam is chosen according to how well it discriminates among student achievement levels. To this end, the test uses achievement differences between students to establish rank

ordering from high achievers to low achievers. In standardized testing, the scores of the "norm" group of students that takes such a test before it is published for general use are the standard by which subsequent test-takers are measured. Once "norm" scores for standardized tests are established, it is not unusual to continue to use these scores for seven years (Bond, 1996).

Criterion-referenced tests, as opposed to norm-referenced ones, are designed to share what and how much a student has learned. Criterion-referenced tests, then, measure how well a student did compared to some predetermined standard of performance. The content for these exams is selected based upon the extent to which such content matches the learning outcomes of the curriculum. In standardized and high-stakes testing, these tests are primarily used to measure student achievement relative to educational goals or objectives set by a school, district, or state curriculum. The test scores are used to determine how well a student is progressing through the curriculum or how well the school is teaching the curriculum (Bond, 1996).

Since norm- and criterion-referenced tests have different purposes, the scoring for these tests is also differentiated. Mehrens & Lehmann (1987, p. 15) summarize these scoring differences as follows:

> If we interpret a score of an individual by comparing that score with those of other individuals (called a norm group), this would be norm referencing. If we interpret a person's performance by comparing it with some specified behavioral domain or criterion of proficiency, this would be criterion referencing. To polarize the distinction, we could say that the focus of a normative score is on how many of Johnny's peers perform (score) less well than he does; the focus of a criterion-referenced score is on what it is that Johnny can do. . . . In norm referencing we might make a statement that "Johnny did better than 80 percent of the students in a test on addition of whole numbers." In criterion referencing we might say that "Johnny got 70 percent of the items correct on a test on addition of whole numbers." Usually we would add further "meaning" to this statement by stating whether or not we thought 70 percent was inadequate, minimally adequate, excellent, or whatever.

It is important to remember that both norm- and criterion-referenced tests can be standardized and can be high-stakes.

Aptitude Versus Achievement

Tests can be further classified by whether they measure aptitude or achievement. Again, the proper label for a test and, more important, the proper subsequent use of test scores is influenced by the test's purpose and content. For example, the purpose of an aptitude test appears to be related to the U.S. Army slogan, "Be all that you can be." How can capacity, potential, or ability be determined? An aptitude test strives to do this by measuring or predicting various kinds of behavior related to these concepts. Among standardized tests, the SAT and the ACT, for example, are used to predict a student's success in college. Intelligence tests, like the Stanford-Binet or Wechsler Intelligence Scale for Children, are further exemplars of this classification. Therefore, aptitude tests "tend to measure or predict (a) the effects of the cumulative influence

of experiences, (b) the effects of learning under relatively uncontrolled and unknown conditions, and (c) the future behavior, achievements, or performance of individuals or groups" (Payne, 1997, p. 380). Aptitude tests are primarily norm-referenced exams, as the aptitude of an individual is compared to those of a norm group.

Conversely, as Payne states, achievement tests "measure (a) the effects of special programs, (b) the effects of a relatively standardized set of experiences, (c) the effects of learning that occur under partially known and controlled conditions, and (d) what the individual student can do at a given point in time" (Payne, 1997, p. 380). Further, "aptitude measures (including readiness tests) are administered before the learning program, and achievement tests are administered after the fact" (Payne, 1997, p. 380).

One particular type of achievement test, introduced in Florida (Beard, 1986), has created a storm of controversy. This type of achievement test is the test of minimum competency. In 1976, Florida mandated by law that all high school students had to pass a minimum competency exam in order to receive a diploma. "Whether such a diploma sanction applies or not, minimum competency testing is precisely what the name implies: a program to test students in terms of, and only in terms of, whatever competencies state or local authorities have decided are the minimally acceptable result of an education" (Lazarus, 1981, p. 2). A minimum competency test, therefore, is a special subset of the achievement test classification, in that it is given after the learning experience, and measures what the student can actually do at a particular point in time. Minimum competency exams, like all achievement tests, may be either norm-referenced or criterion-referenced. Therefore, student performance on these tests may be compared to norm groups or to curriculum standards.

Relevance, Reliability, and Validity

Whether we are examining tests or other assessments, relevance, reliability, and validity are important terms in the assessment language. When assessments are *relevant* they are closely tied to classroom instruction. Teacher-made assessments may fail to be relevant because the teacher is attempting either to assess skills not taught or to assess those not included in the curriculum. For example, one student in an assessment class reported on going to Back to School Night at her daughter's school:

> I was particularly anxious to meet my daughter's science teacher, Ms. Church, as my daughter was reporting academic difficulty in this class. . . . Ms. Church explained her grading practices and revealed that many of her students currently had low marks in science. . . . According to her, most of the low cumulative grades could be attributed to the low scores earned on the pre-test for the current unit. When I questioned Ms. Church about WHY she would "count" scores earned on a PRE-test, she was unable to answer my question. In fact, she seemed to believe that ALL work should "count." That night, when I got home, I talked with my daughter about her grade on this pre-test and she showed me the huge red "53" scrawled across the top of this paper. We celebrated with a trip to Dairy Queen, after I explained to my daughter that she ALREADY KNEW 53 percent of the material Ms. Church had not yet taught! [Butler, 1999].

This experience is not, alas, unique or uncommon. It is, however, a perfect example of irrelevant, misused assessment.

When assessments are *reliable* they show consistency of scores across evaluators, over time, or across different versions of a test. An assessment is reliable when (1) the same answers receive the same score no matter when the assessment occurs or who does the scoring or (2) students receive the same scores no matter which version of the test they take.

When assessments are *valid,* they measure what they are intended to measure, rather than extraneous features. An example of an invalid assessment of the ability to use a microscope correctly would be to give a pencil and paper test on the parts of the microscope. A more valid assessment would be to hand the student a slide and have him or her focus the slide under low and high power.

Conclusion

As we will show throughout the following chapters, the usefulness of classroom assessment depends on understanding what each assessment does and does not reveal about student learning, using multiple and varied assessments to produce a rounded picture, and applying all that assessment information to the design of future classroom instruction and assessment.

We focus mainly on the formative assessment process but realize that there are many factors inside and outside the classroom that affect how we view and use classroom assessment.

Chapter One has laid the groundwork for our thinking about formative assessment and provided some of the language pertinent to understanding the vast concept of assessment.

To continue our study of formative assessment, Part One, Clarifying Learning Targets, begins by outlining the Classroom Assessment Cycle. Chapters Two and Three explore unpacking the targets and defining our expectations for student learning. To understand what we mean by "unpacking," just think about a suitcase full of clothes. Suppose the suitcase is lost and a claim must be made for the items to be replaced. You would certainly want the most important or most essential items to be on this replacement list. When we "unpack the targets," we are taking a standard course of study and determining the most important learning targets embedded in these standards. Chapters Two and Three provide insights into identifying these most important learning targets.