2

The Language of Assessment

"There are three sides to every story—your side, my side, and the truth."

—John Adams

The goal of assessment is to collect objective evidence that represents the truth about student performance. In order to assure objectivity the assessment plan must be well grounded in the principles of assessment. The first step in developing an objective assessment plan is to become familiar with the terminology of assessment to facilitate your understanding of the bigger picture. The purpose of this chapter is to review the basic terminology and principles of assessment and provide you with a basic understanding of the framework on which to base an objective and comprehensive systematic assessment plan. These concepts are discussed in detail in subsequent chapters.

Many of you are familiar with these terms. Some readers may even prefer to move past this chapter and delve right into the strategies for developing assessment tools. However, as further discussion demonstrates, you cannot start collecting data until your assessment plan is established. Unless you consistently work in the area of assessment, you will find this refresher beneficial. Reviewing this chapter will increase your fluency in the Language of Assessment and your understanding of the proposed guidelines.

Assessment

Chapter 1, "The Role of Assessment in Instruction," introduces you to the concept of assessment as the broad and comprehensive process of collecting quantitative and qualitative data to make informed educational decisions about students. It is a process that encompasses the full range of procedures used to obtain information about student learning. Data collection for assessment should be directed by clearly defined learning targets or objectives (Nitko, 2004). In nursing education, assessment answers the question, "How well has the student achieved the instructional objectives?"

Nitko (2004) proposes five guidelines to help teachers select and use classroom assessments. Nitko's principles (p. 6) provide the basis for developing a plan for systematic assessment of learning outcomes:

1. **Identify** the desired learning targets (instructional objectives) and select the behaviors that represent achievement of the objectives, the learning outcomes.
2. **Ensure** that the selected assessment techniques match the learning outcomes. While assessment techniques should be practical and efficient, it is more important that they are derived from the intellectual challenge posed by the learning outcomes.
3. **Provide** assessment opportunities that meet a learner's needs. Students should be given concrete examples of what is expected of them and the assessment techniques should provide meaningful feedback.
4. **Employ** multiple measurement techniques to assess each learning outcome. The validity of assessment is enhanced by using multiple assessment modalities. A variety of measurements may be required to evaluate whether a student has attained a particular learning outcome, especially if the outcome involves higher-order thinking.
5. **Consider** the limitations of assessment techniques when interpreting their results. It is important to remember that the information obtained, even when multiple assessment techniques are used, is only a sample of a student's behavior and that the interpretation of all assessments is subject to measurement error.

Measurement

Measurement is defined as the process of assigning a score that represents the degree to which some trait, characteristic, or behavior is associated with a person (McMillan, 1997). It encompasses a variety of techniques, including tests, ratings, and observations, that are designed to assign a score that represents the degree of a predefined trait an individual possesses. Thus, measurements provide the information that guides decision making. While valid measurements contribute to valid decisions, erroneous measures can lead to inappropriate decisions. Therefore, it is crucial for educators to ensure their measurement instruments are sound.

Objectivity is an essential element of a trustworthy measurement. If a measurement instrument is not objective, the measurement's results depend more on the subjective opinion of the person who is conducting the measurement rather than on the ability of the person who is being measured. A measurement instrument is objective only if it is confined to assigning a number or a rating to a student characteristic based on predefined objective evidence of the characteristic.

One common measurement error is to equate quantification with objective measurement. Numbers have a scientific quality that can be confused with objectivity. Just because a measurement instrument produces a numerical score does not mean the score is an objective one. A score of 80 percent on a test is meaningless and arbitrary if the score is based on a test that was poorly constructed in the first place.

Table 2.1 Steps for Developing Effective Measurement Instruments

1. Identify the instructional objectives and learning outcomes.
2. Develop a blueprint based on course content and objectives.
3. Create items to measure mastery of content and objectives.
4. Quantify the results of the measurement.

Assessment instruments that provide qualitative information are sometimes chosen as the most desirable instruments for classroom measurement. When an assessment involves a procedure that describes student achievement in qualitative terms, extreme care must be taken to ensure objectivity when assigning a number or category as a score. Whatever technique you choose, it is essential that your measurements are never based on subjective judgments.

In addition, it is very important to acknowledge that measurement skills are not intuitive. The ability to produce measurements that provide valid and reliable results is acquired and develops with practice. One way to immediately improve your measurement skills is to follow the four steps for developing effective classroom measurement instruments (Table 2.1). Each of these steps is incorporated when discussing assessment development throughout this book.

Note that step 1 in Table 2.1 reflects the first of Nitko's (2004) assessment guidelines: Identify the desired learning targets and the instructional objectives. This is also the first step in the development of a systematic plan for assessment. As you read this book you will recognize that the steps for developing an assessment plan overlap and that each reflects Nitko's assessment guidelines.

Evaluation

Assessment, measurement, and evaluation are not equivalent. Evaluation is defined as a value judgment that attaches meaning to the data obtained by measurement and gathered through assessment. It is guided by professional judgment and involves interpreting what the accumulated information means and how it can be used. Evaluation compares student performance with a standard and makes a decision based on that comparison. The standard or outcomes that students are expected to achieve must be established at the beginning of the instructional process. Establishing the behavior standards and clearly communicating them to the students facilitates the evaluation of students' achievement of the learning outcomes. Table 2.2 illustrates the difference between measurement and evaluation.

While evaluation involves a judgment about the merit of an individual's performance, it also involves a judgment about the value of the measurement. Although fair evaluation should be objective, classroom evaluation tends to be subjective because

Table 2.2 Difference Between Measurement and Evaluation

Measurement: The student correctly answered 85 of 100 items on the multiple-choice exam.

Evaluation: The student performed at an above average level.

human judgment is subjective. Therefore, it is a teacher's responsibility to verify that evaluation is based on objective assessments. The more judgments are based on carefully constructed and administered classroom measurement instruments, the greater the likelihood they are objectively sound. Furthermore, the more familiarity you have with the principles of assessment, the greater confidence you will have in the objectivity and ultimate fairness of your student evaluations.

Formative Evaluation

Formative evaluation directs future learning by appraising the quality of student achievement while the student is still in the process of learning. It judges student progress toward meeting instructional objectives with the intent of improving teaching and learning. Formative evaluation is diagnostic evaluation; it identifies students' strengths and weaknesses to provide feedback for improvement of teaching and learning. Formative evaluation also involves judgments about the quality of instruction and assessment as they occur. These judgments allow the teacher to revise instructional materials, clarify objects, and update measurement instruments during a course of instruction. Because formative evaluation is a method that shapes the process of teaching and learning while it is in progress, it should not be used for assigning class grades.

Summative Evaluation

The focus of summative evaluation is to describe the quality of student achievement after an instructional process is completed. While a formative evaluation asks, "How are you doing?", a summative evaluation asks, "How did you do?" (Slavin, 1997, p. 491). A summative evaluation is given at the conclusion of a unit or a course of instruction, and it focuses on determining whether learning has occurred and if the desired outcomes have been achieved. Summative evaluation provides a summary of student achievement and is used to determine students' grades and their progress in an educational program.

Summative and formative evaluation should be consistent. This consistency is achieved when both are based on the instructional objectives established at the beginning of the course. In addition, it is imperative that students know whether an evaluation is formative or summative so they understand if the evaluation is for practice or if grades will be assigned. Table 2.3 compares formative and summative evaluation.

Table 2.3 Comparison of Formative and Summative Evaluation

Formative evaluation **How are you doing?**	**Summative evaluation** **How did you do?**
• Occurs during the process of learning	• Occurs at the completion of instruction
• Assesses progress in a course	• Summarizes achievement in a course
• Directs learning to achieve objectives	• Assesses objective achievement
• Grades not assigned	• Assigns grades
• Provides feedback	• Provides feedback

Instructional Objectives

The first step in the development of an assessment plan is to identify what is expected as a result of a student's course and program experience (Connolly & DeYoung, 2004). A variety of terminology is used to describe the statement of learning intent. In fact, the use of that terminology is widely debated, and too often the debate becomes more important than the logical development of the assessment plan. What is important is that the statements clearly communicate the teacher's expectations to the students. The guidelines for developing a comprehensive assessment plan in this book are based on the careful preparation of instructional objectives. Clearly defined objectives identify the student behavior that is going to be assessed and specify what a student should know and be able to do at the end of an instructional course (Gronlund, 2000).

Objectives are also criticized as limiting the learning experience. In fact, while objectives identify the end point, they do not specify the route that must be taken. Others criticize objectives as focusing on minimal learning. In fact, although well-designed objectives do identify the minimum acceptable learning, they guide students to attain their own *personal best*. Educators must clearly communicate what is the minimum acceptable behavior that indicates success or students will not know what is expected of them. When students are involved in learning experiences that inspire them to achieve their own personal best, they are most likely to develop a love of learning, which will compel them to strive for personal excellence throughout life.

As Robert Mager stated, "When clearly defined goals are lacking, it is impossible to evaluate a course or program efficiently, and there is no sound basis for selecting appropriate materials, content, or instructional methods" (1997, p. 3). The development of instructional objectives and learning outcomes is the focus of Chapter 3, "Developing Instructional Objectives."

Learning Outcomes

The most effective way to state instructional objectives is in terms of the behaviors that you expect students to achieve by the end of a course. Gronlund (2000) maintains that defining objectives in terms of desired student learning outcomes shifts the focus from the learning process to the learning outcomes and also provides a basis for the assessment of student learning. Gronlund also suggests that stating the general objective first and then listing a representational sample of learning outcomes clarifies to the student what is deemed to be acceptable by the teacher as evidence that the student has attained the objective. Chapter 3, "Developing Instructional Objectives," expands on this approach for student assessment.

Blueprint

A blueprint, also referred to as a table of specifications or test plan, is the foundation for establishing validity evidence that a test represents the content of the course. A test cannot include the entire instructional domain of a course, yet it should include a representative sample of that domain. A blueprint is defined as a mechanism that guides the systematic selection of a representative sample of the content and objectives of a course. A test based on a carefully planned blueprint enables you to project that a student who receives a score of 90 percent on a 50-item test would receive a score of 90 percent on a 500-item test.

Content validation of a classroom achievement test involves collecting evidence to evaluate the degree to which the test reflects the course's instructional objectives and content. Establishing evidence of validity based on test content must begin with test development. It only makes sense to plan ahead. To make sure that a test represents the desired outcomes of a course, it must be based on a blueprint. The blueprint guides the selection of test questions that reflect achievement of the content and course objectives.

A blueprint answers the question, "What is being measured?" Although a blueprint directs the selection of test items, it is still the teacher's responsibility to carefully plan and develop these items to ensure that they actually measure student ability in the areas specified by the blueprint. Blueprints can be developed for a variety of measurement techniques—to enable the selection of the most appropriate instrument to measure the attainment of the course objectives, content, and skills.

Test development is a time-consuming process. However, using a blueprint as a guide expedites this process and provides the structure for obtaining valid and reliable test results. The effort required for plan development is time well spent. In the long run it facilitates test development and increases your confidence in the decisions you make based on your measurement instruments. Chapter 5, "Implementing Systematic Test Development," provides detailed guidelines for blueprint development.

Item Bank

An item bank is defined as an organized collection of items that can be accessed for test development. Testing experts often distinguish between item *pools* and item *banks.* This distinction defines a bank as a set of items whose difficulty levels have been calibrated on a common scale, while a pool simply consists of a collection of items. Because the term *item bank* is commonly used in test development software designed for classroom use, it is used throughout this book. Although a classroom item bank is designed to accumulate item data, the difficulty levels of the items in the classroom item banks referred to in the following chapters are not calibrated. The implementation of an item banking program is closely examined in Chapter 16, "Instituting Item Banking and Test Development Software."

Test

Tests are measurement instruments: formal events where individuals are asked to demonstrate their achievement of some knowledge or skill in a specific domain. The purpose of an achievement test is to obtain relevant and accurate data needed to make important decisions with a minimum amount of error. Gronlund (2004) describes a test as a tool for measuring a sample of student performance. It can be assumed that students have achieved the course learning objectives in the entire content domain when a designated score is obtained on a test that is designed to sample the content appropriately (Nitko, 2004).

Using a single test or type of measurement instrument is not a satisfactory assessment strategy. Most course objectives require a variety of measurement and evaluation strategies to determine student competency in a particular course. The selection of measurement instruments depends on the outcomes to be measured. It is important to select the most appropriate strategies for measuring each learning outcome. One premise of this book is that multiple-choice exams can be developed to contribute to the assessment of objectives that require higher-level cognitive ability, including the construct of critical thinking.

An achievement test should consist of a sampling of tasks, which represents the larger domain of behavior that is being assessed (Gronlund, 2004). When students complain that an exam did not cover the course content, it may indicate that there is a mismatch between the test items and the larger domain of course content or objectives, or it may indicate that the items did not address the designated content or objectives. It is not possible to measure a student's achievement of objectives with items that do not match those particular objectives. Therefore, the challenge is to develop a blueprint for the test and write items to match the objectives and content being assessed.

Interpreting Test Scores

A raw test score is meaningless without a framework for interpretation. The raw test score is only given meaning within the instructional content domain it represents. Criterion-referenced tests (CRTs) assess an individual's performance based on the percentage of the content mastered. Norm-referenced tests (NRTs) define an individual's performance by comparing it with others. Although both types of interpretation can be applied to the same test, the interpretation is most meaningful when the test is specifically designed for a desired interpretation (Linn & Gronlund, 2000).

Criterion-Referenced Tests

CRTs interpret a student's raw score using a preset standard established by the faculty. Thus, each student's competency in relation to the preset standard is measured without reference to any other student. Student scores are then reported as the percentage correct with each student's performance level determined by the preset, or absolute, standard. Figure 2.1 presents an example of a criterion-referenced objective.

Because CRTs measure a student's attainment of a set of learning outcomes, no attempt should be made to eliminate easy items. The content chosen for a CRT depends only on how well it matches the instructional objectives of the course (Nitko, 2004). If most students in a group meet the standard, the group scores will obviously cluster at the high end of the grading scale.

CRTs are often teacher-made and are closely tied to the objectives and curriculum. They are most meaningful when they are specifically designed to measure student ability in a particular area (Gronlund, 1973). Competency is such a critical requirement in nursing education that the CRT is often the preferred form of classroom testing in nursing education (Reilly & Oermann, 1990).

Gronlund (1973) describes the relationship of criterion-referenced testing to the two levels of learning: mastery and developmental. Designing tests for these two different levels of learning poses different challenges.

Figure 2.1 Example of a criterion-referenced score

The student demonstrated mastery by correctly identifying 90 percent of the terms.

Mastery Learning At the mastery level, CRTs are concerned with measuring the minimum essential skills that indicate mastery of an objective. The scope of learning tasks is limited, which simplifies the process of assessment. A score of the percentage correct is usually used to identify how closely a student's score demonstrates a complete mastery of an objective. One challenge for the faculty is to identify (1) which specific objectives the students are expected to master and (2) which objectives represent learning beyond the mastery level, or developmental learning (Gronlund, 1973). Chapters 3, "Developing Instructional Objectives," and 4, "Assessing Critical Thinking," offer a more in-depth discussion and also provide examples of objectives at the mastery and developmental levels of learning.

Developmental Learning The concept of developmental learning applies to constructs that represent complex higher-order thinking, such as critical thinking. The abilities associated with this level are continuously developing throughout life. Objectives for developmental learning represent goals to work toward, with emphasis focused on continuous development rather than a complete mastery of a set of predetermined skills (Gronlund, 1973).

Learning outcomes at the developmental level represent degrees of progress toward an objective. Because it is impossible to identify all the behaviors that represent a complex construct, only a sample of the behaviors associated with instructional objectives at this level can be identified as learning outcomes. These behaviors should define the construct and provide a representational sample of student performance that will be accepted as evidence of the appropriate progress toward the attainment of the ultimate objective.

Students are not expected to fully achieve objectives at the developmental level. However, they are required to demonstrate the behaviors represented by the learning outcomes, and they are also encouraged to strive for their personal level of maximum achievement toward the ultimate objective—their *personal best.* At this level instructional objectives can be designed to show the development of students as they progress through an instructional program. For example, the same general instructional objectives can be used in every course in a nursing program, with the learning outcomes becoming more complex as the students progress through the program. Developing objectives for mastery and developmental learning is reviewed in Chapter 3, "Developing Instructional Objectives."

Gronlund asserts that the use of CRTs is restricted with the assessment of developmental learning. While test preparation should follow mastery level procedures, he suggests that to adequately describe student performance beyond minimal essentials, tests at the developmental level should include items of varying difficulty and allow for both criterion and norm-referenced interpretations (1973).

Norm-Referenced Tests

While CRTs measure a student's achievement of a program's objectives without reference to other students, the aim of NRTs is to compare a student's achievement with the achievement of the student's peer group. NRTs focus on a student's performance in relation to other students rather than in relation to the attainment of a course's objectives. Norms themselves do not represent levels of performance; they provide a frame of reference to use when comparing the performances of a group of individuals. NRTs interpret a student's raw score as a percentile rank in a group. NRTs do not indicate

Figure 2.2 Example of a norm-referenced score

The student's performance equaled or exceeded 82 percent of the students in the group.

what a student has achieved; the tests only indicate how the student compares with other students in their group. An example of a norm-referenced score is shown in Figure 2.2.

NRTs are designed to discriminate between strong and weak students. The tests are developed to provide a wide range of scores so that the identification of students at different achievement levels is possible. Therefore, items that all students are likely to answer correctly are eliminated.

The content selected for an NRT is based on how well it ranks students from high to low achievers (Nitko, 2004). The NRT format is commonly used on national standardized tests. These tests have a generalized content that is commonly taught in many schools. The norms established by a standardized achievement test are based on nationally accepted educational goals, which enable educators to compare a student's test score with the scores of other students in similar programs in the United States. These scores provide a general indication of the strengths and weaknesses of the students in a particular school and afford faculty an external reference point for comparing their curriculum with a composite national curriculum.

NRTs in the classroom setting identify how students compare with each other. Because strict NRTs are not concerned with the level of individual student achievement, they are usually not appropriate for classroom use. When assessing developmental learning, Gronlund (1973) suggests using NRTs to rank students with the addition of criterion-referenced interpretations applied to the test to assess degrees of student progress toward an objective. Chapter 5, "Implementing Systematic Test Development," elaborates on the use of NRTs and CRTs when determining how difficult a test should be. Table 2.4 compares CRTs and NRTs.

High-Stakes Tests

The term *high stakes* is commonly used among test developers when referring to a test whose results are the basis for making life-altering decisions about people. For example,

Table 2.4 Comparison of Criterion- and Norm-Referenced Tests

Criterion-referenced test	Norm-referenced test
• Compares student performance to pre-established criteria	• Compares student performance to reference group
• Describes the performance	• Discriminates the performance
• Mastery reference	• Relative performance reference
• Narrowly defined content domain	• More diverse content domain
• Larger number of items for each objective	• Smaller number of items for each objective
• Includes easy items	• Eliminates easy items
• Focuses on student competency	• Focuses on student ranking
• Provides percent-correct score	• Provides percentile rank

a licensure examination is a high-stakes test because the examinee' scores on the test determine whether or not they will be allowed to practice nursing. When the results of one test are used to determine whether an individual will be licensed, the test results must have very high evidence of reliability and validity.

Classroom exams in nursing meet the criteria for being high-stakes examinations. Life-altering decisions are certainly made based on the results of these exams. Classroom exams do differ from licensure examinations because decisions are not based on the results of one exam but rather on the accumulation of scores over a semester's worth of exams. However, because the decisions that are made based on the results of classroom exams can have a profound impact on students' lives, it is obvious that faculty must pay careful attention to develop classroom exams that produce trustworthy results.

Grade

While a test score is a numerical indication of what is observed from a single measurement instrument, a grade is a label representing a composite evaluation. A course grade should be derived from the accumulation of scores obtained from several measurement instruments. Because life-altering decisions are associated with student grades, the utmost care must be used when assigning test scores and grades. Chapters 13, "Interpreting Test Results," and 15, "Assigning Grades," both discuss test analysis and grading procedures.

Test Bias

A biased test is one that discriminates against a certain group based on socioeconomic status, disability, race, ethnicity, or gender (Slavin, 1997). When a measurement is biased, students who have the same ability perform differently on the same task because of their affiliation with a particular ethnic, sexual, cultural, or religious group (Hambleton & Rodgers, 1995). Hambleton and Rodgers define stereotyping as another undesirable characteristic of a test that introduces bias. Stereotyping refers to the representation of a group in such a manner that it may be offensive to the group members. They also note that test language that is offensive can obstruct the purpose of a test when it produces negative feelings, which affect the students' attitudes toward the test and thus influences their test scores (Hambleton & Rodgers, 1995). Test bias in a nursing exam refers to the difference in a group's mean performance based on non-nursing elements in the exam, which are elements not familiar to the group.

An assessment is not fair if some students have an advantage because of factors unrelated to purpose of the assessment (McMillan, 2001). The aim of a nursing test is to measure knowledge that is essential to safe nursing practice in the United States. Reading speed, vocabulary ability, or familiarity with cultural practices, which are unrelated to health, should not influence a student's score (Klisch, 1994). Therefore, it is important for teachers to collaborate with each other when developing a nursing exam. Every test should be carefully reviewed by at least two faculty members for items containing language that could offend or be misunderstood. Items with overt cultural or gender bias should be rejected. Items referring to events that are common to one culture but not to another should also be eliminated. All tests should be carefully edited to remove stereotypical language. In fact, even the most innocent vocabulary can introduce

Figure 2.3 Example of a culturally biased item

Biased Question:

A client who is taking a medication that is a sedative says to a nurse, "I am responsible for the *carpool* tomorrow." Which of these directions should the nurse give to the client?

The term "carpool" could be unfamiliar to individuals for whom English is a second language or for those who live in urban areas and depend on public transportation.

bias into a test (Figure 2.3). Although offensive, demeaning, or emotionally charged material may not make an item more difficult, it can cause students to become distracted, thus lowering their overall performance (Hambleton & Rodgers, 1995).

Bosher (2002) defines linguistic bias as resulting from students' inability to understand an item because the language is so complex. Students for whom English is a second language (ESL) are particularly susceptible to linguistic bias. Poorly written test items can introduce structural bias into a test. Items that are grammatically incorrect, ambiguous, or vaguely worded confuse all students, but they particularly confuse ESL and learning disabled students (Klisch, 1994). Each question should be succinctly written so that all students have a clear understanding of its meaning the first time it is read.

Although humor can be a useful tool in the classroom, it can be a distraction in an exam. Students are not inclined to *get the joke* during an exam, particularly ESL students. In fact, test anxiety can increase when students do not understand why others are laughing. Haladyna (2004) points out that humorous items reduce the number of plausible options and therefore make the items easier for those students who understand the joke. The detailed item development guidelines, presented in Chapter 6, "Selected-Response Format: Developing Multiple-Choice Items," will assist you in eliminating bias from your test items.

The following guidelines will help to keep your exams free from language that can be offensive or introduce bias:

- Avoid use of gender or names.
- Refrain from stereotyping members of a group.
- Keep humor in the classroom and out of the test.
- Use references to race, culture, religion, marital status, or sexual orientation only if it pertains to the problem in the question.
- Avoid terminology that refers to popular culture and is unrelated to health issues.
- Eliminate vocabulary that has a different or unfamiliar meaning to different ethnic groups.

Reliability

Test reliability is very important to test developers and users. You would have little confidence in a standardized nursing achievement test that ranked a student in the top five percent last week but places the same student near the mean this week. Reliability refers to the degree of consistency with which an instrument measures an attribute for a particular

Figure 2.4 Reliability requirement for validity

A test can be reliable without being valid.
HOWEVER
A test cannot be valid unless it is reliable.

group. Reliability is not a property of the test itself; the test is not reliable. Reliability refers to the reproducibility of a set of scores obtained from a particular group, on a particular day, under particular circumstances (Frisbe, 1988). Achievement test results that are reliable are consistent, reproducible, and generalizable. That is, a second measurement with the same test on the same individual would obtain the same result. However, because every measurement contains error, you should expect some variation in test performance. It is highly unlikely that your efforts at obtaining a second measurement would produce precisely the same scores as the first measurement.

Reliability can be quantified by several statistical formulas. These estimates provide a reliability coefficient, or a measure of the amount of variation in test performance. While there are several procedures for obtaining a test's reliability estimate, the procedures that are most frequently reported by test analysis software estimate a test's reliability based on the internal consistency of the test. These reliability estimates range from zero to one, with zero indicating no reliability and one indicating perfect reliability. Reliability is discussed at length in Chapter 12, "Establishing Evidence of Reliability and Validity."

Validity

Although a test must be reliable to be valid, a reliable test is not always valid (Figure 2.4). A test can have high reliability and yet not really measure anything of importance or it can fail to be an appropriate measure for a particular use (Burns & Grove, 1997). Therefore, we can have reliable measures that provide the wrong information.

The American Educational Research Association (AERA), American Psychological Association, and the National Council on Measurement in Education agree that, "Validity is the most fundamental consideration in developing and evaluating tests" (AERA et al., 1999, p. 9). Validity is not a property of the test itself. It refers to the appropriateness of the interpretation of the test scores—the extent of the evidence that exists to justify the inferences we make based on the results of the test. A test can have substantial evidence of validity for one interpretation and not for another. For example, an exam can have considerable evidence of validity for interpretations related to acceptance into a city's police department, whereas the same exam can be of no use for admission to the same city's fire department. This is a perfect example of why you cannot use an exam with validity evidence that supports its use to assess theoretical nursing knowledge to also assess a construct such as critical thinking, unless you can collect validity evidence to justify the test's use to measure critical thinking.

Validity does not exist on an all-or-none basis. A test is always valid to some degree—high, moderate, or weak in a particular situation with a particular sample. Validity is a matter of judgment—there are no fixed rules for deciding what is meant by high, moderate, or weak validity. Skill in making these judgments is based on test validation, and it develops with experience in dealing with tests (McMillan, 2001). Test validation is defined as the process of collecting evidence to establish that the inferences, which are based on the test results, are appropriate. The first step in the process of test validation is to have a clear understanding of the evidence that establishes validity.

The traditional approach to establishing validity identified three distinct classifications of validity: content validity, construct validity, and criterion-related validity. Today, however, validity is not viewed as three distinct types. The 1985 edition of the *Standards for Educational and Psychological Testing* identifies validity as a unitary concept that includes each of the categories as evidence of validity:

- Content-related evidence
- Construct-related evidence
- Criterion-related evidence

Content-related evidence, criterion-related evidence, and construct-related evidence are interrelated; ideal validation includes several types of evidence, spanning all three of the traditional categories (AERA et al., 1985, p. 9). This approach emphasizes that validity is not an all-or-none proposition. It is a matter of degree and involves the judgment that you make after considering all the accumulated evidence.

The most recent edition of the *Standards for Educational and Psychological Testing* refers to types of validity evidence rather than categories of validity. In fact, in an attempt to emphasize that validation is a process of collecting a variety of evidence to support a proposed interpretation of a test, this edition does not use the traditional nomenclature. Rather, it outlines the various sources of evidence that can be used for evaluating the proposed interpretation of a test's score for a particular purpose (AERA et al., 1999, p. 11). The types of validity evidence described in the 1999 *Standards* include

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence based on the consequences of testing

When reviewing the different types of validity evidence, it is essential to keep the unitary nature of validity in mind. Types of validity evidence do not exist exclusively or separately; they overlap. They are all essential to a unitary concept of validity. Evidence from each one may be needed when attempting to validate the interpretation of a test score. The importance of selecting, developing, and using tests based on adequate validity evidence for interpreting the test scores for a particular purpose cannot be understated (Goodwin, 2002, p. 101).

Evidence Based on Test Content

Evidence based on test content represents the degree to which the items on a test reflect a course's content domain. Content-related validity is nonstatistical (Lyman, 1998); it cannot be objectively quantified with a number. Rather, the documentation of content-related evidence of validity begins with test development and is established by a detailed examination of the test content. The more closely related a test is to its blueprint, the higher the content validity will be. If a test has content-related evidence of validity, then we can use the test results to make a judgment about the person's knowledge within that specific content domain.

A well-constructed test measures every important aspect of a course, including the subject matter and the course objectives (Anastasi & Urbina, 1997). Because a test measures only a sample of a domain, the degree to which the test items represent the content of the course is the key issue in content validation. No aspect of a course should be under or over represented. The validity of the inferences based on the test results depends on how well the test sample represents the domain being tested (Worthen, Borg & White, 1993, p. 182). A blueprint establishes validity evidence based on test content by ensuring that a test provides a representative sampling of the objectives and content domain of a course. Chapter 5, "Implementing Systematic Test Development," presents detailed guidelines for developing blueprints for your classroom tests.

Content-related evidence of validity is a central concern during test development (AERA et al., 1985, p. 11). Tests that provide content valid results are produced with careful planning. When developing a test to inform decisions about student progression in a course of study, the content domain on the test must be limited to what the students have had the opportunity to learn during the course (AERA et al., 1999, p. 12).

Standardized tests use a national panel of experts in the field being measured to establish validity evidence based on test content. When you develop a classroom test, you do not have access to a panel of experts. However, you can strengthen the evidence for the validity of the decisions you make based on your tests' results by following the steps for enhancing validity evidence based on test content (Figure 2.5).

Evidence Based on Response Processes

This type of validity evidence was formerly a component of construct-related evidence. A construct is an unobservable characteristic of an individual that cannot be measured directly, such as intelligence, creativity, and critical thinking. The 1999 *Standards* focus on whether the questions are in fact measuring the intended construct or are irrelevant factors inherent in the questions influencing the performance of subgroups of examinees. Evidence based on response processes involves the collection of evidence that supports the assertion that a test measures a construct by measuring the observable

Figure 2.5 Steps for enhancing validity evidence based on test content

- State objectives in performance terms
- Identify learning outcomes
- Define the domain to be measured
- Prepare a detailed blueprint
- Write items to fit the blueprint
- Select a representative sample of items for the test
- Ask colleagues to review your blueprint and items
- Provide adequate time for test completion
- Review item and test analysis
- Use the test only for its intended purpose

behaviors, which demonstrate the construct as defined by the test developer (Worthen et al., 1993, p. 187).

Evidence Based on Internal Structure

This type of validity evidence was also a component of construct-related evidence in the 1985 *Standards.* Construct validation begins with test development, and it continues until the evidence establishes a relationship between the test scores and the construct. For example, a test claiming to measure critical thinking would require construct validation. First, a detailed definition of the construct of critical thinking, which is derived from psychological theory, prior research, or systematic observation and analyses of the behavior domain, must be developed (Anastasi & Urbina, 1997, p. 138). The definition should delineate the aspects of the construct that are to be represented in the test. Then, the objectives and learning outcomes that correlate with the definition must be specified. Once this is completed the test is blueprinted, and items are developed that require students to demonstrate the behaviors that define the construct of critical thinking.

A variety of methods can be used to collect data to establish evidence based on response process and internal structure. For this example, a new critical thinking test. Methods include:

- Investigating differential item functioning to identify whether the items are functioning differently for different groups (AERA et al., 1999, p. 13).
- Obtaining intercorrelations of test items to provide evidence of item homogeneity, which supports the assertion that the new test is measuring one construct, in our example, critical thinking (AERA et al., 1999, p. 13).
- Correlating the score on the new critical thinking test with scores from other instruments, which have demonstrated ability to measure critical thinking. A positive correlation would support the identification of critical thinking in the new test.
- Questioning the takers of the new test about their thinking strategies by asking them to think aloud about their mental processes as they answer the questions. This supports the definition of the construct and provides evidence of the cognitive processes involved with critical thinking (AERA et al., 1999, p. 12).
- Asking experts in the area of critical thinking to judge the relationship of the items on the test with the construct of critical thinking (AERA et al., 1999, p. 11) to verify that the items are measuring the construct.

A multitude of commercially prepared examinations maintain that they assess critical thinking. It is very important for you to be an informed consumer when purchasing one of these examinations to administer to your students. A variety of evidence must be collected to establish that a test is measuring the construct it purports to measure. When reviewing a standardized examination that claims to measure critical thinking, or another construct, it is important to closely examine its reliability evidence and to ask the following questions:

- What is the definition of critical thinking?
- What are the objectives of the test?

- What are the learning outcomes on which the questions are based?
- What is the structure of the test's blueprint?

You should expect the test developer to answer these questions and to provide information about the experts who were involved at every level of the test development process. The developer should also provide you with data about the test's reliability coefficient. In addition, you should also expect the developer to report the evidence they have accumulated to support the validity of the proposed interpretations of the test. If the answers to these questions are unavailable or unclear, select another test.

One of the goals of this book is to provide a framework for faculty to develop multiple-choice items that assess critical thinking. Ask yourself the same questions that you would ask when evaluating a standardized test. Once these questions are addressed, you can write multiple-choice items to measure the behaviors that provide evidence of critical thinking abilities. Chapters 4, "Assessing Critical Thinking," and 7, "Writing Critical Thinking Multiple-Choice Items," discuss the development of multiple-choice test items that measure critical thinking abilities in greater detail.

Evidence Based on Relation to Other Variables

This type of evidence examines the relationship of test score to variables that are external to the test (AERA et al., 1999). The 1985 *Standards* referred to this as criterion-related evidence of validity, which demonstrated whether test scores are systematically related to one or more outcome criteria (AERA et al., 1985, p. 11). The focus of predictive evidence is to determine how valid a test is at predicting a second measure of performance—the criteria. A study of concurrent evidence, however, is concerned with estimating present performance when compared to the criterion. The key question with criterion-related validity is "How accurately do test scores estimate criterion performance?" (AERA et al., 1999, p. 14).

As Lyman (1998) explains, concurrent and predictive evidence differ only in their time sequence. Both test scores and criterion values are obtained at about the same time with concurrent validity. However, in predictive validity there is a time lapse between testing and obtaining the criterion values. When criterion-related evidence is high, the test can be used to estimate performance on the criterion.

If you are using a test score to predict future performance, you must be concerned with determining the degree of the relationship between the test and the criterion (the future performance). Support of criterion validity must include empirical evidence on the comparison between test performance and performance of the criterion (Rudner, 1994). Many tests are currently being marketed that claim to predict student success on the National Council Licensure Examination (NCLEX). When evaluating these predictor examinations, it is important for you to determine how they have established criterion-related evidence of validity. You should be able to answer this question: How does the test predict the performance of the students on NCLEX? The predictor test should compare an individual's test scores to NCLEX pass/fail status to provide a basis for predicting the likelihood of passing or failing NCLEX based on the score on the predictor test.

Beware of exams claiming to have an extremely high accuracy rate for predicting the passing rate on the NCLEX examination. Look closely at their statistics. For example, if a company says that it can predict NCLEX success with 98 percent accuracy, what test score does a student need to obtain to qualify for the passing prediction? If, for example, a test predicts success for students who answer more than 90 percent of the test questions correctly, what is the test really predicting? When fewer than 20 percent of a group taking a test are predicted to pass, of what use is the prediction? A conservative estimate of the students who will pass is a safe approach to predicting NCLEX outcomes. When a company predicts that 20 percent of a group of students will pass, does the company also predict how many will fail? Find out how accurate the test is at predicting students who will fail NCLEX. The prediction of failure is much more useful—particularly if the test report delineates the students' weaknesses and proposes a plan for remediation.

Most faculty can accurately identify the top 20 percent of their students, based on their history of classroom test results. In addition, 87.3 percent of first time U.S. educated candidates passed the NCLEX-RN in 2005 and 89.1 percent of first time U.S. educated candidates passed the NCLEX-PN in 2005 (National Council of State Boards of Nursing, 2006g). A test that predicts the success of a very small number of students—who will obviously pass—is really predicting nothing at all! Chapter 11, "Preparing Students for the Licensure Examination: The Importance of NCLEX," presents an in-depth discussion of the issues related to licensure.

Evidence Based on the Consequences of Testing

The 1999 *Standards* focuses attention on the intended and unintended consequences of using test results to make decisions about different groups. This type of evidence answers these questions posed by Goodwin (2002, p. 104):

- To what extent are the anticipated benefits of testing being realized?
- To what extent do unanticipated benefits (positive and negative) occur?
- To what extent are differential consequences observed for different identifiable subgroups of examinees?

The 1999 *Standards* call for the test validation process to provide evidence that the intended benefits of testing are being realized (p. 16). Test developers must support the claims they make for the benefits of using a particular test score as the basis for making decisions that affect peoples' lives.

Face Validity

Face validity is not validity in the technical sense; it refers to what a test appears to measure, not what it actually measures. Face validity means that the appearance of the test coincides with its use (Popham, 1999). While actual validity is far more important than face validity, face validity is still desirable. A test needs face validity so that it *appears* to be valid to the test consumer. Face validity also helps to keep the motivation of the test takers high, because students seem to try harder when a test appears to be reasonable and fair (Lyman, 1998). In fact, a test that appears irrelevant or inappropriate creates a diversion and can even result in poor cooperation from the test takers (Anastasi & Urbina, 1997). Students respond positively to tests that represent the content and objectives of the course. Tests that students perceive as being unrelated to course content can be distracting and therefore decrease the test's reliability.

It is helpful for a test to have face validity, as long as it has demonstrated evidence of actual validity (Polit & Hungler, 1999). Face validity by itself never provides sufficient basis on which to establish validity; the mere appearance of validity is not adequate to establish evidence of validity. We must still establish evidence that enables us to be confident in the decisions we make based on the test's scores.

Usually, when you establish evidence of validity for the interpretation of test scores, face validity is also established. Poor test item construction is a primary cause of inadequate face validity. Thus nursing exams should refer to nursing situations. Developing an exam blueprint and including a nurse and a client in the questions add to the face validity of your nursing exams. Sharing the blueprint with the students before the test alerts them about what to expect on the test and also increases their perception of the test as a valid measurement. Chapter 12, "Establishing Evidence of Reliability and Validity," offers additional discussion related to validity.

Basic Test Statistics

Test analysis is a powerful tool that you can use to increase the quality of your classroom exams and your confidence in the decisions you make based on the test results. In addition, item analysis is an invaluable guide for improving the reliability and validity of the results of future tests by directing the improvement of the individual test items. Before you can analyze test and item data and correctly interpret their meanings, it is important that you understand the basic concepts of test statistics. Appendix A, "Basic Test Statistics," provides a brief reference guide to help familiarize you with the terms related to test and item analysis, which are used throughout the book. Each of these definitions is examined in greater detail in Chapters 13, "Interpreting Test Results," and 16, "Instituting Item Banking and Test Development Software."

Summary

Assessment procedures do not make decisions about students; teachers make decisions about students. To develop procedures that ensure fair decisions, it is important to have a clear understanding of the principles of assessment. This chapter presents an overview of the terminology that is fundamental to a thorough understanding of the concepts underlying valid and reliable assessment procedures. Many of these concepts are explained in greater detail in subsequent chapters. This book explores the entire assessment process and offers guidelines for the development of instruments that provide valid and reliable results, which are an integral component of a plan for the systematic assessment of learning outcomes. Familiarity with the *Language of Assessment* is the basic requirement for establishing a comprehensive assessment plan.