The Language of Assessment

CHAPTER 2

"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. To ensure objectivity, the assessment plan must be 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 principles of assessment and the terminology as it is defined in this text. This review will provide a basic understanding of the framework on which to base an objective, comprehensive, systematic assessment plan. These concepts are elaborated on in subsequent chapters.

Many of you are familiar with these terms. However, the terms used in assessment can be confused or interchanged with one another. This confusion can lead to misinterpretation of assessment results. The review in this chapter is intended to clarify the meaning of assessment terminology as it is referred to in this text and enhance your understanding of the content.

Assessment

Assessment is the broad and comprehensive process of collecting quantitative and qualitative data to make informed educational decisions about students, as introduced in Chapter 1, "The Role of Assessment in Nursing Education." The terms *assessment*, *test*, and *measurement* are often confused because they are all involved in the same process.

Assessment involves collecting information. It is a comprehensive process that involves a wide range of strategies used to gain information to make decisions about student achievement. Assessment data also inform decisions about teaching and learning strategies and the efficacy of the individual courses and the overall curriculum. Data collection for assessment should be directed by clearly defined learning targets or objectives (Brookhart & Nitko, 2019). In nursing education, assessment answers the question, "How well has the student achieved the instructional objectives?"

Brookhart and Nitko (2019) propose five guidelines to help educators select and use assessments. These principles 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. Although 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. (Brookhart & Nitko, 2019, p. 27)

Measurement

Measurement is the process of assigning a score that represents the degree to which an individual possesses a characteristic or behavior according to a specific plan (Miller et al., 2012). 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. Whereas valid measurements contribute to valid decisions, erroneous measures can lead to inappropriate decisions. Therefore, it is crucial for educators to ensure that their measurement instruments are trustworthy.

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 systematic 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 90% on a test is meaningless and arbitrary if the score is based on a test that was poorly constructed in the first place. After all, 90% of nothing is nothing.

Measurement instruments that provide qualitative information are sometimes chosen as the most desirable instruments for measurement. When a measurement instrument 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 measurement instruments 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; it develops with practice. Following the four steps for developing effective measurement instruments, as identified in **Box 2.1**, will give you a head start. Each of these steps is incorporated when discussing assessment development throughout this text.

Note that step 1 in Box 2.1 reflects the first of Brookhart and Nitko's (2019) 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 text, you will recognize that the steps for developing an assessment plan overlap and that each reflects Brookhart and Nitko's assessment guidelines.

Evaluation

Assessment, measurement, and evaluation are not equivalent. *Evaluation* is defined as the process of making a value judgment that attaches meaning to the data obtained by measurement and gathered through assessment (Brookhart & Nitko, 2018). 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

Box 2.1 Steps for Developing Effective Measurement Instruments

1. Create the instructional objectives and learning outcomes.
2. Design a test blueprint based on course content and objectives.
3. Compose items to measure mastery of content and objectives.
4. Assemble the items to generate a test that addresses the blueprint.
5. Quantify the results of the measurement.

Box 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.

facilitates the evaluation of students' achievement of the learning outcomes. **Box 2.2** illustrates the difference between measurement and evaluation.

Although 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, clinical, and online evaluation are at risk of being subjective because human judgment is subjective. Therefore, it is an educator's responsibility to verify that evaluation is based on objective measurement instruments. The more judgments are based on carefully constructed and administered measurement instruments, the greater the likelihood that 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 monitors learning progress during instruction. It directs future learning by appraising the quality of student achievement while the student is still in the process of learning (Yang et al., 2019). 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 constructive feedback.

Formative evaluation provides guidance to both students and educators. It involves judgments about the quality of instruction and learning as they occur (Miller et al., 2012). These judgments allow the educator to revise instructional materials, clarify objectives, update learning outcomes, and revise 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. Whereas a formative evaluation asks how a student is doing in a course, summative evaluation asks how the student did at the end of the unit or course (Slavin, 2018). 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 objectives have been achieved. The main purpose of summative evaluation is to assign a course grade.

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 that they understand if the evaluation is

Table 2.1 Comparison of Formative and Summative Evaluation

Formative Evaluation	Summative Evaluation
• Occurs anytime during the process of learning • Assesses progress in a unit or course • Directs learning to achieve objectives • Low-stakes testing • Grades not assigned • Offers feedback to student and educator	• Occurs at the completion of instruction • Summarizes achievement in a unit or course • Assesses objective achievement • High-stakes testing • Assigns grades • Provides feedback to student and educator

for practice or if grades will be assigned. Summative evaluation should be viewed as a final evaluation of achievement and assignment of grades (Yang et al., 2019). **Table 2.1** compares formative and summative evaluation.

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 (Gronlund & Brookhart, 2009). Various terms are 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 (Glennon, 2006). Whatever term you use (objectives, outcomes, competencies, etc.), what is important is that the statements are consistent and clearly communicate the educator's expectations for what is required to pass the course to the students.

A very effective method for developing instructional objectives is described by Gronlund and Brookhart (2009). They recommend stating the objectives as a broad statement and then defining them in terms of the intended student learning. The definition of *learning outcomes* used in this text follows the suggestion of Gronlund and Brookhart. Learning outcomes specify behaviors and clarify what performance you are willing to accept as evidence that the student has achieved the course objectives.

Objectives are sometimes criticized as limiting students' learning experience. In fact, although objectives identify the end point, they do not specify the route that must be taken. Objectives are also criticized as focusing on minimal learning. Although well-designed objectives do identify the minimum acceptable achievement, they also guide students to attain their own personal best. Educators must clearly communicate what minimum acceptable behaviors students must achieve to demonstrate 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 (1997) 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" (p. 3). The

development of instructional objectives and learning outcomes is elaborated in Chapter 3, "Developing Instructional Objectives."

Learning Outcomes

The most effective way to define an instructional objective is to establish the behaviors that you expect students to achieve by the end of a course. Gronlund and Brookhart (2009) maintain 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. 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 educator as evidence that the student has attained the objective (Gronlund & Brookhart, 2009). Chapter 3, "Developing Instructional Objectives," expands on this approach for student assessment.

Blueprint

To decide whether a student passes or fails a course based on the results of a test, the educator must have evidence that the test actually represents the course. The blueprint, or table of specifications, is the foundation for validity evidence because it is the framework for the test. To establish validity evidence, the blueprint must incorporate only the objectives and content that are included in the course.

A blueprint is most effective when it is represented as a two-dimensional table that relates the objectives to the course content. A two-dimensional table requires that every item on the test be classified in terms of both objectives and content (Gronlund & Brookhart, 2009). A blueprint that is set up as a two-dimensional table is the foundation for establishing validity evidence that a test represents the content, as well as the objectives, of the course. When the blueprint guides the selection of test questions that reflect achievement of both the content and course objectives, evidence of validity is established.

A test cannot include the entire instructional domain of a course, but it must include a sample of that domain that is a true representation of the course in order to provide validity evidence for decisions that are made based on the results of the test. A blueprint is 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% on a 50-item test would receive a score of 90% on a 500-item test.

A blueprint answers the question, "What is being measured?" Although a blueprint directs the selection of test items, it is still the educator's responsibility to plan carefully and develop test items to ensure that they actually measure student ability in the areas specified by the blueprint.

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 blueprint 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 4, "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, whereas a pool simply consists of a collection of items. Because the term *item bank* is commonly used when referring to collections of items for quiz, test, or exam use, it is used throughout this text. Although an item bank can be used to accumulate item data, the difficulty levels of the items in the item banks referred to in this text are not calibrated.

The most efficient way to develop an item bank is to create and store items electronically. Several commercially produced software programs are available that facilitate item banking and test development. Educators can also organize test items electronically with a word-processing or spreadsheet program. The implementation of an item banking program is closely examined in Chapter 18, "Instituting Item Banking and Test Development Software."

Test

A test is a type of assessment that consists of a group of questions that is administered during a fixed time period to a group of students who participated in a class. 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 and Brookhart (2009) describe 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. A test measures how well a student performs either in comparison with a domain of content and objective or in comparison with others (Miller et al., 2012).

Using a single test or type of measurement instrument is not a satisfactory assessment strategy. Most course objectives require a variety of diverse measurement and evaluation strategies to determine student competency. 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 text 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 clinical judgment.

An achievement test should consist of a sampling of tasks that represents the larger domain of behavior included in the course. The number of questions on a test is limited, so the questions you include have to be a representative sample of all the possible questions you could ask. The sample must be relevant and represent the total domain of what was included in the course (Gronlund & Brookhart, 2009). When students complain that an exam did not relate to the course content, it may indicate 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. You are most likely to obtain a representative sample of test items by following a systematic procedure for developing a test blueprint. The challenge is to develop a blueprint for the test and write items to match the objectives and content being assessed. Chapter 4, "Implementing Systematic Test Development," provides guidelines for implementing a procedure for blueprint development.

Interpreting Test Scores

A raw test score is meaningless without a framework for interpretation. A raw score represents the number of correct responses on a test before any review or analysis of the items is done. 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 based on objectives or competencies, whereas norm-referenced tests (NRTs) define an individual's performance by comparing it with others (Furby, 2020). 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 (Miller et al., 2012).

Criterion-Referenced Tests (CRTs)

A criterion is a measurable behavior, attitude, or bit of knowledge, so CRTs assess a student's mastery of a criterion. A criterion reference approach interprets 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. A CRT score is listed as a percentage, with the number of correctly answered questions divided by the total number of questions (Furby, 2020). **Exhibit 2.1** presents an example of a criterion-referenced score.

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 (Brookhart & Nitko, 2019; Furby, 2020). If most students in a group meet the standard, the group scores will obviously cluster at the high end of the grading scale or be skewed to the right.

CRTs are often educator 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 & Brookhart, 2009).

Exhibit 2.1 Example of a Criterion-Referenced Score

The student demonstrated mastery by correctly identifying 90%, or 90/100, of the terms.

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.

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 the 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). Chapter 3, "Developing Instructional Objectives," offers a more in-depth discussion and also provides 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 clinical judgment. 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 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 attain full mastery of objectives at the developmental level. However, they are required to demonstrate the behaviors described 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 (1973) asserts that the use of CRTs is restricted to the assessment of developmental learning. Although test preparation should follow mastery-level procedures, he suggests that adequate assessment of student performance beyond minimal essentials requires tests at the developmental level to include items of varying difficulty and allow for both criterion- and norm-referenced interpretations. Robinson Kurpius and Stafford (2006) suggest that educators can designate multiple cutoff scores with CRTs. The syllabus would explain, for example, that a student who demonstrates mastery of 95% of the course content and objectives would receive a grade of A for the course. Students who achieve 85% would earn a B; 75%, a C; and below

75%, a D. In this case, 75% is the minimum for passing, and students are rewarded for achieving beyond the minimum.

Norm-Referenced Tests (NRTs)

Whereas CRTs measure a student's achievement of a program's objectives or competencies without reference to other students, the aim of an NRT 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 (Furby, 2020). 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 and do not indicate what a student has achieved; the tests indicate only how the student compares with other students in his or her group (Furby, 2020). An example of a norm-referenced score is shown in **Exhibit 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 (Brookhart & Nitko, 2019). 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 members an external reference point for comparing their curriculum with a composite national curriculum.

NRTs 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, clinical, or online use. Chapter 4, "Implementing Systematic Test Development," elaborates on the use of NRTs and CRTs when determining how difficult a test should be. **Table 2.2** compares CRTs and NRTs.

High-Stakes Test

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, a licensure examination is a high-stakes test because the examinees' scores on the test determine whether or not they will be allowed to practice their

Exhibit 2.2 Example of a Norm-Referenced Score

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

Table 2.2 Comparison of Criterion- and Norm-Referenced Tests

Criterion-Referenced Test	Norm-Referenced Test
• Compares student performance to preestablished criteria	• Compares student performance to reference group
• Describes the performance	• Rates the performance
• Mastery reference	• Relative performance reference
• Narrowly defined content domain	• 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 percentage-correct score	• Provides percentile rank

profession. 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.

Exams in nursing meet the criteria for being designated as high-stakes examinations. Brodersen and Lorenz (2020) examined the relationship between high-stakes tests and perceived stress. High levels of perceived stress were found in students, along with sympathetic activation. Life-altering decisions are certainly made based on the results of these exams. Classroom, clinical, and online 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 decisions that are made based on the results of exams can have a profound impact on students' lives, it is obvious that faculty must pay careful attention to developing exams that produce trustworthy results.

Grade

Whereas 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. Chapter 14, "Interpreting Test Results," and Chapter 17, "Assigning Grades," both discuss test analysis and grading procedures.

A cutoff score is the lowest grade a student can achieve to demonstrate proficiency in a course. Every course syllabus in a nursing program should spell out what cutoff score is required to pass the course. Suppose the pass score, or cutoff score, in a nursing program is 75%. The students would have to demonstrate an average of 75% across all the assessments in a course to pass. Every course syllabus should describe what scores correlate to each grade. If a passing grade of C requires an average of 75%, then an A might require a grade of 95%, a B an average of 85%, and a failing grade of D would be an average below 75%. The important issue is to make the grade requirements clear to the students.

Test Bias

A biased test is one that discriminates against a certain group based on socioeconomic status, disability, race, ethnicity, and/or gender (Slavin, 2018). 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 (Ahmad et al., 2018). *Stereotyping* refers to the representation of a group in a way that may be offensive to the group members. 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 influence their test scores (Ahmad et al., 2018). Diversity, equity, inclusion, and belonging are also key aspects of test language. *Test bias* in a nursing exam refers to the difference in a group's mean performance based on nonnursing 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 the purpose of the assessment. The aim of a nursing test is to measure knowledge that is essential to safe nursing practice after licensing examination. Reading speed, vocabulary ability, or familiarity with cultural practices that are unrelated to health should not influence a student's score (Miller et al., 2012). Therefore, it is important for educators 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 edited to remove stereotypical language. In fact, even the most innocent vocabulary can introduce bias into a test, as **Exhibit 2.3** illustrates. 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 (Miller et al., 2012).

Bosher (2002) defines linguistic bias as resulting from students' inability to understand an item because the language is so complex. Students who are English language learners (ELLs) 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. Each question should be written succinctly 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 nursing education, it can be a distraction in an exam. Students are not inclined to get the joke during an exam, particularly

Exhibit 2.3 Example of a Culturally Biased Stem

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 new language or for those who live in urban areas and depend on public transportation.*

ELL 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, "Writing Clinical Judgment Multiple-Choice Items," provides guidelines to assist you in eliminating bias from your test items.

Reliability

Test reliability is very important to test developers and test takers. You would have little confidence in a standardized nursing achievement test that ranked a student in the top 5% 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 group (Schrieber & Turk, 2023). 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 (Schrieber & Turk, 2023). 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, which is a measure of the amount of variation in test performance. Although 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 0 to 1, with 0 indicating no reliability and 1 indicating perfect reliability. Reliability is discussed at length in Chapter 13, "Establishing Evidence of Reliability and Validity."

Validity

Although a test must be reliable to be valid, a reliable test is not always valid. 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, 2020). Therefore, we can have reliable measures that provide the wrong information (**Exhibit 2.4**).

Frisbie (2005) notes that the term *validity* is one of the most misused and misunderstood concepts in educational measurement. It is important to the development

Exhibit 2.4 Reliability Requirement for Validity

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

and evaluation of a test. Validity is not a property of the test itself. It refers to the appropriateness of the interpretation and use 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 clinical judgment unless you can collect validity evidence to justify the test's use to measure clinical judgment or another nursing construct.

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 (Miller et al., 2012). *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 viewed as a unitary concept, not as three distinct types. 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* (American Educational Research Association [AERA] et al., 2014) refers to types of validity evidence rather than categories of validity. Validity is referred to as the most fundamental consideration when interpreting a test score. It is described as a process of collecting a variety of evidence to support a proposed interpretation of a test score. The 2014 edition 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., 2014, p. 11). The sources of validity evidence described in the 2014 *Standards* include the following:

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence related to 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.

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. 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 overrepresented. The validity of the inferences based on the test results depends on how well the test sample represents the domain being tested (Gronlund & Brookhart, 2009). 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 4, "Implementing Systematic Test Development," presents detailed guidelines for developing blueprints for your tests.

Content-related evidence of validity is essential during test development. 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.

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 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 (see **Exhibit 2.5**).

Exhibit 2.5 Steps for Enhancing Validity Evidence Based on Test Content

- State objectives in performance terms.
- Identify learning outcomes.
- Define the domain(s) 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.
- Review items for test bias.
- Provide adequate time for test completion.
- Review item and test analysis.
- Use the test only for its intended purpose.

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 clinical judgment. The 2014 *Standards* (AERA et al., 2014) 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 behaviors.

Evidence Based on Internal Structure

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 clinical judgment would require construct validation. Hence, a detailed definition of the construct of clinical judgment should be derived from prior evidence, theory, and research.

Evidence Based on Relation to Other Variables

This type of evidence examines the relationship of test scores to variables that are external to the test (AERA et al., 2014). 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., 2014, p. 17).

As Schreiber and Turk (2023) explain, 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. In predictive validity, however, 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). 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.

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 (Miller et al., 2012). Although 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 (Schrieber & Turk, 2023). 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 reliability of the test's results.

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 instrument. Chapter 13, "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 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 B, "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 this book. Each of these definitions is examined in greater detail in Chapter 14, "Interpreting Test Results," and Chapter 18, "Instituting Item Banking and Test Development Software."

Summary

Assessment procedures do not make decisions about students; educators 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 as proposed in this text. Many of these concepts are explained in greater detail in subsequent chapters. This text 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.

Learning Activities

1. Identify one instructional objective from a course you have taught or taken. Use Brookhart and Nitko's (2019) five guidelines to outline a plan for assessing student achievement of the learning outcomes associated with the objective.
2. Explain how a blueprint establishes validity evidence for the decisions made based on the results of a test.
3. Compare norm-referenced to CRT score interpretations. Explain why norm-referenced score interpretation is inappropriate in classroom and online settings.
4. Describe a situation that would result in test bias.
5. Compare reliability to validity when interpreting a test score.

Web Links

Association for the Assessment of Learning in Higher Education
http://www.aalhe.org/
Educational Resources Information Center
https://eric.ed.gov/

References

Ahmad, H., Mokshein, S. E., & Husin, M. R. (2018). Detecting item bias in an anatomy & physiology test for nursing students using item response theory. *International Journal of Academic Research in Progressive Education and Development*, *7*(1), 97–109.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bosher, S. (2002). Barriers to creating a more culturally diverse nursing profession: Linguistic bias in multiple-choice nursing exams. *Nursing Education Perspectives*, *24*, 25–34.

Brodersen, L., & Lorenz, R. (2020). Perceived stress, physiological stress reactivity, and exit exam performance in a prelicensure bachelor of science nursing program. *International Journal of Nursing Education Scholarship*, *17*(1).

Brookhart, S. M., & Nitko, A. J. (2018). *Educational assessment of students* (8th ed.). Pearson Education.

Brookhart, S. M., & Nitko, A. J. (2019). *Assessment and grading in classrooms.* Pearson Education.

Burns, N., & Grove, S. K. (2020). *The practice of nursing research: Appraisal, synthesis, and generation of evidence* (9th ed.). W. B. Saunders.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, *24*(3), 21–28.

Furby, Leanne. (2020). Implementing educational testing standards in nursing education. *Nursing Education Perspectives*, *41*, 70–71.

Glennon, C. D. (2006). Reconceptualizing program outcomes. *Journal of Nursing Education*, *45*, 55–58.

Gronlund, N. E. (1973). *Preparing criterion-referenced tests for classroom instruction*. Macmillan.

Gronlund, N. E., & Brookhart, S. M. (2009). *Gronlund's writing instructional objectives* (8th ed.). Pearson Education.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates.

Lyman, H. L. (1998). *Test scores and what they mean* (6th ed.). Allyn and Bacon.

Mager, R. F. (1997). *Preparing instructional objectives* (3rd ed.). Fearon.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2012). *Measurement and assessment in teaching* (11th ed.). Pearson.

Robinson Kurpius, S. E., & Stafford, M. E. (2006). *Testing and measurement: A user friendly guide*. Sage.

Schreiber J., & Turk M. T. (2023). *Statistics and data analysis literacy for nurses*. Springer.

Slavin, R. E. (2018). *Educational psychology: Theory and practice* (12th ed.). Pearson.

Yang, B. W., Razo, J., & Persky, A. M. (2019). Using testing as a learning tool. *American Journal of Pharmaceutical Education*, *83*(9), 7324.