SEVENTH EDITION

THE PRACTICE OF NURSING RESEARCH

Appraisal, Synthesis, and Generation of Evidence

SUSAN K. **GROVE**
NANCY **BURNS**
JENNIFER R. **GRAY**

ELSEVIER

http://evolve.elsevier.com

CHAPTER 16

Measurement Concepts

Measurement is the process of assigning numbers to objects, events, or situations in accord with some rule (Kaplan, 1963). The numbers assigned can indicate numerical values or categories for the objects being measured for research or practice. **Instrumentation**, a component of measurement, is the application of specific rules to develop a measurement device such as a scale or questionnaire. Quality instruments are essential for obtaining trustworthy data when measuring outcomes for research and practice (Doran, 2011; Melnyk & Fineout-Overhold, 2011; Waltz, Strickland, & Lenz, 2010).

The rules of measurement were developed so that the assigning of values or categories might be done consistently from one subject (or event) to another and eventually, if the measurement method is found to be meaningful, from one study to another. The rules of measurement established for research are similar to the rules of measurement implemented in nursing practice. For example, when nurses measure the urine output from patients, they use an accurate measurement device, observe the amount of urine in the device or container in a consistent way, and precisely record the urine output in the medical record. This practice promotes accuracy and precision and reduces the amount of error in measuring physiological variables such as urine output.

When measuring a subjective concept such as pain experienced by a child, researchers and nurses in practice need to use an instrument that captures the pain the child is experiencing. A commonly used scale to measure a child's pain is the Wong-Baker FACES Pain Rating Scale (Hockenberry & Wilson, 2009). By using this valid and reliable rating scale to measure the child's pain, any change in the measured value can be attributed to a change in the child's pain rather than measurement error. A copy of the Wong-Baker FACES Pain Rating Scale is provided in Chapter 17. Selecting accurate and precise physiological measurement methods and valid and reliable scales and questionnaires is essential in measuring study variables and outcomes in practice (Bannigan & Watson, 2009; Bialocerkowski, Klupp, & Bragge, 2010; DeVon, et al., 2007).

Researchers need to understand the logic within measurement theory so that they can select and use existing instruments or develop new quality measurement methods for their studies. Measurement theory, as with most theories, uses terms with meanings that can be best understood within the context of the theory. The following explanation of the logic of measurement theory includes definitions of directness of measurement, measurement error, levels of measurement, and reference of measurement. The reliability and validity of measurement methods, such as scales and questionnaires, are detailed. The accuracy, precision, and error of physiological measures are described. The chapter concludes with a discussion of sensitivity; specificity; and likelihood ratios examined to determine the quality of diagnostic tests and instruments used in healthcare research and practice.

Directness of Measurement

Measurement begins by clarifying the object, characteristic, or element to be measured. Only then can one identify or develop strategies or methods to measure it. In some cases, identification of the measurement object and measurement strategies can be objective, specific, and straightforward, as when we are measuring concrete factors, such as a person's weight or waist circumference; this is referred to as **direct measurement**. Healthcare technology has made direct measures of objective elements—such as height, weight, temperature, time, space, movement, heart rate, and respiration—familiar to us. Technology is also available to measure many biological and chemical characteristics, such as laboratory values, pulmonary

functions, and sleep patterns. Nurses are also experienced in gathering direct measures of demographic variables, such as age, gender, ethnicity, diagnosis, marital status, income, and education.

However, in nursing, the characteristic we want to measure often is an abstract idea or concept, such as pain, stress, depression, anxiety, caring, or coping. If the element to be measured is abstract, it is best clarified through a conceptual definition (see Chapter 8). The conceptual definition can be used to select or develop appropriate means of measuring the concept. The instrument or measurement strategy used in the study must match the conceptual definition. An abstract concept is not measured directly; instead, indicators or attributes of the concept are used to represent the abstraction. This is referred to as **indirect measurement**. For example, the complex concept of coping might be defined by the frequency or accuracy of identifying problems, the creativity in selecting solutions, and the speed or effectiveness in resolving the problem. A single measurement strategy rarely, if ever, can completely measure all aspects of an abstract concept. Multi-item scales have been developed to measure abstract concepts, such as the Spielberger State-Trait Anxiety Inventory developed to measure individuals' innate anxiety trait and their anxiety in a specific situation (Spielberger, Gorsuch, & Lushene, 1970).

Measurement Error

There is no perfect measure. Error is inherent in any measurement strategy. **Measurement error** is the difference between what exists in reality and what is measured by an instrument. Measurement error exists in both direct and indirect measures and can be random or systematic. Direct measures, which are considered to be highly accurate, are subject to error. For example, the weight scale may not be accurate, laboratory equipment may be precisely calibrated but may change with use, or the tape measure may not be placed in the same location or held at the same tension for each measurement.

There is also error in indirect measures. Efforts to measure concepts usually result in measuring only part of the concept or measures that identify an aspect of the concept but also contain other elements that are not part of the concept. Figure 16-1 shows a Venn diagram of the concept *A* measured by instrument *A-1*. In this figure, *A-1* does not measure all of concept *A*. In addition, some of what *A-1* measures is outside the concept of *A*. Both of these situations are examples of errors in measurement and are shaded in Figure 16-1.

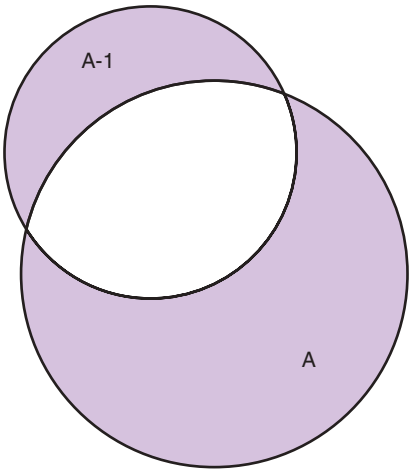

Figure 16-1 Measurement error when measuring a concept.

Types of Measurement Errors

Two types of errors are of concern in measurement: random error and systematic error. To understand these types of errors, we must first understand the elements of a score on an instrument or an observation. According to measurement theory, there are three components to a measurement score: true score, observed score, and error score. The **true score** (T) is what we would obtain if there was no error in measurement. Because there is always some measurement error, the true score is never known. The **observed score** (O) is the measure obtained for a subject using a selected instrument during a study. The **error score** (E) is the amount of random error in the measurement process. The theoretical equation of these three measures is as follows:

$$\text{Observed score} = \text{true score} + \text{random error}$$

This equation is a means of conceptualizing random error and not a basis for calculating it. Because the true score is never known, the random error is never known but only estimated. Theoretically, the smaller the error score, the more closely the observed score reflects the true score. Therefore, using instruments that reduce error improves the accuracy of measurement (Waltz et al., 2010).

Several factors can occur during the measurement process that can increase random error. These factors include (1) transient personal factors, such as fatigue, hunger, attention span, health, mood, mental status, and motivation; (2) situational factors, such as a hot stuffy room, distractions, the presence of significant

others, rapport with the researcher, and the playfulness or seriousness of the situation; (3) variations in the administration of the measurement procedure, such as interviews in which wording or sequence of questions is varied, questions are added or deleted, or researchers code responses differently; and (4) processing of data, such as errors in coding, accidentally marking the wrong column, punching the wrong key when entering data into the computer, or incorrectly totaling instrument scores (Devon et al., 2007; Waltz et al., 2010).

Random error causes individuals' observed scores to vary in no particular direction around their true score. For example, with random error, one subject's observed score may be higher than his or her true score, whereas another subject's observed score may be lower than his or her true score. According to measurement theory, the sum of random errors is expected to be zero, and the random error score (E) is not expected to correlate with the true score (T). Random error does not influence the mean to be higher or lower but rather increases the amount of unexplained variance around the mean. When this occurs, estimation of the true score is less precise.

If you were to measure a variable for three subjects and diagram the random error, it might appear as shown in Figure 16-2. The difference between the true score of subject 1 (T_1) and the observed score (O_1) is two positive measurement intervals. The difference between the true score (T_2) and observed score (O_2) for subject 2 is two negative measurement intervals. The difference between the true score (T_3) and observed score (O_3) for subject 3 is zero. The random error for these three subjects is zero (+2 − 2 + 0 = 0). In viewing this example, one must remember that this is only a means of conceptualizing random error.

Measurement error that is not random is referred to as **systematic error**. A scale that weighs subjects 3 pounds more than their true weights is an example of systematic error. All of the body weights would be higher, and, as a result, the mean would be higher than it should be. Systematic error occurs because something else is being measured in addition to the concept. A conceptualization of systematic error is presented in Figure 16-3. Systematic error (represented by the shaded area in the figure) is due to the part of *A-1* that is outside of *A*. This part of *A-1* measures factors other than *A* and biases scores in a particular direction.

Systematic error is considered part of T (true score) and reflects the true measure of *A-1,* not *A*. Adding the true score (with systematic error) to the random error

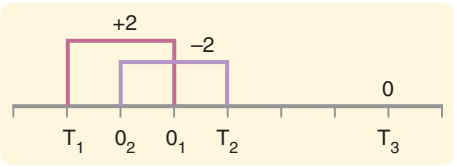

Figure 16-2 Conceptualization of random error.

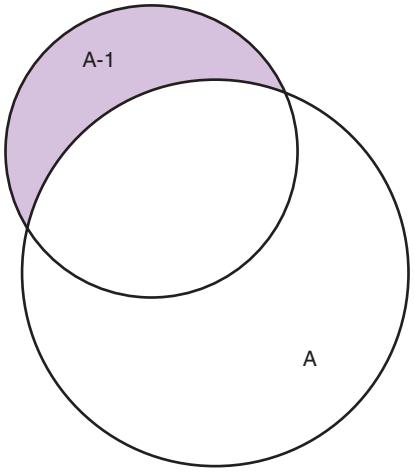

Figure 16-3 Conceptualization of systematic error.

(which is 0) yields the observed score, as shown by the following equations:

$$\text{T (true score with systematic error)} + \text{E (random error of 0)} = \text{O (observed score)}$$

or

$$T + E = O$$

Some systematic error is incurred in almost any measure; however, a close link between the abstract theoretical concept and the development of the instrument can greatly decrease systematic error. Because of the importance of this factor in a study, researchers spend considerable time and effort in selecting and developing quality measurement methods to decrease systematic error.

Another effective means of diminishing systematic error is to use more than one measure of an attribute or a concept and to compare the measures. To make this comparison, researchers use various data collection methods, such as scale, interview, and observation. Campbell and Fiske (1959) developed a technique

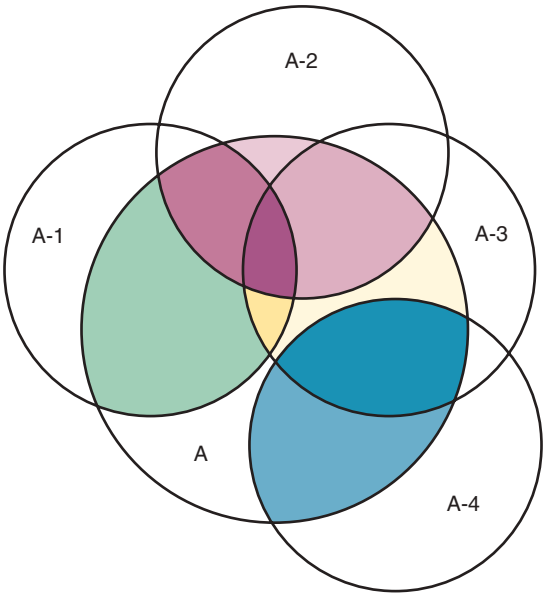

Figure 16-4 Multiple measures of an abstract concept.

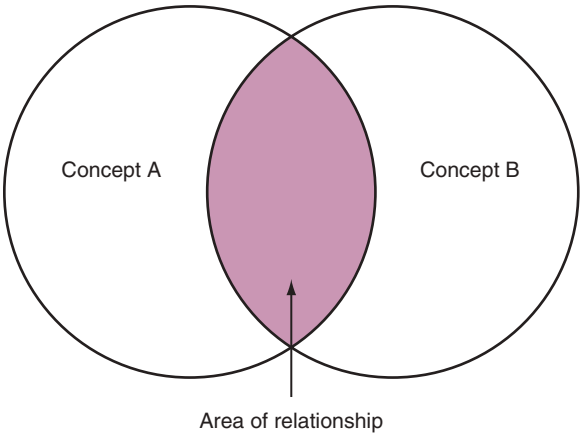

Figure 16-5 True relationship of concepts *A* and *B*.

of using more than one method to measure a concept, referred to as the **multimethod-multitrait technique**. More recently, the technique has been described as a version of mixed methodology, as discussed in Chapter 10. These techniques allow researchers to measure more dimensions of abstract concepts, and the effect of the systematic error on the composite observed score decreases. Figure 16-4 illustrates how more dimensions of concept *A* are measured through the use of four instruments, designated *A-1, A-2, A-3,* and *A-4*.

For example, a researcher could decrease systematic error in measures of anxiety by (1) administering the Spielberger State-Trait Anxiety Inventory, (2) recording blood pressure readings, (3) asking the subject about anxious feelings, and (4) observing the subject's behavior. Multimethod measurement strategies decrease systematic error by combining the values in some way to give a single observed score of anxiety for each subject. However, sometimes it may be difficult logically to justify combining scores from various measures, and a mixed-methods approach might be the most appropriate to use in the study. Mixed-methods study uses a combination of quantitative and qualitative approaches in their implementation (Creswell, 2009).

In some studies, researchers use instruments to examine relationships. Consider a hypothesis that tests the relationship between concept *A* and concept *B*. In Figure 16-5, the shaded area enclosed in the dark lines represents the true relationship between concepts *A* and *B,* such as the relationship between anxiety and depression. For example, two instruments, *A-1* (Spielberger State Anxiety Scale) and *B-1* (Center for Epidemiological Studies Depression Scale, Radloff, 1977), are used to examine the relationship between concepts *A* and *B*. The part of the true relationship actually reflected by *A-1* and *B-1* measurement methods is represented by the colored area in Figure 16-6. Because two instruments provide a more accurate measure of concepts *A* and *B,* more of the true relationship between concepts *A* and *B* can be measured.

If additional instruments (*A-2* and *B-2*) are used to measure concepts *A* and *B,* more of the true relationship might be reflected. Figure 16-7 demonstrates with different colors the parts of the true relationship between concepts *A* and *B* that is measured when concept *A* is measured with two instruments (*A-1* and *A-2*) and concept *B* is measured with two instruments (*B-1* and *B-2*).

Levels of Measurement

The traditional levels of measurement have been used for so long that the categorization system has been considered absolute and inviolate. In 1946, Stevens organized the rules for assigning numbers to objects so that a hierarchy in measurement was established called the **levels of measurement**. The levels of measurement, from lower to higher, are nominal, ordinal, interval, and ratio.

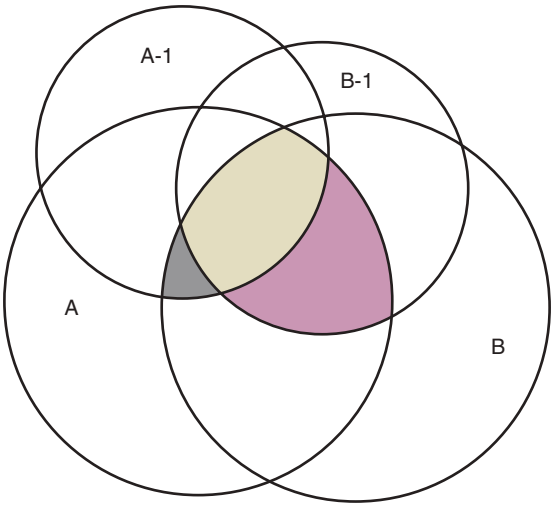

Figure 16-6 Examining a relationship using one measure of each concept.

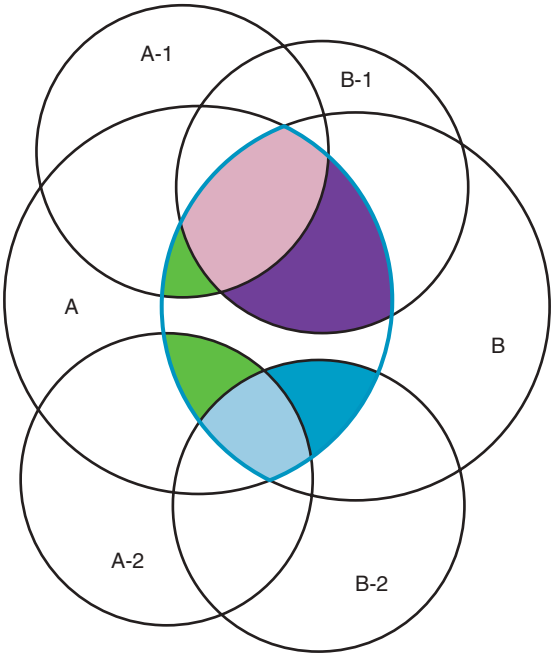

Figure 16-7 Examining a relationship using two measures of each concept.

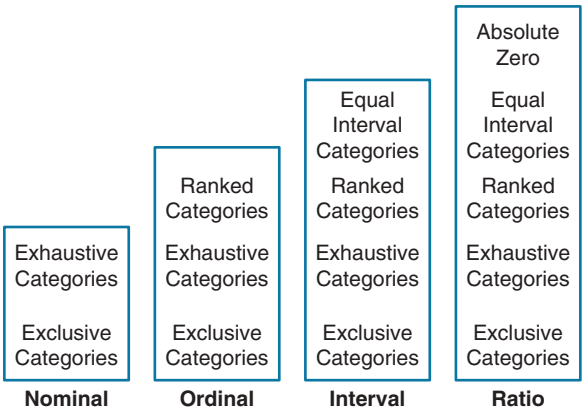

Figure 16-8 Summary of the rules for levels of measurement.

Nominal Level of Measurement

Nominal level of measurement is the lowest of the four measurement levels or categories. It is used when data can be organized into categories of a defined property but the categories cannot be ordered. For example, diagnoses of chronic diseases are nominal data with categories such as hypertension, type 2 diabetes, and dyslipidemia. One cannot say that one category is higher than another or that category A (hypertension) is closer to category B (diabetes) than to category C (dyslipidemia). The categories differ in quality but not quantity. One cannot say that subject A possesses more of the property being categorized than does subject B. **(*Rule:* The categories must be unorderable.)** Categories must be established so that each datum fits into only one of the categories. **(*Rule:* The categories must be exclusive.)** All the data must fit into the established categories. **(*Rule:* The categories must be exhaustive.)**

Figure 16-8 provides a summary for the rules for the four levels of measurement—nominal, ordinal, interval, and ratio. Data such as ethnicity, gender, marital status, religion, and diagnoses are examples of nominal data. When data are coded for entry into the computer, the categories are assigned numbers. For example, gender may be classified as 1 = male and 2 = female. The numbers assigned to categories in nominal measurement are used only as labels and cannot be used for mathematical calculations.

Ordinal Level of Measurement

Data that can be measured at the **ordinal level** can be assigned to categories of an attribute that can be ranked. There are rules for how one ranks data. As

with nominal-scale data, the categories must be exclusive and exhaustive. With ordinal level data, the quantity of the attribute possessed can be identified. However, it cannot be shown that the intervals between the ranked categories are equal (see Figure 16-8). Ordinal data are considered to have unequal intervals. Scales with unequal intervals are sometimes referred to as ordered metric scales.

Many scales used in nursing research are ordinal levels of measure. For example, one could rank intensity of pain, degrees of coping, levels of mobility, ability to provide self-care, or daily amount of exercise on an ordinal scale. For daily exercise, the scale could be 0 = no exercise; 1 = moderate exercise, no sweating; 2 = exercise to the point of sweating; 3 = strenuous exercise with sweating for at least 30 minutes per day; 4 = strenuous exercise with sweating for at least 1 hour per day. This type of scale may be referred to as a **metric ordinal scale**.

Interval Level of Measurement

In **interval level of measurement**, distances between intervals of the scale are numerically equal. Such measurements also follow the previously mentioned rules: mutually exclusive categories, exhaustive categories, and rank ordering. Interval scales are assumed to be a continuum of values (see Figure 16-8). The researcher can identify the magnitude of the attribute much more precisely. However, it is impossible to provide the absolute amount of the attribute because of the absence of a zero point on the interval scale.

Fahrenheit and Celsius temperatures are commonly used as examples of interval scales. A difference between a temperature of 70° F and one of 80° F is the same as the difference between a temperature of 30° F and one of 40° F. We can measure changes in temperature precisely. However, it is impossible to say that a temperature of 0° C or 0° F means the absence of temperature because these indicate very cold temperatures.

Ratio Level of Measurement

Ratio level of measurement is the highest form of measure and meets all the rules of the lower forms of measures: mutually exclusive categories, exhaustive categories, rank ordering, equal spacing between intervals, and a continuum of values. In addition, ratio level measures have absolute zero points (see Figure 16-8). Weight, length, and volume are common examples of ratio scales. Each has an **absolute zero point**, at which a value of zero indicates the absence of the property being measured: Zero weight means the absence of weight. In addition, because of the absolute zero point, one can justifiably say that object A weighs twice as much as object B, or that container A holds three times as much as container B. Laboratory values are also an example of ratio level of measurement where the individual with a fasting blood sugar (FBS) of 180 has an FBS twice that of an individual with a normal FBS of 90. To help expand understanding of levels of measurement (nominal, ordinal, interval, and ratio) and to apply this knowledge, Grove (2007) developed a statistical workbook focused on examining the levels of measurement, sampling methods, and statistical results in published studies.

Importance of Level of Measurement for Statistical Analyses

An important rule of measurement is that one should use the highest level of measurement possible. For example, you can collect data on age (measured) in a variety of ways: (1) you can obtain the actual age of each subject (ratio level of measurement); (2) you can ask subjects to indicate their age by selecting from a group of categories, such as 20 to 29, 30 to 39, and so on (ordinal level of measurement); or (3) you can sort subjects into two categories of younger than 65 years of age and 65 years of age and older (nominal level of measurement). The highest level of measurement in this case is the actual age of each subject, which is the preferred way to collect these data. If you need age categories for specific analyses in your research, the computer can be instructed to create age categories from the initial age data (Waltz et al., 2010).

The level of measurement is associated with the types of statistical analyses that can be performed on the data. Mathematical operations are limited in the lower levels of measurement. With nominal levels of measurement, only summary statistics, such as frequencies, percentages, and contingency correlation procedures, can be used. However, if a variable such as age is measured at the ratio level (actual age of the subject), the data can be analyzed with more sophisticated analysis techniques. Variables measured at the interval or ratio level can be analyzed with the strongest statistical techniques available, which are more effective in identifying relationships among variables or determining differences between groups (Corty, 2007; Grove, 2007).

Controversy over Measurement Levels

There is controversy over the system that is used to categorize measurement levels, dividing researchers into two factions: fundamentalists and pragmatists.

Pragmatists regard measurement as occurring on a continuum rather than by discrete categories, whereas *fundamentalists* adhere rigidly to the original system of categorization (Nunnally & Bernstein, 1994; Stevens, 1946).

The primary focus of the controversy relates to the practice of classifying data into the categories ordinal and interval. This controversy developed because, according to the fundamentalists, many of the current statistical analysis techniques can be used only with interval and ratio data. Many pragmatists believe that if researchers rigidly adhered to rules developed by Stevens (1946), few if any measures in the social sciences would meet the criteria to be considered interval-level data. They also believe that violating Stevens' criteria does not lead to serious consequences for the outcomes of data analysis. Pragmatists often treat ordinal data from multi-item scales as interval data, using statistical methods (parametric analysis techniques) to analyze them, such as Pearson's product-moment correlation coefficient, *t*-test, and analysis of variance (ANOVA), which are traditionally reserved for interval or ratio level data (Armstrong, 1981; Knapp, 1990). Fundamentalists insist that the analysis of ordinal data be limited to statistical procedures designed for ordinal data, such as nonparametric procedures. Parametric statistical analysis techniques were developed to analyze interval and ratio level data, and nonparametric techniques were developed to analyze nominal and ordinal data (see Chapter 21).

The Likert scale uses scale points such as "strongly disagree," "disagree," "uncertain," "agree," and "strongly agree." Numerical values (e.g., 1, 2, 3, 4, and 5) are assigned to these categories. Fundamentalists claim that equal intervals do not exist between these categories. It is impossible to prove that there is the same magnitude of feeling between "uncertain" and "agree" as there is between "agree" and "strongly agree." Therefore, they hold this is ordinal level data, and parametric analyses cannot be used. Pragmatists believe that with many measures taken at the ordinal level, such as scaling procedures, an underlying interval continuum is present that justifies the use of parametric statistics (Knapp, 1990; Nunnally & Bernstein, 1994).

Our position agrees more with the pragmatists than with the fundamentalists. Many nurse researchers analyze data from Likert scales and other rating scales as though the data were interval level (Waltz et al., 2010). However, some of the data in nursing research are obtained through the use of crude measurement methods that can be classified only into the lower levels of measurement (ordinal or nominal). Therefore, we have included the nonparametric statistical procedures needed for their analysis in Chapters 22 to 25 on statistics.

Reference Testing of Measurement

Referencing involves comparing a subject's score against a standard. Two types of testing involve referencing: norm-referenced testing and criterion-referenced testing. **Norm-referenced testing** addresses the question, "How does the average person score on this test or instrument?" This testing involves standardization of scores for an instrument that is accomplished by data collection over several years, with extensive reliability and validity information available on the instrument. Standardization involves collecting data from thousands of subjects expected to have a broad range of scores on the instrument. From these scores, population parameters such as the mean and standard deviation (described in Chapter 22) can be developed. Evidence of the reliability and validity of the instrument can also be evaluated through the use of the methods described later in this chapter. The best-known norm-referenced test is the Minnesota Multiphasic Personality Inventory (MMPI), which is used commonly in psychology and occasionally in nursing research and practice to diagnosis personality disorders. The Graduate Record Examination (GRE) is another norm-referenced test commonly used as one of the admission criteria for graduate study.

Criterion-referenced testing asks the question, "What is desirable in the perfect subject?" It involves comparing a subject's score with a criterion of achievement that includes the definition of target behaviors. When the subject has mastered these behaviors, he or she is considered proficient in the behavior (DeVon et al., 2007; Sax, 1997). The criterion might be a level of knowledge or desirable patient outcomes. Criterion measures have been used for years to evaluate outcomes in healthcare agencies and to determine clinical expertise of students. For example, a clinical evaluation form would include the critical behaviors the nurse practitioner (NP) student is expected to master in a pediatric course to be clinically competent to care for pediatric patients at the end of the course. Criterion-reference testing is also used in nursing research. Criterion-referenced testing might be used to measure the clinical expertise of a nurse or the self-care of a cardiac patient after cardiac rehabilitation.

Reliability

The **reliability** of an instrument denotes the consistency of the measures obtained of an attribute, item, or situation in a study or clinical practice. The greater the reliability or consistency of the measures of a particular instrument, the less random error in the measurement method (Bannigan & Watson, 2009; Bialocerkowski et al., 2010; DeVon et al., 2007). If the same measurement scale is administered to the same individuals at two different occasions, the measurement is reliable if the individuals' responses to the items remain the same (assuming that nothing has occurred to change their responses). For example, if you use a scale to measure the anxiety levels of 10 individuals at two points in time 30 minutes apart, you would expect the individuals' anxiety levels to be relatively unchanged from one measurement to the next if the scale is reliable. If two data collectors observe the same event and record their observations on a carefully designed data collection instrument, the measurement would be reliable if the recordings from the two data collectors are comparable. The equivalence of their results would indicate the reliability of the measurement technique. If responses vary each time a measure is performed, there is a chance that the instrument is unreliable, meaning that it yields data with a large random error.

Reliability plays an important role in the selection of measurement methods for use in a study. Researchers need instruments that are reliable and provide values with only a small amount of error. Reliable instruments enhance the power of a study to detect significant differences or relationships actually occurring in the population under study. It is important to examine the reliability of an instrument from previous research before using it in a study. Estimates of instrument reliability are specific to the population and sample being studied. High reported reliability values on an established instrument do not guarantee that its reliability would be satisfactory in another sample from a different population (Waltz et al., 2010). Researchers need to perform reliability testing on each instrument used in their study before performing other statistical analyses. The reliability values must be included in the published report of a study to document that the instruments used were reliable for the study sample (Bialocerkowski et al., 2010; DeVon et al., 2007).

Reliability testing examines the amount of measurement error in the instrument being used in a study. Reliability is concerned with the dependability, consistency, stability, precision, reproducibility, and comparability of a measurement method (Bartlett & Frost, 2008). The strongest measure of reliability is obtained from heterogeneous samples versus homogeneous samples. Heterogeneous samples have more between-participant variability, and this is a stronger evaluation of reliability than homogeneous samples with little between-participant variation. When critically appraising the reliability of an instrument in a study, you need to examine the sample for heterogeneity by determining the variability of the scores among study participants (Bartlett & Frost, 2008; Bialocerkowski et al., 2010).

All measurement techniques contain some random error, and the errors might be due to the measurement method used, the study participants, or the researchers gathering the data. Reliability exists in degrees and is usually expressed as a form of correlation coefficient, with 1.00 indicating perfect reliability and 0.00 indicating no reliability (Bialocerkowski et al., 2010). For example, reliability coefficients of 0.80 or higher are considered strong values for an established psychosocial scale such as the State-Trait Anxiety Inventory by Spielberger et al. (1970). With test-retest, the closer that reliability coefficient is to 1.00, the more stable the measurement method. The reliability coefficient varies based on the type of reliability being examined. The three most common types of reliability discussed in healthcare studies are (1) stability reliability, (2) equivalence reliability, and (3) internal consistency (Bannigan & Watson, 2009; Bialocerkowski et al., 2010; DeVon et al., 2007; Waltz et al., 2010).

Stability Reliability

Stability reliability is concerned with the consistency of repeated measures of the same attribute with the use of the same scale or instrument over time. It is usually referred to as **test-retest reliability**. This measure of reliability is generally used with physical measures, technological measures, and paper-and-pencil scales. The technique requires an assumption that the factor to be measured remains the same at the two testing times and that any change in the value or score is a consequence of random error.

The optimal time period between test-retest measurements depends on the variability of the variable being measured, complexity of the measurement process, and characteristics of the participants (Bialocerkowski et al., 2010). Physical measures and equipment can be tested and then immediately retested, or the equipment can be used for a time and then retested to determine the necessary frequency of recalibration. For example, in measuring blood pressure (BP), researchers often take two to three BP readings 5

minutes apart and average the readings to obtain a reliable or precise measure of BP. Researchers can follow the standards for recalibration of equipment or be more conservative. The standard requirements might be to recalibrate the BP equipment every 6 months, but researchers might choose to recalibrate the equipment every month or even every week if multiple BP readings are being taken each day for a study. The test-retest of a measurement method might have a longer period of time if the variable being measured changes slowly. For example, the diagnosis of osteoporosis is made by bone mineral density (BMD) study of the hip, spine, and wrist. The BMD score is determined with a dual-energy x-ray absorptiometry (DEXA) scan. Because the BMD does not change rapidly in people even with treatment, test-retest over a 1- to 2-month time period could be used to show reliable or consistent DEXA scan scores for patients.

With paper-and-pencil educational tests, a period of 2 to 4 weeks is recommended between the two testing times, but the time period for retesting does depend on what is measured and the instrument used (Sax, 1997). After the same participants have been retested with the same instrument, the investigators perform a correlational analysis on the scores from the two measurement times. This correlation is called the **coefficient of stability**, and the closer the coefficient is to 1.00, the more stable the instrument (Waltz et al., 2010). For some scales, test-retest reliability has not been as effective as originally anticipated. The procedure presents numerous problems. Subjects may remember their responses from the first testing time, leading to overestimation of the reliability. Subjects may be changed by the first testing and may respond to the second test differently, leading to underestimation of the reliability (Bialocerkowski et al., 2010).

Test-retest reliability requires the assumption that the factor being measured has not changed between the measurement points. Many of the phenomena studied in nursing, such as hope, coping, pain, and anxiety, do change over short intervals. Thus, the assumption that if the instrument is reliable, values will not change between the two measurement periods may not be justifiable. If the factor being measured does change, the test is not a measure of reliability. If the measures stay the same even though the factor being measured has changed, the instrument may lack reliability. If researchers are going to examine the reliability of an instrument with test-retest, they need to determine the optimum time between administrations of the instrument based on the variable being measured and the study participants (Devon et al., 2007).

Stability of a measurement method is important and needs to be examined as part of instrument development and discussed when the instrument is used in a study. When describing test-retest results, researchers need to discuss the process and the time period between administering an instrument and the rationale for this time frame. If a scale was administered twice 30 minutes apart or there was 1 month between test and retesting, the consistency of the subjects' scores need to be discussed in terms of the timing for retesting (Bannigan & Watson, 2009; Bialocerkowski et al., 2010; DeVon et al., 2007).

Equivalence Reliability

Equivalence reliability compares two versions of the same paper-and-pencil instrument or two observers measuring the same event. Comparison of two observers is referred to as **interrater reliability**. Comparison of two paper-and-pencil instruments is referred to as **alternate-forms reliability** or **parallel-forms reliability**. Alternative forms of instruments are of more concern in the development of normative knowledge testing. However, when repeated measures are part of the design, alternative forms of measurement, although not commonly used, would improve the design. Demonstrating that one is actually testing the same content in both tests is extremely complex, and the procedure is rarely used in clinical research (Bialocerkowski et al., 2010).

The procedure for developing parallel forms involves using the same objectives and procedures to develop two like instruments. These two instruments when completed by the same group of study participants on the same occasion or two different occasions should have approximately equal means and standard deviations. In addition, these two instruments should correlate equally with another variable. For example, if two instruments were developed to measure pain, the scores from these two scales should correlate equally with perceived anxiety score. If both forms of the instrument are administered during the same occasion, a reliability coefficient can be calculated to determine equivalence. A coefficient of 0.80 or higher indicates equivalence (Waltz et al., 2010).

Determining interrater reliability is a concern when studies include observational measurement, which is common in qualitative research. Interrater reliability values need to be reported in any study in which observational data are collected or judgments are made by two or more data gatherers. Two techniques determine interrater reliability. Both techniques require that two or more raters independently observe and record the same event using the protocol

developed for the study or that the same rater observes and records an event on two occasions. To judge interrater reliability adequately, the raters need to observe at least 10 subjects or events (DeVon et al., 2007; Waltz et al., 2010). A digital recorder can be used to record the raters to determine their consistency in recording essential study information. Every data collector used in the study must be tested for interrater reliability and trained to a consistency in data collection.

One procedure for calculating interrater reliability requires a simple computation involving a comparison of the agreements obtained between raters on the coding form with the number of possible agreements. This calculation is performed through the use of the following equation:

Number of agreements ÷ number of possible agreements = interrater reliability

This formula tends to overestimate reliability, a particularly serious problem if the rating requires only a dichotomous judgment, such as present or absent. In this case, there is a 50% probability that the raters will agree on a particular item through chance alone. If more than two raters are involved, a statistical procedure to calculate coefficient alpha (discussed later in this chapter) may be used. ANOVA may also be used to test for differences among raters. There is no absolute value below which interrater reliability is unacceptable. However, any value less than 0.80 (80%) should generate serious concern about the reliability of the data because there is 20% chance of error. The more ideal interrater reliability value is 0.90, which means 90% reliability and 10% error. The process for determining interrater reliability and the value achieved need to be included in the research report (DeVon et al., 2007).

When raters know they are being watched, their accuracy and consistency are considerably better than when they believe they are not being watched. Interrater reliability declines (sometimes dramatically) when the raters are assessed covertly (Topf, 1988). You can develop strategies to monitor and reduce the decline in interrater reliability, but they may entail considerable time and expense.

Internal Consistency

Tests of instrument **internal consistency** or **homogeneity**, used primarily with paper-and-pencil tests or scales, address the correlation of various items within the instrument. The original approach to determining internal consistency was **split-half reliability**. This strategy was a way of obtaining test-retest reliability without administering the test twice. The instrument items were split in odd-even or first-last halves, and a correlational procedure was performed between the two halves. In the past, researchers generally reported the Spearman-Brown correlation coefficient in their studies (Nunnally & Bernstein, 1994; Sax, 1997). One of the problems with the procedure was that although items were usually split into odd-even items, it was possible to split them in a variety of ways. Each approach to splitting the items would yield a different reliability coefficient. The researcher could continue to split the items in various ways until a satisfactorily high coefficient was obtained.

More recently, testing the internal consistency of all the items in the instrument has been seen as a better approach to determining reliability. Although the mathematics of the procedure are complex, the logic is simple. One way to view it is as though one conducted split-half reliabilities in all the ways possible and then averaged the scores to obtain one reliability score. Internal consistency testing examines the extent to which all the items in the instrument consistently measure a concept. **Cronbach's alpha coefficient** is the statistical procedure used for calculating internal consistency for interval and ratio level data. This reliability coefficient is essentially the mean of the interitem correlations and can be calculated using most data analysis programs such as the Statistical Program for the Social Sciences (SPSS). If the data are dichotomous, such as a symptom list that has responses of present or absent, the Kuder-Richardson formulas (*KR 20* or *KR 21*) can be used to calculate the internal consistency of the instrument (DeVon et al., 2007). The *KR 21* assumes that all the items on a scale or test are equally difficult; the *KR 20* is not based on this assumption. Waltz et al. (2010) provided the formulas for calculating both *KR 20* and *KR 21*.

Cronbach's alpha coefficients can range from 0.00, indicating no internal consistency or reliability, to 1.00, indicating perfect internal reliability with no measurement error. Alpha coefficients of 1.00 are not obtained in study results because all instruments have some measurement error. However, many respected psychosocial scales used for 15 to 30 years to measure study variables in a variety of populations have strong 0.8 or greater internal reliability coefficients. The coefficient of 0.80 (or 80%) indicates the instrument is 80% reliable with 20% random error (DeVon et al., 2007; Fawcett & Garity, 2009; Grove, 2007). Scales with 20 or more items usually have stronger internal consistency coefficients than scales with 10 to 15 items or less. Often scales that measure complex constructs such as quality of life (QOL) have subscales

that measure different aspects of QOL, such as health, physical functioning, and spirituality. Some of these complex scales with distinct subscales, such as the QOL scale, might have lower Cronbach's alpha coefficients since the scale is measuring different aspects of QOL. The subscales have fewer items than the total scale and usually lower Cronbach's alpha coefficients but do need to show internal consistency in measuring a concept (Bialocerkowski et al., 2010; Waltz et al., 2010).

Newer instruments, developed in the last 5 years, might show moderate internal reliability (0.70 to 0.79) when used in measuring variables in a variety of samples. The subscales of these new instruments might have internal reliability from 0.60 to 0.69. The authors of these scales might continue to refine them based on additional reliability and validity information to improve the reliability of the total scale and subscales. Reliability coefficients less than 0.60 are considered low and indicate limited instrument reliability or consistency in measurement with high random error. Higher levels of reliability or precision (0.90 to 0.99) are important for physiological measures that are used to determine critical physiological functions such as arterial pressure and oxygen saturation (Bialocerkowski et al., 2010; DeVon et al., 2007).

The quality of the instrument reliability needs to be examined in terms of the type of study, measurement method, and population (DeVon et al., 2007; Kerlinger & Lee, 2000). In published studies, researchers need to identify the reliability coefficients of an instrument from previous research and for their particular study. Because the reliability of an instrument can vary from one population or sample to another, it is important that the reliability of the scale and subscales be determined and reported for the sample in each study (Bialocerkowski, et al., 2010).

Dickerson, Kennedy, Wu, Underhill, and Othman (2010) conducted a study of QOL and anxiety levels of patients with implantable defibrillators. They provided the following discussion of the reliability of the scales they used in their study.

> "***Anxiety.*** Anxiety was measured by the Spielberger State-Trait Anxiety Inventory (STAI), which determines a subject's current state of anxiety. This instrument differentiates between the temporary condition of 'state anxiety' and the longstanding quality of 'trait anxiety.' The STAI is a 40-item instrument that gages emotional reactions to the environment (e.g., 'I am tense,' 'I feel upset,' or 'I am worried'). Subjects rate themselves on a 4-point scale of 1 (not at all) to 4 (very much so). The median α [alpha] reliability of the inventory was reported to be 0.93 [from previous studies].... Cronbach's α scores for the present study ranged from 0.90 to 0.96 for the 3 time periods. Only the state anxiety score was used in this analysis.
>
> "***Quality-of-life measure.*** Quality-of-life was measured using the Ferrans and Powers Quality of Life Index, Cardiac Version (QLI: CV) (Bliley & Ferrans, 1993). The QLI: CV measures the subject's perception of QOL, according to a 72-item scale consisting of 2 parts. The first part measures satisfaction with various aspects of life and the second part measures the importance of these aspects to a subject. In part 1, subjects respond to a 6-point scale, ranging from 'very important' (6 points) to 'very unimportant' (1 point). Scores are calculated by weighing the satisfaction responses with the importance responses. They reflect how satisfied subjects are with the aspects of life that are important to them. The 4 subscales of QOL scored include health and functioning, social and economic, psychological/spiritual, and family. The reliability of the total scale's internal consistency was supported by α coefficients ranging from 0.90 to 0.95. Stability and reliability were supported by a test-retest correlation of 0.87 at a 2-week interval, and 0.81 at a 1-month interval.... Cronbach's α scores for the present study ranged from 0.95 to 0.96 for total QLI: CV, and for the subscales, Cronbach's α scores ranged from 0.88 to 0.94." (Dickerson et al., 2010, p. 468)

Dickerson et al. (2010) used two very reliable scales to measure their study variables and documented this in their article. They measured anxiety with the Spielberger STAI, which was developed more than 40 years ago (Spielberger et al., 1970), has shown strong internal consistency in previous research (median alpha = 0.93), and was reliable in this study (Cronbach's alpha = 0.90 to 0.96). In previous studies, the QLI: CV had strong internal consistency for the total scale (alpha = 0.90-0.95) and stability reliability with test-retest over 2 weeks and 1 month. In addition to the strong stability reliability coefficients, the researchers also provided the time frames for the test-retests that were run on the scale. Another strength is that the QLI: CV showed strong internal consistency for the total scale (alpha = 0.95 to 0.96) and the four subscales (alpha = 0.88 to 0.94) with the population in this study.

Other approaches to testing internal consistency are (1) Cohen's kappa statistic, which determines the

percentage of agreement with the probability of chance being taken out; (2) correlating each item with the total score for the instrument; and (3) correlating each item with each other item in the instrument. This procedure, often used in instrument development, allows researchers to identify items that are not highly correlated and delete them from the instrument. Factor analysis may also be used to develop instrument reliability. The number of factors being measured influences the reliability of the instrument, and total instrument scores may be more reliable than the scores of the subscales. After performing the factor analysis, the researcher can delete instrument items with low factor weights. After these items have been deleted, reliability scores on the instrument are higher. For instruments with more than one factor, correlations can be performed between items and factor scores (see Chapter 23 for a discussion of factor analysis).

It is essential that an instrument be both reliable and valid for measuring a study variable in a population. If the instrument has low reliability values, it cannot be valid because its measurement is inconsistent and has high measurement error (DeVon et al., 2007; Waltz et al., 2010). An instrument that is reliable cannot be assumed to be valid for a particular study or population. You need to determine the validity of the instrument you are using for your study, which you can accomplish in a variety of ways.

Validity

The **validity** of an instrument determines the extent to which it actually reflects or is able to measure the construct being examined. Several types of validity are discussed in the literature, such as content validity, predictive validity, criterion validity, and construct validity. Within each of these types, subtypes have been identified. These multiple types of validity are very confusing, especially because the types are not discrete but are interrelated (Bannigan & Watson, 2009; DeVon et al., 2007; Fawcett & Garity, 2009).

In this text, validity is considered a single broad measurement evaluation that is referred to as **construct validity** and includes various types, such as content validity, validity from factor analysis, convergent and divergent validity, validity from contrasting groups, and validity from prediction of future and current events (DeVon et al., 2007). All of the previously identified types of validity are now considered evidence of construct validity. In 1999, in its *Standards for Educational and Psychological Testing,* the American Psychological Association's Committee to Develop Standards published standards used to judge the evidence of validity. This important work greatly extends our understanding of what validity is and how to achieve it. According to the American Psychological Association's Committee to Develop Standards (1999), validity addresses the appropriateness, meaningfulness, and usefulness of the specific inferences made from instrument scores. It is the inferences made from the scores, not the scores themselves, that are important to validate (Devon et al., 2007; Goodwin & Goodwin, 1991).

Validity, similar to reliability, is not an all-or-nothing phenomenon but rather a matter of degree. No instrument is completely valid. One determines the degree of validity of a measure rather than whether or not it has validity. Determining the validity of an instrument often requires years of work. Many authors equate the validity of the instrument with the rigorousness of the researcher. The assumption is that because the researcher develops the instrument, the researcher also establishes the validity. However, this is an erroneous assumption because validity is not a commodity that researchers can purchase with techniques. Validity is an ideal state—to be pursued, but not to be attained. As the roots of the word imply, *validity* includes truth, strength, and value. Some authors might believe that validity is a tangible "resource," which can be acquired by applying enough appropriate techniques. However, we reject this view and believe measurement validity is similar to integrity, character, or quality, to be assessed relative to purposes and circumstances and built over time by researchers conducting a variety of studies (Brinberg & McGrath, 1985).

Figure 16-9 illustrates validity (the shaded area) by the extent to which the instrument *A-1* reflects concept

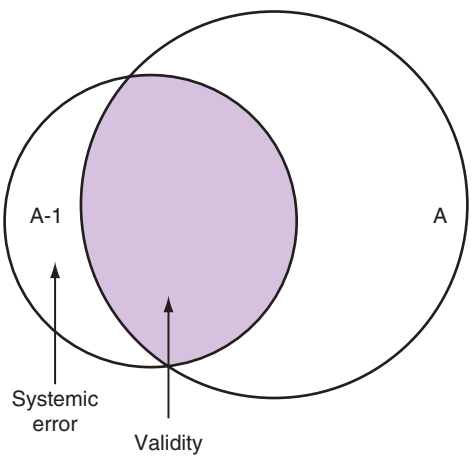

Figure 16-9 Representation of instrument validity.

A. As measurement of the concept improves, validity improves. The extent to which the instrument *A-1* measures items other than the concept is referred to as systematic error (identified as the unshaded area of *A-1* in Figure 16-9). As systematic error decreases, validity increases.

Validity varies from one sample to another and from one situation to another; therefore, validity testing affirms the appropriateness of an instrument for a specific group or purpose rather than the instrument itself (DeVon et al., 2007; Waltz et al., 2010). An instrument may be valid in one situation but not valid in another. Instruments used in nursing studies that were developed for use in other disciplines need to be examined for validity in terms of nursing knowledge. An instrument developed to measure cognitive function in educational studies might not capture the cognitive function level of elderly adults measured in a nursing study. Researchers are encouraged to reexamine validity in each of their study situations. The types of validity covered in this section include face and content validity, readability of an instrument, validity from factor analysis, validity from structural analysis, validity from contrasting (or known) groups, convergent and divergent validity, validity from discriminant analysis, validity from prediction of future and concurrent events, and successive verification validity.

Face and Content Validity

In the 1960s and 1970s, the only type of validity that most studies addressed was referred to as **face validity**, which verified basically that the instrument looked like it was valid or gave the appearance of measuring the construct it was supposed to measure. Face validity is a subjective assessment that might be made by the researchers or potential subjects. Because this is a subjective judgment with no clear guidelines for making the judgment, this is considered the weakest form of validity (DeVon et al., 2007). However, it is still an important aspect of the usefulness of the instrument because the willingness of subjects to complete the instrument relates to their perception that the instrument measures the construct they agreed to provide (Thomas, 1992). Face validity is often considered a step before or an aspect of content validity.

Content validity examines the extent to which the measurement method includes all the major elements relevant to the construct being measured. This evidence is obtained from the following three sources: the literature, representatives of the relevant populations, and content experts (DeVon et al., 2007; Fawcett & Garity, 2009; Waltz et al., 2010).

Documentation of content validity begins with development of the instrument. The first step of instrument development is to identify *what* is to be measured; this is referred to as the *universe* or *domain* of the construct. You can determine your domain through a concept analysis or an extensive literature search. Qualitative methods can also be used for this purpose. Johnson and Rogers (2006) developed the Medication-Taking Questionnaire (MTQ) based on purposeful action dimensions to determine the decision-making process of individuals for adherence to medication treatment for hypertension. They described their initial instrument development process as follows.

> "A total of 20 items (need, $n = 8$; effectiveness, $n = 6$; and safety, $n = 6$) were initially developed to tap the three underlying dimensions of purposeful action based on the statements given by participants in a qualitative study (Johnson, 2002; Johnson, Williams, & Marshall, 1999). The method for item construction was guided by the principles outlined in DeVellis (1991) and Streiner and Norman (1995).... The MTQ: Purposeful Action items were arranged in a 7-point, Likert-type format describing responses based on agreement (7 = *always agree*, 6 = *very frequently agree*, 5 = *usually agree*, and 4 = *occasionally agree*, 3 = *rarely agree*, 2 = *almost never agree*, 1 = *never agree*). The 7-response option was used in an attempt to obtain optimal variance while discouraging a ceiling effect (Steiner & Norman, 1995). Higher scores for the MTQ: Purposeful Action indicated greater intent to take medications based on perceived need, effectiveness, and safety." (Johnson & Rogers, 2006, p. 339)

Researchers need to describe the procedures used to develop or select items for the instrument that represent the domain of the construct. One helpful strategy commonly used is to develop a blueprint or matrix, such as was used in developing test items for an examination that was done by Johnson (2002) in her dissertation focused on development of the MTQ. However, before developing such items, the blueprint specifications must be submitted to an expert panel to validate that they are appropriate, accurate, and representative. At least five experts are recommended, although a minimum of three experts is acceptable if you cannot locate additional individuals with expertise in the area. Researchers might seek out individuals with expertise in various fields—for example, one

individual with knowledge of instrument development, a second with clinical expertise in an appropriate field of practice, and a third with expertise in another discipline relevant to the content area.

The experts need specific guidelines for judging the appropriateness, accuracy, and representativeness of the specifications. Berk (1990) recommended that the experts first make independent assessments and then meet for a group discussion of the specifications. The instrument specifications then can be revised and resubmitted to the experts for a final independent assessment. Davis (1992) recommended that the researcher provide expert reviewers with theoretical definitions of concepts and a list of which instrument items are expected to measure each of the concepts. The researcher asks the reviewers to judge how well each of the concepts has been represented in the instrument.

Researchers need to determine how to measure the domain. The item format, item content, and procedures for generating items must be carefully described. Items are then constructed for each cell in the matrix, or observational methods are designated to gather data related to a specific cell. Researchers are expected to describe the specifications used in constructing items or selecting observations. Sources of content for items must be documented. Then researchers can assemble, refine, and arrange the items in a suitable order before submitting them to the content experts for evaluation. Specific instructions for evaluating each item and the total instrument must be given to the experts.

In developing content validity for an instrument, researchers can calculate a **content validity ratio** (CVR) for each item on a scale by rating it 0 (not necessary), 1 (useful), or 3 (essential). A method for calculating the CVR was developed by Lawshe (1975) and is presented in Table 16-1 (DeVon et al., 2007). Minimum CVR scores for including items in an instrument can be based on a one-tailed test with a 0.05 level of significance.

The content validity score calculated for the complete instrument is called the **content validity index** (CVI). The CVI was developed to obtain a numerical value that reflects the level of content-related validity evidence for a measurement method (Waltz & Bausell, 1981). In calculating CVI, experts rate the content relevance of each item in an instrument using a 4-point rating scale. Lynn (1986, p. 384) recommended standardizing the options on this scale to read as follows: "1 = not relevant; 2 = unable to assess relevance without item revision or item is in need of such revision that it would no longer be relevant; 3 = relevant but needs minor alteration; 4 = very relevant and succinct." In addition to evaluating existing items, the experts were asked to identify important areas not included in the instrument. The calculation for the CVI is presented in Table 16-1 using the format developed by Lynn (1986). Complete agreement needs to exist among the expert reviewers to retain an item with seven or fewer reviewers. If few reviewers are used and many of the experts support most of the items on an instrument, this often results in an inflated CVI and an inflation in the content validity of the instrument (DeVon et al., 2007).

As presented earlier, Johnson and Rogers (2006) developed the MTQ: Purposeful Action and described their content validity testing process and outcomes as follows.

> "Content validity testing was undertaken to determine clarity and relevance of content. Participants and experts were given verbal instructions and a packet consisting of a consent form, written instructions, clarity instrument, content validity instrument, and demographic questionnaire. The clarity instrument asked participants to rate items as clear or unclear (Imle & Atwood, 1988). Participants were given a definition of each subscale and asked to rate each item's relevancy using a 4-point scale from 1 (irrelevant) to 4 (extremely relevant; Lynn, 1986). Space was provided to make comments after each rating procedure. (p. 339)
>
> Items met clarity criterion if 70% of participants rated the item as clear and the content validity criterion if 80% of participants rated the item as 3 or 4 (Imle & Atwood, 1988; Lynn, 1986). The comments from the clarity and content validity criterion were used to revise the MTQ: Purposeful Action items and subscales.…
>
> Of the 20 MTQ: Purpose Action items, 19 achieved clarity and content validity agreement. The 1 item that had an unacceptable clarity agreement was eventually eliminated from the questionnaire. Professionals expressed a concern about the lack of specificity in the questions, but that was not an issue for the hypertensive participants. For example, one professional indicated that the item, 'Blood pressure pills keep me from having problems,' lacked specificity. Because the purpose of this questionnaire was to establish a general screening tool for individuals who potentially may choose not to take their medications rather than to create a diagnostic tool, the participants' scores were given priority. Of the 20 items [see

TABLE 16-1 Two Methods of Calculating the Content Validity Ratio (CVR) and the Content Validity Index (CVI)

	Lawshe (1975)	Lynn (1986)
Rating scale	Scale used for rating items 0 1 3 Not necessary Useful Essential	Scale used for rating items 1 2 3 4 Irrelevant Extremely Relevant
Calculations	To calculate CVR (a score for individual scale items) $CVR = (n_e - N/2)/(N/2)$ *Note:* n_e = The number of experts who rated an item as "essential" N = the total number of experts. *Example:* If 8 of 10 experts rated an item as essential, CVR would be (8 − 5/5) = 0.60	CVI for each scale item is the proportion of experts who rate the item as a 3 or 4 on a 4-point scale. *Example:* If 4 of 6 content experts rated an item as relevant (3 or 4), CVI would be: 4/6 = 0.67 This item would not meet the 0.83 level of endorsement required to establish content validity using a panel of 6 experts at the 0.05 level of significance. Therefore, it would be dropped CVI for the entire scale is the proportion of the total number of items deemed content valid. *Example:* If 77 of 80 items were deemed content valid, CVI would be: 77/80 = 0.96
Acceptable range	Depends on number of reviewers	Depends on number of reviewers

From DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*(2), 158.

> Table 16-2 for the 20 items in the original questionnaire], 12 underwent minor grammatical revisions guided by the comments of both the participants and professionals. For example, items were made specific to blood pressure and the term medication was changed to pills. Several items were reworded, or the tense of the verb was changed." (Johnson & Rogers, 2006, pp. 341-342)

Before sending the instrument to experts for evaluation, researchers need to decide how many experts must agree on each item and on the total instrument for the content to be considered valid. Items that do not achieve minimum agreement by the expert panel must be either eliminated from the instrument or revised (DeVon et al., 2007; Lynn, 1986). Johnson and Rogers (2006) described their panel of reviewers, who were health professionals and patients prescribed antihypertensive medications, for the MTQ in the following excerpt.

> "Content validity testing was conducted in a sample of five hypertensive patients and five health care professionals who examined the MTQ for clarity and content relevance (Imle & Atwood, 1988; Lynn, 1986). Professionals were invited to participate in the study based on their known experience with antihypertensive treatment and included two family physicians, a cardiology nurse practitioner, a nurse working with a statewide cardiovascular disease program, and a nurse researcher who had published articles on adherence. All professionals were Anglo American and were nearly equally divided with regard to gender.
>
> Participants for the content validity phase who had been prescribed antihypertensive medications and lived in a situation in which they managed their own medications were recruited through healthy aging clinics, worksite wellness programs, hospital outpatient clinics, and hospital emergency departments in the intermountain west. The five hypertensive participants were Anglo American, had at least a high school education and ranged in age from 48 to 90 years (M = 62.0 ± 16.4)." (Johnson & Rogers, 2006, p. 338)

Johnson and Rogers (2006) provided excellent detail about the development of their instrument and

TABLE 16-2 Medication-Taking Questionnaire: Purposeful Action—Initial 20 Items Statistics

	M	SD	Item-Total Correlation	Mann-Whitney Adherence *p* Values*
Perceived Need				
My blood pressure pills keep me from having a stroke.	5.8	1.5	0.58	0.08
I need to take my blood pressure pills.	6.4	1.4	0.77	0.01
I take my blood pressure pills for my health.	6.5	1.3	0.75	0.01
Blood pressure pills keep me from having health-related problems.	5.7	1.5	0.63	0.17
I could have health problems if I do not take my blood pressure pills.	6.1	1.3	0.74	0.13
It's not a problem if I miss my blood pressure pills.†	5.1	2.0	0.30	0.02
I would rather treat my blood pressure without pills.†	4.1	2.3	0.37	0.26
I am OK if I do not take my blood pressure pills.†	5.6	1.8	0.64	0.012
Perceived Effectiveness				
My blood pressure will come down enough without pills.†	5.4	1.8	0.40	0.10
I will have problems if I don't take my blood pressure pills.	6.1	1.4	0.63	0.001
My blood pressure pills control my blood pressure.	6.0	1.4	0.66	0.46
Blood pressure pills benefit my health.	6.1	1.4	0.74	0.01
I feel better when I take my blood pressure pills.	5.4	1.8	0.56	0.01
I have problems finding pills that will control my blood pressure.†	5.7	1.8	0.09	0.059
Perceived as Safe				
The side effects from my blood pressure pills are a problem.†	5.2	1.9	0.40	0.10
The side effects from my blood pressure pills are harmful.†	5.6	1.8	0.63	0.27
My blood pressure pills are safe.	5.8	1.4	0.66	0.47
Taking my blood pressure pills is not a problem because they benefit my health.	6.0	1.4	0.74	0.02
My blood pressure pills cause other health problems.†	5.4	1.8	0.56	0.35
I will become dependent on my blood pressure pills.†	3.9	2.3	–0.5	0.20

From Johnson, M. J., & Rogers, S. (2006). Development of the Purposeful Action Medication-Taking Questionnaire. *Western Journal of Nursing Research, 28*(3), 344.

*Difference between low (scored 1-3) versus high (scored 7-10) adherence.

†Reverse coded.

M, Mean; *SD,* standard deviation.

the process for determining content validity. They also provided extensive information about the expert review panel for conducting the content validity testing. The strength of the review panel is that it included both health professionals and patients taking medications for hypertension. However, since all the reviewers were Anglo American, there was no ethnic diversity in the review process. The MTQ was a Likert scale with 7-point response options (described earlier), so it would be clearer if the researchers had called the MTQ a Likert scale versus a questionnaire (Waltz et al., 2010).

With some modifications, the content validity procedure previously described can be used with existing instruments, many of which have never been evaluated for content-related validity. With the permission of the author or researcher who developed the instrument, you could revise the instrument to improve its content-related validity (Lynn, 1986). In addition, the panel of experts or reviewers evaluating the items of the instrument for content validity might also examine it for readability and language acceptability to possible subjects or data gatherers (Berk, 1990; DeVon et al., 2007).

Readability of an Instrument

Readability is an essential element of the validity and reliability of an instrument. Assessing the level of readability of an instrument is simple and takes seconds with the use of a computer. There are more

than 30 readability formulas. These formulas count language elements in the document and use this information to estimate the degree of difficulty a reader may have in comprehending the text. Readability formulas are now a standard part of word-processing software. Box 16-1 provides instructions for using the Fog formula to determine the readability of a measurement method.

Although readability has never been formally identified as a component of content validity, it is essential that the items of an instrument be comprehended by subjects. Miller and Bodie (1994) suggested that the researcher should directly assess the reading comprehension level of the study population before using a formula to calculate an instrument's readability. They indicated that it is a mistake to assume that someone's literacy is equivalent to the last grade level the individual completed. Miller and Bodie (1994) recommended that researchers use the Classroom Reading Inventory (CRI), which is based on the Flesch, Space, Dale, and Fry reading comprehension scales (Flesch 1984; Silvaroli, 1986). This instrument determines the level at which an individual can comprehend written material without assistance. Johnson and Rogers (2006) described the readability of their MTQ: Purposeful Action as follows.

"Items were worded at approximately a sixth-grade reading level, evaluated by using the Flesch-Kincaid grade-level assessment program in Microsoft Word (2000) (Rasin, 1997). Items ranged from a 1.0 to 6.2 grade level, with a 3.5 grade level readability score for the overall questionnaire." (Johnson & Rogers, 2006, p. 339)

Box 16-1 How to Find the Fog Index (Fog Formula)

1. Pick a sample of writing 100 to 125 words long. Count the average number of words per sentence. In counting, treat independent clauses as separate sentences. "In school we studied; we learned; we improved" is three sentences.
2. Count the words of three syllables or more. Do not count: (a) capitalized words, (b) combinations of short words such as butterfly or manpower, or (c) verbs made into three syllables by adding "-es" or "-ed" such as trespasses or created. Divide the count of long words by the number of words in the passage to get the percentage.
3. Add the results from no. 1 (average sentence length) and no. 2 (percentage of long words). Multiply the sum by 0.4. Ignore the numbers after the decimal point.
4. The result is the years of schooling needed to understand the passage tested easily. Few readers have more than 17 years of schooling, so give any passage higher than 17 a Fog Index of 17-plus.

Adapted from Gunning, R., & Kallan, R. A. (1994). *How to take the fog out of business writing*. Chicago, IL: Dartnell. The Fog Index is a service mark licensed exclusively to RK Communication Consultants by D. and M. Mueller.

Validity from Factor Analysis

Factor analysis is a valuable approach for determining evidence of an instrument's construct validity. This analysis technique is used to determine the various dimensions or subcomponents of a phenomenon of interest. To employ factor analysis, the instrument must be administered to a large, representative sample of participants at one time. Usually the data are initially analyzed with exploratory factor analysis (EFA) to examine relationships among the various items of the instrument. Items that are closely related are clustered into a factor. The researcher needs to preset the minimum loading for an item to be included in a factor. The minimum loading is usually set at 0.30 but might be as high as 0.50 (Waltz et al., 2010). The factors identified are the subcomponents of the construct the instrument was developed to measure. Determining and naming the factors identified through EFA require detailed work on the part of the researcher. The researcher can validate the number of factors or subcomponents in the instrument and measurement equivalence among comparison groups through the use of confirmatory factor analysis (CFA). Items that do not fall into a factor (because they do not correlate with other items) may be deleted (DeVon et al., 2007; Munro, 2005; Stommel, Wang, Given, & Given, 1992; Waltz et al., 2010). A more extensive discussion of EFA and CFA is presented in Chapter 23.

Johnson and Rogers (2006) conducted an EFA to determine the factor structure for their MTQ: Purposeful Action scale. The EFA identifies the specific factors or subscales for the scale and the items that fit each of these subscales. The original scale had 20 items sorted into three subscales (labeled perceived need, perceived effectiveness, and perceived as safe) that are identified in Table 16-2. The EFA and the results are presented in Table 16-3 and described as follows.

"Factor analysis is a grouping technique that allows for evaluation of the dimensionality of scales (Munro, 2001; Nunnally & Bernstein, 1994). A principal axis factoring solution with an oblimen rotation, considered the best analysis for achieving a theoretical solution uncontaminated by unique and random error variability was undertaken.…

The EFA yielded two interpretable factors [see Table 16-3], which eliminated six additional items because of factor loadings < 0.40. The first factor merged the need and effectiveness items along with one item from the Safe subscale. This factor was renamed treatment benefits (benefits). The second factor, renamed medication safety (safety), was reduced to three of the original safe subscales items.

The Benefits subscale retained nine items that focused on the actual perceived benefits of treatment, such as preventing a stroke, controlling blood pressure, preventing further health problems, and feeling better when taking medications, which indicated a desire to control blood pressure to maintain and promote health and well-being. The subscale had an eigenvalue of 5.5 and a total item variance explained by the factor of 46%.…

The Safety subscale (three items) focused on side effects of medications. This subscale had an eigenvalue of 1.9 and a total item variance explained by the factor of 16%.… Together, the two factor solution had a coefficient alpha [Cronbach alpha] of 0.87 and an explained variance of 62%." (Johnson & Rogers, 2006, pp. 343-346)

TABLE 16-3 Principal Axis Factor Analysis with Oblimen Rotation Pattern (and Structure in Parentheses) Coefficients for the Medication-Taking Questionnaire: Purposeful Action Two-Factor Solution

	Factor Loadings					Eigenvalue	% Variance Explained	Coefficient Alpha
	1			2	h^2			
Treatment benefits						5.5	45.9	0.90
I need to take my blood pressure pills.	0.84	(0.85)	(0.34)		0.73			
Taking my blood pressure pills is not a problem because they benefit my health.	0.82	(0.84)	(0.35)		0.72			
I could have problems if I do not take my blood pressure pills.	0.81	(0.84)	(0.21)		0.70			
Blood pressure pills keep me from having health-related problems.	0.81	(0.79)	(0.16)		0.63			
My blood pressure pills keep me from having a stroke.	0.75	(0.75)	(0.23)		0.55			
I feel better when I take my blood pressure pills.	0.74	(0.74)	(0.21)		0.55			
My blood pressure pills control my blood pressure.	0.74	(0.74)	(0.26)		0.55			
I am OK if I do not take my blood pressure pills.*	0.72	(0.71)			0.52			
My blood pressure will come down enough without pills.*	0.54	(0.48)			0.30			
Medication safety						1.9	15.6	0.80
The side effects from my blood pressure pills are harmful.*	(0.19)		0.87	(0.86)	0.74			
The side effects from my blood pressure pills are a problem.*	(0.27)		0.84	(0.86)	0.71			
My blood pressure pills cause other health problems.*	(0.29)		0.82	(0.83)	0.70			
Total						7.4	61.5	0.88

From Johnson, M. J., & Rogers, S. (2006). Development of the Purposeful Action Medication-Taking Questionnaire. *Western Journal of Nursing Research, 28*(3), 345.

Note: $n = 229$.

*Item required reverse coding. Factor loadings in parentheses represent structure coefficients. If patterned or structure coefficient is not listed, the value was <0.15.

Johnson and Rogers (2006) provided a clear, concise rationale for the revisions that they made in their MTQ: Purposeful Action scale. In this study, the revised scale showed internal consistency (Cronbach alpha = 0.87) and construct validity obtained through content analysis and EFA. Johnson and Rogers (2006, p. 348) also conducted CFA that "supported the hypothesis that benefits and safety [factors or subscales] underlie the cognitive component of medication taking in hypertensive medications."

Validity from Structural Analysis

Structural analysis is used to examine the structure of relationships among the various items of an instrument. This approach provides insights beyond that provided by factor analysis. Factor analysis determines what items group together. Structural analysis determines how each item is related to other items. Structural analysis goes a step beyond factor analysis. The exact relationship of each item in a factor is examined through correlational analyses.

Convergent Validity

In some cases, instruments are available to measure a construct, such as depression. However, for many possible reasons, the existing instruments may be unsatisfactory for a particular purpose or a particular population, such as measuring major depression in young children, and the researcher may choose to develop a new instrument for a study. In examining the validity of the new instrument, it is important to determine how closely the existing instruments measure the same construct as the newly developed instrument (**convergent validity**). One can administer all of the instruments (the new one and the existing ones) to a sample concurrently and evaluate the results using correlational analyses. If the measures are highly positively correlated, the validity of each instrument is strengthened.

Johnson and Rogers (2006) strengthened the validity of their 12-item MTQ: Purposeful Action scale and its subscales (benefit and safety) by correlating them with a variety of other instruments (Hamilton Health Belief Model Hypertension [HBM] Scale with the HBM subscales of Susceptibility, Severity, Benefits, and Barriers; Lifestyle Busyness Questionnaire with Busyness and Routine subscales; and Blood Pressure Feedback Log). The results of these correlations are presented in Table 16-4. The significant positive correlations of 0.3 to 0.63 between the existing scales (Hamilton HBM Scale with Susceptibility and Benefits subscales and the Blood Pressure Feedback Log for adherent group) and the MTQ and the benefits subscale add to the construct validity of these instruments. This is an example of examining convergent validity for this scale, which was strong for the

TABLE 16-4 Validity Correlation Coefficients for the Medication-Taking Questionnaire: Purposeful Action and Subscales

	MTQ: Purposeful Action	MTQ Benefit Subscale	MTQ Safe Subscale
Hamilton HBM Scale[a]	0.30**	0.43**	–0.12
HBM: Susceptibility subscale	0.36**	0.41**	0.01
HBM: Severity subscale	0.00	0.12	–0.27**
HBM: Benefits subscale	0.58**	0.63**	0.19
HBM: Barriers subscale	–0.49**	–0.42**	–0.41**
Lifestyle Busyness Questionnaire[b]	0.08	0.11	–0.02
Busyness subscale	0.10	0.13	0.01
Routine subscale	–0.07	–0.06	–0.06
Blood Pressure Feedback Log[c]			
Adherent	0.53**	0.54**	0.25*
Nonadherent	–0.60**	–0.50**	–0.53**

From Johnson, M. J., & Rogers, S. (2006). Development of the Purposeful Action Medication-Taking Questionnaire. *Western Journal of Nursing Research, 28*(3), 347.

HBM, Health Belief Model Hypertension Scale.

[a] $n = 107$.

[b] $n = 104$.

[c] $n = 102$.

*$p < 0.05$, two-tailed.

**$p < 0.01$, two-tailed.

MTQ and the benefit subscale but not the safety subscale.

Divergent Validity

Sometimes, instruments can be located that measure a construct opposite to the construct measured by the newly developed instrument (**divergent validity**). For example, if the newly developed instrument measures hope, you could search for an instrument that measures hopelessness or despair. If possible, you could administer this instrument and the instruments used to test convergent validity at the same time. This approach of combining convergent and divergent validity testing of instruments is called **multitrait-multimethod** (MT-MM).

The MT-MM approach can be used when researchers are examining two or more constructs being measured by two or more measurement methods (DeVon et al., 2007). Correlational procedures are conducted with the different scales and subscales. If the convergent measures positively correlate and the divergent measures negatively correlate with other measures, validity for each of the instruments is strengthened. Johnson and Rogers (2006) used an MT-MM approach in examining convergent and divergent validity related to their MTQ: Purposeful Action. The convergent validity findings were discussed in the previous section. Table 16-4 shows that the MTQ and the subscales benefits and safety were significantly, negatively correlated with HBM Barriers subscale and the Blood Pressure Feedback Log for nonadherent hypertensive patients. These scales measure the opposite construct from the MTQ and its subscales, so these significant negative correlations indicated that the construct validity was strengthened for these instruments. The correlations with the Lifestyle Busyness Questionnaire were too low (–0.07 to 0.13) to add to the convergent or divergent validity of the MTQ: Purposeful Action scale and subscales (see Table 16-4).

Validity from Contrasting (or Known) Groups

To test the validity of an instrument, identify groups that are expected (or known) to have contrasting scores on the instrument. Generate hypotheses about the expected response of each of these known groups to the construct. Next, select samples from at least two groups that are expected to have opposing responses to the items in the instrument. Hagerty and Patusky (1995) developed a measure called the Sense of Belonging Instrument (SOBI). They tested the instrument on the following three groups: community college students, clients diagnosed with major depression, and retired Roman Catholic nuns, as described in the following excerpt.

> "The community college sample was chosen for its heterogeneous mix of students and ease of access. Depressed clients were included based on the literature and the researcher's clinical experience that interpersonal relationships and feeling 'connected' are difficult when one is depressed. It was hypothesized that the depressed group would score significantly lower on the SOBI than the student group. The nuns were selected to examine the performance of the SOBI with a group that, in accordance with the theoretical basis of the instrument, should score significantly higher than the depressed and student groups." (Hagerty & Patusky, 1995, p. 10)

The nuns had the highest sense of belonging, the student groups followed, and the depressed group had the lowest sense of belonging. This test increased the validity of the instrument in that the scores of groups were as anticipated.

Evidence of Validity from Discriminant Analysis

Instruments sometimes have been developed to measure constructs closely related to the construct measured by a newly developed instrument. For example, an instrument might exist to measure medication management in patients with diabetes that is similar to the MTQ: Purposeful Action developed by Johnson and Rogers (2006) for patients with hypertension. If such instruments can be located, you can strengthen the validity of the MTQ instrument and the other medication management instrument by testing the extent to which the two instruments can finely discriminate between these related concepts. Testing of this discrimination involves administering the two instruments simultaneously to a sample and performing a discriminant analysis (see Kerlinger & Lee, 2000, for a discussion of discriminant analysis).

Validity from Prediction of Future Events and Concurrent Events

The ability to predict future performance or attitudes on the basis of instrument scores adds to the validity of an instrument. Nurse researchers often want to determine the ability of scales developed to measure selected health behaviors to predict the future health status of individuals. One approach might be to examine reported stress levels of selected individuals in highly stressful careers such as nursing and see if stress is

linked to the nurses' future incidence of hypertension. If study analysis links stress to future hypertension, measuring a nurse's stress could be used to predict his or her future likelihood of becoming hypertensive. For example, the validity of the Nursing Stress Scale (NSS) could be tested in this manner. French, Lenton, Walters, and Eyles (2000) did an expanded evaluation of the reliability and validity of the NSS with a random sample of 2280 nurses working in a wide range of healthcare settings. They noted that the NSS included nine subscales: death and dying, conflict with physicians, inadequate preparation, problems with supervisors, workload, problems with peers, uncertainty concerning treatment, patients and their families, and discrimination. Confirmatory factor analyses supported the factor structure. Cronbach alpha coefficients of eight of the subscales were 0.70 or higher. The NNS showed reliability and validity in measuring stress in nurses and could be used in a study to determine the link to hypertension. The accuracy of **predictive validity** is determined through regression analysis.

Validity can be tested by examining the ability to predict the current value of one measure on the basis of the value obtained on the measure of another concept. For example, you might be able to predict the self-esteem score of an individual who had a high score on an instrument to measure coping. A person who received a high score on coping might be expected also to have a high self-esteem score. If these results held true in a study in which both measures were obtained concurrently, the two instruments would have evidence of **concurrent validity**.

Successive Verification of Validity

After the initial development of an instrument, it is hoped that other researchers would begin using the instrument in additional studies. Each of these studies could add to the validity and reliability information on the instrument. There is a successive verification of the validity of the instrument over time when used in a variety of studies with different populations and settings. For example, additional researchers are using the MTQ: Purposeful Action in their studies, which has the potential to add to the validity of this questionnaire (Lehane & McCarthy, 2007).

Accuracy, Precision, and Error of Physiological Measures

Accuracy and precision of physiological and biochemical measures tend not to be reported in published studies. These routine **physiological measures** are assumed to be accurate and precise, an assumption that is not always correct. The most common physiological measures used in nursing studies are blood pressure, heart rate, weight, and temperature. These measures are often obtained from the patient's record with no consideration given to their accuracy. It is important to consider the possibility of differences between the obtained value and the true value of physiological measures. Thus, researchers using physiological measures need to provide evidence of the accuracy and precision of their measures (Ryan-Wenger, 2010).

The evaluation of physiological measures may require a slightly different perspective from that applied to behavioral measures, in that standards for most biophysical measures are defined by national and international organizations such as the International Organization of Standardization (IOS) (2011a) and the Clinical Laboratory Standards Institute (CLSI) (2011). CLSI develops standards for laboratory and other healthcare-related biophysical measures. The IOS is the world's largest developer and publisher of international standards and includes a network of 160 countries (see IOS website for details at http://www.iso.org/iso/home.htm). The ISO standards were developed to accomplish the following:

- Make the development, manufacturing, and supply of products and services more efficient, safer, and cleaner
- Facilitate trade between countries and make it fairer
- Provide governments with a technical base for health, safety, and environmental legislation and conformity assessment
- Share technological advances and good management practice
- Disseminate innovations
- Safeguard consumers and users in general of products and services
- Make life simpler by providing solutions to common problems (ISO, 2011b)

You can locate the standards for different biophysical equipment, products, or services that you might use in a study or in clinical practice. Within IOS, the Joint Committee for Guides in Metrology (JCGM) has two major areas of focus: (1) Guide to the Expression of Uncertainty in Measurement (GUM) and (2) International Vocabulary of Basic and General Terms in Metrology (VIM) (JCGM, 2011). VIM is a document that standardizes terminology related to biophysical measurements, such as accuracy, precision, error, sensitivity, specificity, and likelihood ratio that are described in this section.

Accuracy

Accuracy involves determining the closeness of the agreement between the measured value and the true value of the quantity being measured (JCGM, 2011). Accuracy is similar to validity, in which evidence of content-related validity addresses the extent to which the instrument measured the construct or domain defined in the study. New measurement devices are compared with existing standardized methods of measuring a biophysical property or concept. For example, measures of oxygen saturation with a pulse oximeter were correlated with arterial blood gas measures of oxygen saturation to determine the accuracy of the pulse oximeter. Thus, there should be a very strong, positive correlation (≥0.95) between pulse oximeter and blood gas measures of oxygen saturation to support the accuracy of the pulse oximeter (CLSI, 2011).

Accuracy of physiological measures depends on the (1) quality of the measurement equipment or device, (2) detail of the data collection plan, and (3) expertise of the data collector (Ryan-Wenger, 2010). The data collector or person conducting the biophysical measures must do the measurements in a standardized way that is usually directed by a measurement protocol. For example, BP readings in a study need to be taken using a protocol: (1) place the subject in a chair and allow 5 minutes of rest; (2) remove restrictive clothing from the subject's arm; (3) measure the subject's upper arm and select the appropriate cuff size; (4) instruct the subject to place his or her feet flat on the floor; (5) support the subject's arm when taking the BP reading; and (6) take three BP readings each 5 minutes apart, average the readings, and enter the averaged BP reading into a computer. Some measurements, such as arterial pressure, can be obtained by the biomedical device producing the reading and automatically recorded in a computerized database. This type of data collection greatly reduces the potential for error and increases accuracy and precision.

The biomedical device or equipment used to measure a study variable must be examined for accuracy. Researchers need to document the extent to which the biophysical measure is an accurate measurement of a study variable and the level of error expected. Reviewing the ISO (2011b) and CLSI (2011) standards could provide essential accuracy information and information about the company that developed the device or equipment.

Selectivity, an element of accuracy, is "the ability to identify correctly the signal under study and to distinguish it from other signals" (Gift & Soeken, 1988, p. 129). Because body systems interact, the researcher must choose instruments that have selectivity for the dimension being studied. For example, electrocardiographic readings allow one to differentiate electrical signals coming from the myocardium from similar signals coming from skeletal muscles.

To determine the accuracy of biochemical measures, review the standards set by CLSI (2011) and determine if the laboratory where the measures are going to be obtained is certified. Most laboratories are certified, so researchers could contact experts in the agency on the laboratory procedure and ask them to describe the process for collection, analysis, and values obtained for specimens. You might also ask these experts to judge the appropriateness of the biophysical device for the construct being measured in the study. Use contrasted groups' techniques by selecting a group of subjects known to have high values on the biochemical measures and comparing them with a group of subjects known to have low values on the same measure. In addition, to obtain concurrent validity, compare the results of the test with results from the use of a known standard (CLSI, 2011), such as the example of the comparison of pulse oximeter values with blood gas values for oxygen saturation.

Precision

Precision is the degree of consistency or reproducibility of measurements made with physiological instruments or devices. There should be close agreement in the replicated measures of the same variable or object under specified conditions (Ryan-Wenger, 2010). Precision is similar to reliability. The precision of most physiological devices or equipment is determined by the manufacturer and is part of quality control testing done in the agency using the device. Similar to accuracy, precision depends on the collector of the biophysical measures and the consistency of the measurement equipment or device. The protocol for collecting the biophysical measures improves precision and accuracy (see the previous example of protocol to measure BP).

The data collectors need to be trained to ensure consistency, which is documented with intrarater (within a single data collector) and interrater (among data collectors) percentages of agreements. The kappa coefficient of agreement is one of the most common and simplest statistics to determine intrarater and interrater accuracy and precision for nominal level data (Cohen, 1960; Ryan-Wenger, 2010). The equipment and devices used to measure physiological variables need to be maintained according to the standards set by IOS and the manufacturers of the devices. Many devices need to be recalibrated according to set criteria

to ensure consistency in measurements. Because of fluctuations in some physiological measures, test-retest reliability might be inappropriate.

Two procedures are commonly used to determine the precision of biochemical measures. One is the Levy-Jennings chart. For each analysis method, a control sample is analyzed daily for 20 to 30 days. The control sample contains a known amount of the substance being tested. The mean, the standard deviation, and the known value of the sample are used to prepare a graph of the daily test results. Only 1 value of 22 is expected to be greater than or less than 2 standard deviations from the mean. If two or more values are more than 2 standard deviations from the mean, the method is unreliable in that laboratory. Another method of determining the precision of biochemical measures is the duplicate measurement method. The same technician performs duplicate measures on randomly selected specimens for a specific number of days. The results are essentially the same each day if there is high precision. Results are plotted on a graph, and the standard deviation is calculated on the basis of difference scores. The use of correlation coefficients is not recommended (DeKeyser & Pugh, 1990).

Sensitivity

Sensitivity of physiological measures relates to "the amount of change of a parameter that can be measured precisely" (Gift & Soeken, 1988, p. 130). If changes are expected to be small, the instrument must be very sensitive to detect the changes. Thus, sensitivity is associated with effect size (see Chapter 15). With some instruments, sensitivity may vary at the ends of the spectrum. This is referred to as the *frequency response*. The stability of the instrument is also related to sensitivity. This feature may be judged in terms of the ability of the system to resume a steady state after a disturbance in input. For electrical systems, this feature is referred to as *freedom from drift* (Gift & Soeken, 1988).

Error

Sources of **error in physiological measures** can be grouped into the following five categories: environment, user, subject, machine, and interpretation. The environment affects both the machine and the subject. Environmental factors include temperature, barometric pressure, and static electricity. User errors are caused by the person using the instrument and may be associated with variations by the same user, different users, changes in supplies, or procedures used to operate the equipment. Subject errors occur when the subject alters the machine or the machine alters the subject. In some cases, the machine may not be used to its full capacity. Machine error may be related to calibration or to the stability of the machine. Signals transmitted from the machine are also a source of error and can cause misinterpretation (Ryan-Wenger, 2010).

Sources of error in biochemical measures are biological, preanalytical, analytical, and postanalytical. Biological variability in biochemical measures is due to factors such as age, gender, and body size. Variability in the same individual is due to factors such as diurnal rhythms, seasonal cycles, and aging. Preanalytical variability is due to errors in collecting and handling of specimens. These errors include sampling the wrong patients; using an incorrect container, preservative, or label; lysis of cells; and evaporation. Preanalytical variability may also be due to patient intake of food or drugs, exercise, or emotional stress. Analytical variability is associated with the method used for analysis and may be due to materials, equipment, procedures, and personnel used. The major source of postanalytical variability is transcription error. This source of error can be greatly reduced by entering data into the computer directly (DeKeyser & Pugh, 1990).

When the scores obtained in a study are at the interval or ratio level, a commonly used method of evaluating precision and accuracy errors is the Bland-Altman chart (Bland & Altman, 1986). This chart is a scatter plot of the differences between observed scores on the *Y*-axis and the combined mean of the two methods on the *X*-axis. The distribution of the difference scores is examined in context of the limits of agreement that are drawn as a horizontal line across the chart or scatter plot (see Chapter 23). The limits are set by the researchers and might include 1 or 2 standard deviations from the mean or might be the clinical standards of the maximum amount of error that is safe. The data points are examined for level of agreement (congruence) and for level of bias (systematic error). Outliers are readily visible from the chart, and each outlier case should be examined to identify the cause of such a large discrepancy. Clinical laboratory standards indicate that "more than 3 outliers per 100 observations suggest there are major flaws in the measurement system" (Ryan-Wenger, 2010, p. 381).

Schell et al. (2011) conducted a study to compare upper arm and calf automatic noninvasive BPs in children in a pediatric intensive care unit (PICU). The researchers documented the accuracy of their BP monitoring equipment, training of their data collectors, and the procedures for taking the BPs in their study. The errors in precision and accuracy are documented with Bland-Altman charts for systolic BP, diastolic BP, and

mean arterial pressure readings. The chart of the systolic BP is included as an example in Figure 16-10. This study was conducted to determine an alternative method of obtaining BP when the injuries of the child prevent BP readings using the upper arm.

"BP Monitor

"BP was obtained using a Spacelabs Ultraview SL monitoring system (Spacelabs Healthcare, Issaquah, WA), which consists of hemodynamic parameter modules that can be inserted into stationary bedside and portable monitor housings. All monitoring functions were controlled through the modules. During data collection, each set of arm and calf BP measurements was obtained simultaneously using two identical parameter modules: one inserted into the subject's stationary bedside housing and the other inserted in to a portable monitor housing brought to the subject's bedside. Modules and housings are inspected and tested annually by Biomedical Support Services to ensure accurate functioning. The accuracy of these monitors for arm BPs meets or exceeds SP10-1992 Association for the Advancement of Medical Instrumentation standards (mean error = ±4.5 mm Hg, *SD* = ±7.3 mm Hg) for arm measurements (White et al., 1993). Spacelabs Healthcare did not report data regarding accuracy of calf BPs.

"Training of Data Collectors

"Data were collected by five pediatric intensive care nurses who attended a data training session that addressed location of arm and calf sites, measurement of limb circumference, and use of the RASS [Richmond Agitation Sedation Scale]. The nurses also attended a BP monitor in-service offered by the Spacelab representative when the monitors were adopted in the PICU in January 2006.…

"Procedure

"Subjects were placed in a supine position with the head of bed elevated 30° as determined by a handheld protractor or the degree indicator incorporated into the bed frame. Subjects remained in this position for at least 5 minutes prior to data collection. Cuff sizes were selected based on limb circumferences measured to the nearest 0.5 cm. Spacelabs cuff sizes were as follows: neonate, 6-11 cm; infant, 8-11 cm; child, 12-19 cm; small adult, 17-26 cm; and adult, 24-32 cm. Per manufacturer's recommendations, if circumference overlapped two categories of cuff size, the larger cuff was selected. Using a paper tape measure, arm circumference was obtained at the point halfway between the elbow and the shoulder. Calf circumference was measured at the point midway between the ankle and the knee. The BP cuffs were applied to the arm and calf on the same side. Subjects' extremities were positioned at the side of their bodies, resting on the bed, for all measurements.… Systolic, diastolic, and mean BP values for the arm and calf as well as a simultaneous heart rate were documented. Data collectors notified the child's nurse or physician if an abnormal arm reading was obtained." (Schell et al., 2011, pp. 6-7)

"To promote best practice, clinicians should base treatment choices on individual patient data, not group data. Therefore, Bland-Altman analyses were used to determine agreement between arm and calf oscillometric BPs for individual subjects. Perfect agreement occurs when all data points lie on the line of equality of the *X*-axis. The bias (mean difference between arm and calf pressures) systolic BP was 8.0 mm Hg with the limits of agreement −18.9 and 34.9 mm Hg. Limits of agreement indicated that 95% of the sample falls between these values [see Figure 16-10]. The limits of agreement for diastolic BP were −22.7 and 25.0 mm Hg with a bias of 1.1 mm Hg." (Schell et al., 2011, p. 9)

Schell et al. (2011) provided evidence of the accuracy, precision, and error of the BP monitoring equipment used in their study. They also provided a detailed discussion of the procedures for data collection that followed a rigorous protocol to ensure accurate and precise BP readings were obtained for all ages of children based on their measured arm and calf sizes. The data collectors were trained in BP monitoring by the Spacelab representative, which would increase their expertise in the use of the equipment. However, the study would have been strengthened by a discussion of the intrarater and interrater percentage of agreement for the data collectors. The use of the Bland-Altman plot to identify the error in precision and accuracy for systolic BPs, diastolic BPs, and mean arterial pressures added to the credibility of the findings. The researchers found that the arm and calf BPs were not interchangeable for many of the children 1 to 8 years old. "Clinical BP differences were the greatest in children between ages 2 and less than 5 years. Calf BPs are not recommended for this population. If the calf is unavoidable due to medical reasons, trending of BP

from this site should remain consistent during the child's stay" (Schell et al., 2011, p. 10).

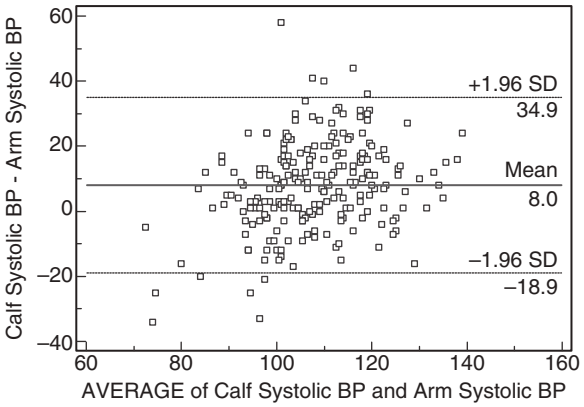

Figure 16-10 Bland-Altman plot of systolic BP. (From Schell, K., Briening, E., Lebet, R., Pruden, K., Rawheiser, S., & Jackson, B. [2011]. Comparison of arm and calf automatic noninvasive blood pressures in pediatric intensive care patients. *Journal of Pediatric Nursing, 26*[1], 9.)

Sensitivity, Specificity, and Likelihood Ratios

An important part of building evidence-based practice (EBP) is the development, refinement, and use of quality diagnostic tests and measures in research and practice. Researchers want to use the most accurate and precise measure or test in their study to promote quality outcomes. If a quality diagnostic test does not exist, some nurses have participated in the development and refinement of new biophysical tests. Clinicians want to know what diagnostic test, such as laboratory or imaging study, to order to help screen for and accurately determine the absence or presence of an illness (Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). When you order a diagnostic test, how can you be sure that the results are valid or accurate? This question is best answered by current, quality research to determine the sensitivity and specificity of the test.

Sensitivity and Specificity

The **accuracy of a screening test** or a test used to confirm a diagnosis is evaluated in terms of its ability to assess correctly the presence or absence of a disease or condition as compared with a gold standard. The **gold standard** is the most accurate means of currently diagnosing a particular disease and serves as a basis for comparison with newly developed diagnostic or screening tests (Campo, Shiyko, & Lichtman, 2010). If the test is positive, what is the probability that the disease is present? If the test is negative, what is the probability that the disease is not present? When you talk to the patient about the results of their tests, how sure are you that they do or do not have the disease? **Sensitivity** and **specificity** are the terms used to describe the accuracy of a screening or diagnostic test (Table 16-5). There are four possible outcomes of a screening test for a disease: (1) **true positive**, which accurately identifies the presence of a disease; (2) **false positive**, which indicates a disease is present when it is not; (3) **true negative**, which indicates accurately that a disease is not present; or (4) **false negative**, which indicates that a disease is not present when it is (Campo et al., 2010; Grove, 2007). The 2 × 2 contingency table shown in Table 16-5 should help you to visualize sensitivity and specificity and these four outcomes (Craig & Smyth, 2012; Sackett et al., 2000).

Sensitivity and specificity can be calculated based on research findings and clinical practice outcomes to determine the most accurate diagnostic or screening tool to use in identifying the presence or absence of a disease for a population of patients. The calculations for sensitivity and specificity are provided as follows:

TABLE 16-5 Results of Sensitivity and Specificity of Screening Tests

Diagnostic Test Result	Disease Present	Disease Not Present or Absent	Total
Positive test	a (true positive)	b (false positive)	a + b
Negative test	c (false negative)	d (true negative)	c + d
Total	a + c	b + d	a + b + c + d

From Grove, S. K. (2007). *Statistics for health care research: A practical workbook*. Philadelphia, PA: Saunders, p. 335.
a = The number of people who have the disease and the test is positive (true positive).
b = The number of people who do not have the disease and the test is positive (false positive).
c = The number of people who have the disease and the test is negative (false negative).
d = The number of people who do not have the disease and the test is negative (true negative).

Sensitivity calculation = probability of disease
= a/(a + c)
= true positive rate

Specificity calculation = probability of disease
= d/(b + d)
= true negative rate

Sensitivity is the proportion of patients with the disease who have a positive test result or true positive. The ways the researcher or clinician might refer to the test sensitivity include the following:

- *Highly sensitive test* is very good at identifying the patient with a disease.
- If a test is highly sensitive, it has a low percentage of false negatives.
- *Low sensitivity test* is limited in identifying the patient with a disease.
- If a test has low sensitivity, it has a high percentage of false negatives.
- If a sensitive test has negative results, the patient is less likely to have the disease.
- Use the acronym *SnNout:* High sensitivity (*Sn*), test is negative (*N*), rules the disease out (*out*) (Campo et al., 2010; Grove, 2007).

Specificity of a screening or diagnostic test is the proportion of patients without the disease who have a negative test result or true negative. The ways the researcher or clinician might refer to the test specificity include the following:

- *Highly specific test* is very good at identifying patients without a disease.
- If a test is very specific, it has a low percentage of false positives.
- *Low specificity test* is limited in identifying patients without a disease.
- If a test has low specificity, it has a high percentage of false positives.
- If a specific test has positive results, the patient is more likely to have the disease.
- Use the acronym *SpPin:* High specificity (*Sp*), test is positive (*P*), rules the disease in (*in*).

Sarikaya, Aktas, Ay, Cetin, and Celikmen (2010) conducted a study to determine the sensitivity and specificity of rapid antigen diagnostic testing (RADT) for diagnosing pharyngitis in patients in the emergency department. Acute pharyngitis is primarily a viral infection, but in 10% of the cases it is caused by bacteria. Most cases of bacterial pharyngitis are caused by group A beta-hemolytic streptococci (GABHS). One laboratory method for diagnosing GABHS is RADT, which has become more popular than a throat culture because it can be processed rapidly during an emergency department and primary care visit.

> "We conducted a study to define the sensitivity and specificity of RADT, using throat culture results as the gold standard, in 100 emergency department patients who presented with symptoms consistent with streptococcal pharyngitis. We found that RADT had a sensitivity of 68.2% (15 of 22), a specificity of 89.7% (70 of 78), a positive predictive value of 65.2% (15 of 23), and a negative predictive value of 90.9% (70 of 77). We conclude that RADT is useful in the emergency department when the clinical suspicion is GABHS, but results should be confirmed with a throat culture in patients whose RADT results are negative." (Sarikaya et al., 2010, p. 180)

The results of the study by Sarikaya et al. (2010) were put into Table 16-6 so that you might see how the sensitivity and specificity were calculated in this study.

Sensitivity calculation = probability of disease
= a/(a + c) = true positive rate

Sensitivity = probability of GABHS pharyngitis
= 15/(15 + 7) = 15 / 22 = 68.18% = 68.2%

TABLE 16-6 Results of Sensitivity and Specificity of Rapid Antigen Diagnostic Testing (RADT)

RADT Result	GABHS Disease Present	GABHS Disease Absent	Total
Positive test	a (true positive) = 15	b (false positive) = 8	a + b = 15 + 8 = 23
Negative test	c (false negative) = 7	d (true negative) = 70	c + d = 7 + 70 = 77
Total	a + c = 15 + 7 = 22	b + d = 8 + 70 = 78	a + b + c + d = 100

GABHS, Group A beta-hemolytic streptococci.
a = The number of people who have GABHS pharyngitis disease and the test is positive (true positive).
b = The number of people who do not have GABHS pharyngitis disease and the test is positive (false positive).
c = The number of people who have GABHS pharyngitis disease and the test is negative (false negative).
d = The number of people who do not have GABHS pharyngitis disease and the test is negative (true negative).

Specificity calculation = probability of disease
= d/(b + d) = true negative rate

Specificity = probability no GABHS pharyngitis
= 70/(8 + 70) = 70/78 = 89.74% = 89.7%

The sensitivity of 68.2% indicates the percentage of patients with a positive RADT who had GABHS pharyngitis (true positive rate). The specificity of 89.7% indicates the percentage of patients with a negative RADT who did not have GABHS pharyngitis (true negative rate). In developing a diagnostic or screening test, researchers need to achieve the highest sensitivity and specificity possible. In selecting screening tests to diagnose illnesses, clinicians need to determine the most sensitive and specific screening test but also need to examine cost and ease of access to these tests in making their final decision (Craig & Smyth, 2012; Grove, 2007; Sackett et al., 2000).

Likelihood Ratios

Likelihood ratios (LRs) are additional calculations that can help researchers to determine the accuracy of diagnostic or screening tests, which are based on the sensitivity and specificity results. LRs are calculated to determine the likelihood that a positive test result is a true positive and a negative test result is a true negative. The ratio of the true positive results to false positive results is known as the positive LR (Campo et al., 2010). The **positive LR** is calculated as follows using the data from the study by Sarikaya et al. (2010):

Positive LR = sensitivity ÷ 100% − specificity

Positive LR for GABHS pharyngitis
= 68.2% ÷ 100% − 89.7% = 68.2% ÷ 10.3% = 6.62

The **negative LR** is the ratio of true negative results to false negative results, and it is calculated as follows:

Negative LR = 100% − sensitivity ÷ specificity

Negative LR for GABHS pharyngitis
= 100% − 68.2% ÷ 89.7% = 31.8% ÷ 89.7% = 0.35

The very high LRs (or LRs that are >10) rule in the disease or indicate that the patient has the disease. The very low LRs (or LRs that are <0.1) virtually rule out the chance that the patient has the disease (Campo et al., 2010; Craig & Smyth, 2012; Melnyk & Fineout-Overholt, 2011; Sackett et al., 2000). Understanding sensitivity, specificity, and LR increases your ability to read clinical studies and to determine the most accurate diagnostic test to use in research and clinical practice.

KEY POINTS

- Measurement is the process of assigning numbers to objects, events, or situations in accord with some rule.
- Instrumentation is the application of specific rules to develop a measurement device or instrument.
- Measurement theory and the rules within this theory have been developed to direct the measurement of abstract and concrete concepts.
- There is direct measurement and indirect measurement.
- Healthcare technology has made researchers familiar with direct measures of concrete elements, such as height, weight, heart rate, temperature, and blood pressure.
- Indirect measurement is used with abstract concepts, when the concepts are not measured directly, but when the indicators or attributes of the concepts are used to represent the abstraction. Common abstract concepts measured in nursing include anxiety, stress, coping, quality of life, and pain.
- Measurement error is the difference between what exists in reality and what is measured by a research instrument.
- The levels of measurement, from lower to higher, are nominal, ordinal, interval, and ratio.
- Reliability refers to how consistently the measurement technique measures the concept of interest and includes stability reliability, equivalence reliability, and internal consistency.
- The validity of an instrument is determined by the extent to which the instrument actually reflects the abstract construct being examined and includes such types as face and content validity, validity from factor analysis, validity from structural analysis, convergent validity, divergent validity, validity from contrasting groups, validity from discriminant analysis, validity from prediction of future and concurrent events, and successive verification validity.
- Evaluation of physiological measures requires a different perspective from that of behavioral measures and requires evaluation for accuracy, precision, and error.
- The accuracy of screening or diagnostic tests is determined by calculating the sensitivity, specificity, and likelihood ratios for the test.

REFERENCES

American Psychological Association's Committee to Develop Standards. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Armstrong, G. D. (1981). Parametric statistics and ordinal data: A pervasive misconception. *Nursing Research*, *30*(1), 60–62.

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237–3243.

Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. *Ultrasound Obstetric Gynecology*, *31*(4), 466–475.

Berk, R. A. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, *12*(5), 659–671.

Bialocerkowski, A., Klupp, N., & Bragge, P. (2010). Research methodology series: How to read and critically appraise a reliability article. *International Journal of Therapy & Rehabilitation*, *17*(3), 114–120.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *1*(8476), 307–310.

Bliley, A. V., & Ferrans, C. E. (1993). Quality of life after coronary angioplasty. *Heart & Lung*, *22*(3), 193–199.

Brinberg, D., & McGrath, J. E. (1985). *Validity and the research process*. Beverly Hills, CA: Sage.

Campbell, D. T, & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Campo, M., Shiyko, M. P., & Lichtman, S. W. (2010). Sensitivity and specificity: A review of related statistics and controversies in the context of physical therapist education. *Journal of Physical Therapy Education*, *24*(3), 69–78.

Clinical and Laboratory Standards Institute (CLSI). (2011). *Harmonized terminology database*. Retrieved from http://www.clsi.org/Content/NavigationMenu/Resources/HarmonizedTerminologyDatabase/Harmonized_Terminolo.htm.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Education & Psychological Measurement*, *20*(1), 37–46.

Corty, E. W. (2007). *Using and interpreting statistics: A practical text for the health, behavioral, and social sciences*. St. Louis, MO: Mosby.

Craig, J. V., & Smyth, R. L. (2012). *The evidence-base practice manual for nurses* (3rd ed.). Edinburgh, Scotland: Churchill Livingstone.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Los Angeles, CA: Sage.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194–197.

DeKeyser, F. G, & Pugh, L. C. (1990). Assessment of the reliability and validity of biochemical measures. *Nursing Research*, *39*(5), 314–317.

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., et al. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155–164.

Dickerson, S. S., Kennedy, M., Wu, Y. B., Underhill, M., & Othman, A. (2010). Factors related to quality-of-life pattern changes in recipients of implantable defibrillators. *Heart & Lung*, *39*(6), 466–476.

Doran, D. M. (2011). *Nursing outcomes: The state of the science* (2nd ed.). Sudbury, MA: Jones & Bartlett Learning.

Fawcett, J., & Garity, J. (2009). *Evaluating research for evidence-based nursing practice*. Philadelphia: F.A. Davis.

Flesch, R. (1984). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

French, S. E., Lenton, R., Walters, V., & Eyles, J. (2000). An empirical evaluation of an expanded nursing stress scale. *Journal of Nursing Measurement*, *8*(2), 161–178.

Gift, A. G., & Soeken, K. L. (1988). Assessment of physiologic instruments. *Heart & Lung*, *17*(2), 128–133.

Goodwin, L. D., & Goodwin, W. L. (1991). Estimating construct validity. *Research in Nursing & Health*, *14*(3), 235–243.

Grove, S. K. (2007). *Statistics for health care research: A practical workbook*. Philadelphia, PA: Saunders.

Gunning, R., & Kallan, R. A. (1994). *How to take the fog out of business writing*. Chicago, IL: Dartnell.

Hagerty, B. M. K., & Patusky, K. (1995). Developing a measure of sense of belonging. *Nursing Research*, *44*(1), 9–13.

Hockenberry, M. J., & Wilson, D. (2009). *Wong's essentials of pediatric nursing* (8th ed.). St. Louis, MO: Mosby.

Imle, M. A., & Atwood, J. R. (1988). Retaining qualitative validity while gaining reliability and validity: Development of the Transition to Parenthood Concerns Scale. *Advanced Nursing Science*, *11*(1), 61–75.

International Organization for Standardization (ISO). (2011a). *Standards development*. Retrieved from http://www.iso.org/iso/standards_development.htm.

International Organization for Standardization (ISO). (2011b). *Discover IOS: What standards do.* Retrieved from http://www.iso.org/iso/about/discover-iso_what-standards-do.htm.

Johnson, M. J. (2002). *The development and testing of three medication-taking questionnaires for the medication adherence model constructs for hypertensive patients*. Unpublished doctoral dissertation. University of Utah, Salt Lake City, UT.

Johnson, M. J., & Rogers, S. (2006). Development of the Purposeful Action Medication-Taking Questionnaire. *Western Journal of Nursing Research*, *28*(3), 335–351.

Johnson, M. J., Williams, M., & Marshall, E. S. (1999). Adherent and nonadherent medication-taking elderly hypertensive patients. *Clinical Nursing Research*, *8*(4), 318–335.

Joint Committee for Guides in Metrology. (2011). *JCGM: Joint Committee for Guides in Metrology*. Retrieved from http://www.iso.org/sites/JCGM/GUM-introduction.htm.

Kaplan, A. (1963). *The conduct of inquiry: Methodology for behavioral science*. New York, NY: Harper & Row.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.) Fort Worth, TX: Harcourt College Publishers.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research*, *39*(2), 121–123.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.

Lehane, E., & McCarthy, G. (2007). An examination of intentional and unintentional aspects of medication non-adherence in patients diagnosed with hypertension. *Journal of Clinical Nursing*, *16*(4), 698–706.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–385.

Melnyk, B. M., & Fineout-Overholt, E. (2011). *Evidence-based practice in nursing & healthcare: A guide to best practice* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Miller, B., & Bodie, M. (1994). Determination of reading comprehension level for effective patient health-education materials. *Nursing Research*, *43*(2), 118–119.

Munro, B. H. (2001). *Statistical methods for health care research* (4th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Munro, B. H. (2005). *Statistical methods for health care research* (5th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measures*, *1*, 385–394.

Rasin, J. H. (1997). Measurement issues with the elderly. In M. Frank-Stromberg & S. J. Olsen (Eds.), *Instruments for clinical health-care research* (2nd ed., pp. 44–53). Boston, MA: Jones & Bartlett.

Ryan-Wenger, N. A. (2010). Evaluation of measurement precision, accuracy, and error in biophysical data for clinical research and practice. In C. F. Waltz, O. L. Strickland, & E. R. Lenz (Eds.), *Measurement in nursing and health research* (4th ed.) (pp. 371–383). New York, NY: Springer Publishing Company.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM* (2nd ed.). Edinburgh: Churchill Livingstone.

Sarikaya, S., Aktas, C., Ay, D., Cetin, A., & Celikmen, F. (2010). Sensitivity and specificity of rapid antigen detection testing for diagnosing pharyngitis in emergency department. *Ear Nose & Throat Journal*, *89*(4), 180–182.

Sax, G. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth Publishing Company.

Schell, K., Briening, E., Lebet, R., Pruden, K., Rawheiser, S., & Jackson, B. (2011). Comparison of arm and calf automatic noninvasive blood pressures in pediatric intensive care patients. *Journal of Pediatric Nursing*, *26*(1), 3–12.

Silvaroli, N. J. (1986). *Classroom reading inventory* (5th ed.). Dubuque, IA: William C. Brown.

Spielberger, C. D., Gorsuch, R. L., & Lushene, P. R. (1970). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Palo Alto, CA: Consulting Psychologists Press.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stommel, M., Wang, S., Given, C. W., & Given, B. (1992). Confirmatory factor analysis (CFA) as a method to assess measurement equivalence. *Research in Nursing & Health*, *15*(5), 399–405.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). Oxford, UK: Oxford University Press.

Thomas, S. (1992). Face validity. *Western Journal of Nursing Research*, *14*(1), 109–112.

Topf, M. (1988). Interrater reliability decline under covert assessment. *Nursing Research*, *37*(1), 47–49.

Waltz, C. W., & Bausell, R. B. (1981). *Nursing research: Design, statistics and computer analysis*. Philadelphia: F. A. Davis.

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2010). *Measurement in nursing and health research* (4th ed.). New York, NY: Springer Publishing Company.

White, W. B., Berson, A. S., Robbins, C., Jamieson, M. J., Prisant, L. M., Roccella, E., et al. (1993). National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers. *Hypertension*, *21*(4), 504–509.