NINTH EDITION

NURSING RESEARCH

Generating and Assessing Evidence *for* Nursing Practice

Denise F. Polit • Cheryl Tatano Beck

Wolters Kluwer
Health

Lippincott
Williams & Wilkins

thePoint

14

Measurement and Data Quality

An ideal data collection procedure is one that captures a construct in a way that is accurate, truthful, and sensitive. Biophysiologic methods have a higher chance of success in attaining these goals than self-report or observational methods, but no method is flawless. In this chapter, we discuss criteria for evaluating the quality of data obtained with structured instruments.

We begin by discussing principles of measurement. Our discussion is based primarily on **classical measurement theory (CMT)**, the leading theory with regard to the measurement of affective constructs (i.e., constructs such as self-esteem or depression). An alternative measurement theory (*item response theory* or *IRT*) has gained in popularity, especially for measuring cognitive constructs (e.g., knowledge). We discuss IRT briefly in Chapter 15.

MEASUREMENT

Quantitative studies derive data through the measurement of variables. **Measurement** involves assigning numbers to represent the amount of an attribute present in an object or person, using a specified set of rules. Quantification and measurement go hand in hand. Attributes are not constant; they vary from day to day or from one person to another. Variability is presumed to be capable of a numeric expression signifying *how much* of an attribute is present. The purpose of assigning numbers is to differentiate between people with varying degrees of the attribute.

Rules and Measurement

Measurement involves assigning numbers according to rules. Rules for measuring temperature, weight, and other physical attributes are familiar to us. Rules for measuring many variables for nursing studies, however, have to be invented. Whether the data are collected by observation, self-report, or some other method, researchers must specify criteria for assigning numeric values to the characteristic of interest.

As an example, suppose we were studying parental attitudes toward dispensing condoms in school clinics, and we asked parents their extent of agreement with the following statement:

> Teenagers should have access to contraceptives in school clinics.
>
> □ Strongly disagree
> □ Disagree
> □ Slightly disagree
> □ Neither agree nor disagree
> □ Slightly agree
> □ Agree
> □ Strongly agree

Responses to this question can be quantified by developing a system for assigning numbers to them. Note that *any* rule would satisfy the definition of measurement. We could assign the value of 30 to "strongly agree," 28 to "agree," 20 to "slightly agree," and so on, but there is no justification for doing so. In measuring attributes, researchers strive to use good, meaningful rules. Without *a priori* knowledge of the "distance" between response options, the most practical approach is to assign a 7 to "strongly agree" and a 1 to "strongly disagree." This rule would quantitatively differentiate, in increments of one point, among people with seven different opinions. Researchers seldom know in advance if their rules are the best possible. New measurement rules reflect hypotheses about how attributes vary. The adequacy of the hypotheses—that is, the worth of the instruments—needs to be assessed empirically.

Researchers try to link numeric values to reality. To state this goal more technically, measurement procedures are ideally isomorphic to reality. The term *isomorphism* signifies equivalence or similarity between two phenomena. An instrument cannot be useful unless the measurements resulting from it correspond with the real world.

To illustrate the concept of isomorphism, suppose a standardized test was administered to 10 students, who obtained the following scores: 345, 395, 430, 435, 490, 505, 550, 570, 620, and 640. These values are shown at the top of Figure 14.1. Suppose that in reality the students' true scores on a hypothetically perfect test were as follows: 360, 375, 430, 465, 470, 500, 550, 610, 590, and 670, shown at the bottom of Figure 14.1. Although not perfect, the test came close to representing true scores; only two people (H and I) were improperly ordered. This example illustrates a measure whose isomorphism with reality is high but improvable.

Researchers work with fallible measures. Instruments that measure psychosocial phenomena are less likely to correspond to reality than physical measures, but few instruments are error free.

Advantages of Measurement

What exactly does measurement accomplish? Consider how handicapped healthcare professionals would be in the absence of measurement. What would happen, for example, if there were no measures of blood pressure or temperature? Subjective evaluations of clinical outcomes would have to be used. A principal strength of measurement is that it removes subjectivity and guesswork. Because measurement is based on explicit rules, resulting information tends to be objective—that is, it can be independently verified. Two people measuring the weight of a person using the same scale would likely get identical results. Most measures incorporate mechanisms for minimizing subjectivity.

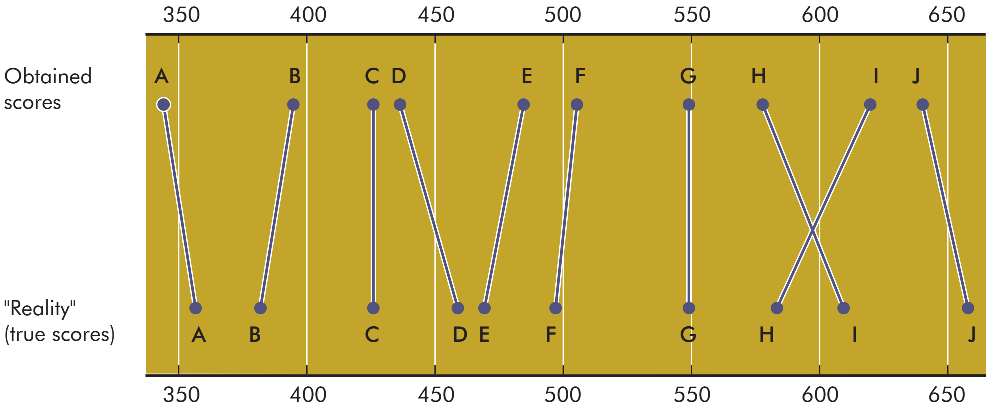

FIGURE 14.1 Relationship between obtained and true scores for a hypothetical set of test scores.

Measurement also makes it possible to obtain reasonably precise information. Instead of describing Nathan as "rather tall," we can depict him as being 6 feet 3 inches tall. With precise measures, researchers can differentiate among people with different degrees of an attribute.

Finally, measurement is a language of communication. Numbers are less vague than words and can communicate information more accurately. If a researcher reported that the average oral temperature of a sample of patients was "somewhat high," different readers might make different inferences about the sample's physiologic state. However, if the researcher reported an average temperature of 99.6°F, there would be no ambiguity.

Errors of Measurement

Procedures for obtaining measurements, as well as the objects being measured, are susceptible to influences that can alter the resulting data. Some influences can be controlled to a certain degree, and attempts should be made to do so, but such efforts are rarely completely successful.

Instruments that are not perfectly accurate yield measurements containing some error. Within classical measurement theory, an **observed** (or **obtained**) **score** can be conceptualized as having two parts—an error component and a true component. This can be written symbolically as follows:

$$\text{Obtained score} = \text{True score} \pm \text{Error}$$

or

$$X_O = X_T \pm X_E$$

The first term in the equation is an observed score—for example, a score on an anxiety scale. X_T is the value that would be obtained with an infallible measure. The **true score** is hypothetical—it can never be known because measures are *not* infallible. The final term is the **error of measurement**. The difference between true and obtained scores is the result of factors that distort the measurement.

Decomposing obtained scores in this manner highlights an important point. When researchers measure an attribute, they are also *measuring* attributes that are not of interest. The true score component is what they hope to isolate; the error component is a composite of other factors that are also being measured, contrary to their wishes. This concept can be illustrated with an exaggerated example. Suppose a researcher measured the weight of 10 people on a spring scale. As participants step on the scale, the researcher places a hand on their shoulders and applies pressure. The resulting measures (the X_Os) will be biased upward because scores reflect both actual weight (X_T) and pressure (X_E). Errors of measurement are problematic because their value is unknown and also because they often are variable. In this example, the amount of pressure applied likely would vary from one person to the next. In other words, the proportion of true score component in an obtained score varies from one person to the next.

Many factors contribute to errors of measurement. Some errors are random while others are systematic, reflecting *bias*. Common influences on measurement error include the following:

1. *Situational contaminants*. Scores can be affected by the conditions under which they are produced. A participant's awareness of an observer's presence (reactivity) is one source of bias. Environmental factors, such as temperature, lighting, and time of day, are potential sources of measurement error.
2. *Transitory personal factors*. A person's score can be influenced by such personal states as fatigue or mood. In some cases, such factors directly affect the measurement, as when anxiety affects pulse rate measurement. In other cases, personal factors alter scores by influencing people's motivation to cooperate, act naturally, or do their best.
3. *Response-set biases*. Relatively enduring characteristics of people can interfere with accurate measurements. Response sets such as social desirability or acquiescence are potential biases in self-report measures, particularly in psychological scales (Chapter 13).
4. *Administration variations*. Alterations in the methods of collecting data from one person to the next can result in score variations unrelated

to variations in the target attribute. For example, if some physiologic measures are taken before a feeding and others are taken after a feeding, then measurement errors can potentially occur.

5. *Instrument clarity*. If the directions on an instrument are poorly understood, then scores may be affected. For example, questions in a self-report instrument may be interpreted differently by different respondents, leading to a distorted measure of the variable.
6. *Item sampling*. Errors can be introduced as a result of the sampling of items used in the measure. For example, a nursing student's score on a 100-item test of critical care nursing knowledge will be influenced by *which* 100 questions are included. A person might get 95 questions correct on one test but only 92 right on another similar test.
7. *Instrument format*. Technical characteristics of an instrument can influence measurements. For example, the ordering of questions in an instrument may influence responses.

TIP: The Toolkit section of Chapter 14 of the *Resource Manual* includes a list of suggestions for enhancing data quality and minimizing measurement error in quantitative studies.

RELIABILITY OF MEASURING INSTRUMENTS

The reliability of a quantitative instrument is a major criterion for assessing its quality. An instrument's **reliability** is the consistency with which it measures the target attribute. If a scale weighed a person at 120 pounds one minute and 150 pounds the next, it would be unreliable. The less variation an instrument produces in repeated measurements, the higher its reliability. Thus, reliability can be equated with a measure's stability, consistency, or dependability.

Reliability also concerns accuracy. An instrument is reliable to the extent that its measures reflect true scores—that is, to the extent that measurement errors are absent from obtained scores. Reliable measures maximize the true score component and minimize error.

These two ways of explaining reliability (consistency and accuracy) are not so different as they might appear. Errors of measurement that impinge on an instrument's accuracy also affect its consistency. The example of the scale with variable weight readings illustrates this point. Suppose that the true weight of a person is 125 pounds, but that two independent measurements yielded 120 and 150 pounds. In terms of the equation presented in the previous section, we could express the measurements as follows:

$$120 = 125 - 5$$

$$150 = 125 + 25$$

The errors of measurement for the two trials (–5 and +25, respectively) resulted in scores that are inconsistent *and* inaccurate.

The reliability of an instrument can be assessed in various ways, and the appropriate method depends on the nature of the instrument and on the aspect of reliability of greatest concern. Three key aspects are stability, internal consistency, and equivalence.

Stability

The **stability** of an instrument is the extent to which similar scores are obtained on separate occasions. The reliability estimate focuses on the instrument's susceptibility to extraneous influences over time, such as participant fatigue.

Assessments of stability involve procedures that evaluate **test–retest reliability**. Researchers administer the same measure to a sample twice and then compare the scores. The comparison is performed objectively by computing a **reliability coefficient**, which is an index of the magnitude of the test's reliability.

To explain reliability coefficients, we must discuss a statistic called a **correlation coefficient**. We have pointed out that researchers seek to detect and explain relationships among phenomena. For example, is there a relationship between patients' gastric acidity levels and degree of stress? The correlation coefficient is a tool for quantitatively describing the magnitude and direction of a relationship between two variables. The computation of this index does not concern us here. It is more important to understand how to read a correlation coefficient.

Two variables that are obviously related are people's height and weight. Tall people tend to be heavier than short people. We would say that there was a **perfect relationship** if the tallest person in a population were the heaviest, the second tallest person were the second heaviest, and so forth. Correlation coefficients summarize how perfect a relationship is. The possible values for a correlation coefficient range from –1.00 through .00 to +1.00. If height and weight were perfectly correlated, the correlation coefficient expressing this relationship would be 1.00. Because the relationship exists but is not perfect, the correlation coefficient is in the vicinity of .50 or .60. The relationship between height and weight can be described as a **positive relationship** because *increases* in height tend to be associated with *increases* in weight.

When two variables are totally unrelated, the correlation coefficient equals zero. One might expect that women's dress sizes are unrelated to their intelligence. Large women are as likely to perform well on IQ tests as small women. The correlation coefficient summarizing such a relationship would presumably be in the vicinity of .00.

Correlation coefficients running from .00 to –1.00 express **inverse** or **negative relationships**. When two variables are inversely related, increases in one variable are associated with *decreases* in the second variable. Suppose that there is an inverse relationship between people's age and the amount of sleep they get. This means that, on average, the older the person, the fewer the hours of sleep. If the relationship were perfect (e.g., if the oldest person in a population got the least sleep, and so on), the correlation coefficient would be –1.00. In actuality, the relationship between age and sleep is probably modest—in the vicinity of –.15 or –.20. A correlation coefficient of this magnitude describes a weak relationship: older people *tend* to sleep fewer hours and younger people *tend* to sleep more, but nevertheless some younger people sleep few hours, and some older people sleep a lot.

Now, we can discuss the use of correlation coefficients to compute reliability estimates. With test–retest reliability, an instrument is administered twice to the same people. Suppose we wanted to assess the stability of a self-esteem scale. Self-esteem is a fairly stable attribute that does not fluctuate much from day to day, so we would expect a reliable measure of it to yield consistent scores on two occasions. To check the instrument's stability, we administer the scale 2 weeks apart to 10 people. Fictitious data for this example are presented in Table 14.1. It can be seen that, in general, differences in scores on the two testings are not large. The reliability coefficient for test–retest estimates is the correlation coefficient between the two sets of scores. In this example, the reliability coefficient is .95, which is high.

The value of the reliability coefficient theoretically can range between –1.00 and +1.00, like other correlation coefficients. A negative coefficient would have been obtained in our example if those with high self-esteem scores at time 1 had low scores at time 2, and vice versa. In practice, reliability coefficients usually range between .00 and 1.00. The higher the coefficient, the more stable the

TABLE 14.1 Fictitious Data for Test–Retest Reliability of Self-Esteem Scale

PARTICIPANT NUMBER	TIME 1	TIME 2	
1	55	57	
2	49	46	
3	78	74	
4	37	35	
5	44	46	
6	50	56	
7	58	55	
8	62	66	
9	48	50	
10	67	63	$r = .95$

measure. Reliability coefficients above .80 usually are considered good.

The test–retest method is easy, and can be used with self-report, observational, and physiologic measures. Yet, this approach has certain disadvantages. One issue is that many traits *do* change over time, independently of the measure's stability. Attitudes, knowledge, perceptions, and so on can be modified by experiences between testings. Test–retest procedures confound changes from measurement error with true changes in the attribute. Still, there are many relatively enduring attributes for which a test–retest approach is suitable.

Stability estimates suffer from other problems, however. One possibility is that people's responses (or observers' coding) on the second administration will be influenced by their memory of initial responses, regardless of the actual values the second day. Such memory interference results in spuriously high reliability coefficients. Another difficulty is that people may actually change *as a result of* the first administration. Finally, people may not be as careful using the same instrument a second time. If they find the process boring on the second occasion, then responses could be haphazard, resulting in a spuriously low estimate of stability.

On the whole, reliability coefficients tend to be higher for short-term retests than for long-term retests (those greater than 1 month) because of actual changes in the attribute being measured. Stability indexes are most appropriate for relatively stable characteristics such as personality, abilities, or certain physical attributes such as adult height.

It might be noted that while most test–retest efforts involve the calculation of a standard correlation coefficient, as just described, other methods are sometimes used. For example, Yen and Lo (2002) describe how an *intraclass correlation* (ICC) approach offers advantages because of the ability of this index to detect systematic error.

Example of test–retest reliability: Kao and Lynn (2009) developed the Family Caregiver Medication Administration Hassles Scale for use with Mexican American family caregivers of older relatives. The 3-week test–retest reliability for the scale was .64.

Internal Consistency

Scales and tests that involve summing item scores are typically evaluated for their internal consistency. Scales designed to measure an attribute ideally are composed of items that measure that attribute and nothing else. On a scale to measure nurses' empathy, it would be inappropriate to include an item that measures diagnostic competence. An instrument may be said to be **internally consistent** or *homogeneous* to the extent that its items measure the same trait.

Internal consistency reliability is the most widely used reliability approach. Its popularity reflects the fact that it is economical (it requires only one administration) and is the best means of assessing an especially important source of measurement error in psychosocial instruments, the sampling of items.

TIP: Many scales contain multiple **subscales**, each of which taps distinct but related concepts (e.g., a measure of fatigue might include subscales for mental and physical fatigue). The internal consistency of each subscale should be assessed. If subscale scores are summed for a total score, the scale's overall internal consistency is also computed.

The most widely used method for evaluating internal consistency is **coefficient alpha** (or **Cronbach's alpha**). Coefficient alpha can be interpreted like other reliability coefficients: the normal range of values is between .00 and +1.00, and higher values reflect higher internal consistency. It is beyond the scope of this text to explain this method in detail, but information is available in psychometric textbooks (e.g., Nunnally & Bernstein, 1994; Waltz, et al. 2010). Most statistical software can be used to calculate alpha. The research example at the end of Chapter 15 presents some computer output for a reliability analysis.

In summary, coefficient alpha is an index of internal consistency to estimate the extent to which different subparts of an instrument (i.e., items) are reliably measuring the critical attribute. Cronbach's alpha does not, however, evaluate fluctuations over time as a source of unreliability.

Example of internal consistency reliability: Villanueva and colleagues (2009) developed and evaluated a scale to measure nonpsychiatric healthcare providers' attitudes toward pediatric patients with mental illness. The 18-item scale had good internal consistency, alpha = .85.

Equivalence

Equivalence, in the context of reliability assessment, primarily concerns the degree to which two or more independent observers or coders agree about scoring. If there is a high level of agreement, then the assumption is that measurement errors have been minimized. Nurse researchers are especially likely to use this approach with observational measures, although it can be used in other applications—for example, for evaluating the consistency of coding open-ended questions or the accuracy of extracting data from records.

The reliability of ratings and classifications can be enhanced by careful training and the specification of clearly defined, nonoverlapping categories. Even when such care is taken, researchers should assess the reliability of observational instruments and coding systems. In this case, "instrument" includes both the category or rating system *and* the observers or coders making the measurements.

Interrater (or *interobserver*) **reliability** can be assessed using various approaches, which can be categorized as consensus, consistency, and measurement approaches (Stemler, 2004). Many interrater reliability indexes used by nurse researchers are of the consensus type, in which the goal is to have observers share a common interpretation of a construct, and to reach consensus (exact agreement). Consensus measures of interrater reliability for observational coding involve having two or more trained observers watching an event simultaneously, and independently recording data. The data are then used to compute an index of agreement between observers. (For coders, information would be independently coded into categories and then intercoder agreement would be assessed.) When ratings are dichotomous, one procedure is to calculate the proportion of agreements, using the following equation:

$$\frac{\text{Number of agreement}}{\text{Number of agreement + disagreements}}$$

This formula unfortunately tends to overestimate agreements because it fails to account for agreement by chance. If a behavior being observed were coded for absence versus presence, the observers would agree 50% of the time by chance alone. A widely used statistic in this situation is Cohen's **kappa**, which adjusts for chance agreements. Different standards have been proposed for acceptable levels of kappa, but there is some agreement that a value of .60 is minimally acceptable, and that values of .75 or higher are very good.

For certain types of data (e.g., ratings on a multipoint scale), correlation techniques are suitable, and these typically capture consistency rather than consensus. For example, a correlation coefficient can be computed to demonstrate the strength of the relationship between one rater's scores and another's. The **intraclass correlation coefficient** (ICC) can also be used to assess interrater reliability (Shrout & Fleiss, 1979).

Example of interrater reliability: Voepel-Lewis and colleagues (2010) assessed the FLACC Behavioral Scale, an observational tool to assess pain in critically ill patients. Exact agreement, kappa values, and intraclass correlation coefficients suggested strong interrater reliability of the measure.

Interpretation of Reliability Coefficients

Reliability coefficients are important indicators of an instrument's quality. Unreliable measures reduce statistical power and hence affect statistical conclusion validity. If data fail to support a hypothesis, one possibility is that the instruments were unreliable—not necessarily that the expected relationships do not exist. Knowing an instrument's reliability thus is critical in interpreting research results, especially if hypotheses are not supported.

For group-level comparisons, coefficients in the vicinity of .70 may be adequate (especially for subscales), but coefficients of .80 or greater are highly desirable. By group-level comparisons, we mean that researchers compare scores of groups, such as male versus female or experimental versus control participants. The reliability coefficients for measures used for making decisions about individuals ideally should be .90 or better. For instance, if a test score was used as a criterion for admission to a nursing program, then the test's accuracy would be of critical importance to both the applicants and the school of nursing.

Reliability coefficients have a special interpretation that relates to our discussion of decomposing observed scores into error and true score components. Suppose we administered a scale that measures hopefulness to 50 patients with cancer. The scores would vary from one person to another—that is, some people would be more hopeful than others. Some variability in scores is true variability, reflecting real individual differences in hopefulness; some variability, however, is error. Thus,

$$\mathrm{V_O} = \mathrm{V_T} + \mathrm{V_E}$$

where $\mathrm{V_O}$ = observed total variability in scores
$\mathrm{V_T}$ = true variability
$\mathrm{V_E}$ = variability owing to errors

A reliability coefficient is directly associated with this equation. *Reliability is the proportion of true variability to the total obtained variability,* or

$$r = \frac{V_\mathrm{T}}{V_\mathrm{O}}$$

If, for example, the reliability coefficient were .85, then 85% of the variability in obtained scores would represent true individual differences, and 15% of the variability would reflect extraneous fluctuations. Looked at in this way, it should be clear why instruments with reliability lower than .70 are risky to use.

Factors Affecting Reliability

Various things affect an instrument's reliability, and these factors are useful to keep in mind in selecting an instrument. First, the reliability of composite self-report and observational scales is partly a function of their length (i.e., number of items). To improve reliability, more items tapping the same concept should be added. Items that have no discriminating power (i.e., that elicit similar responses from everyone) should, however, be removed. Item analysis procedures for guiding decisions about item retention, modification, or deletion are outlined in Chapter 15.

With observational scales, reliability can be improved by greater precision in defining categories, or greater clarity in explaining the underlying construct for rating scales. The best means of enhancing reliability in observational studies, however, is thorough observer training.

An instrument's reliability is related in part to the heterogeneity of the sample with which it is used. The more homogeneous the sample (i.e., the more similar their scores), the lower the reliability coefficient will be. This is because instruments are designed to measure differences among those being measured. If the sample is homogeneous, then it is more difficult for the instrument to discriminate reliably among those who possess varying degrees of the attribute. For example, a depression scale will be less reliable when administered to a homeless sample than when it is used with a general population.

An instrument's reliability is not a fixed entity. *The reliability of an instrument is a property not of the instrument but rather of the instrument when administered to certain people under certain conditions*. A scale that reliably measures dependence in hospitalized adults may be unreliable with nursing homes residents. This means that in selecting an instrument, it is important to know the characteristics of the group with which it was developed. If the group is similar to the population for a new study, then the reliability estimate calculated by the scale developer is probably a reasonably good index of the instrument's accuracy in the new research.

TIP: You should not be satisfied with an instrument that will *probably* be reliable in your study. The recommended procedure is to compute new estimates of reliability whenever research data are collected.

Finally, reliability estimates vary according to the procedures used to obtain them. A scale's test–retest reliability is rarely the same value as its internal consistency reliability. In selecting an instrument, researchers need to determine which aspect of reliability (stability, internal consistency, or equivalence) is relevant.

Example of different reliability estimates: Schilling and colleagues (2009) developed a scale to measure self-management of type I diabetes among adolescents. They evaluated the scale's reliability using test–retest and internal consistency approaches. As an example of their findings, the coefficient alpha for the 7-item Goals subscale was .75. The subscale's test–retest reliability was .60 at 2 weeks and .59 at 3 months.

VALIDITY

A second key criterion for evaluating an instrument is its validity. **Validity** is the degree to which an instrument measures what it is supposed to measure. When researchers develop an instrument to measure hopelessness, they need to be sure that resulting scores validly reflect this construct and not something else, like depression.

Reliability and validity are not independent qualities of an instrument. *A measuring device that is unreliable cannot be valid.* An instrument cannot validly measure an attribute if it is inconsistent and inaccurate. An unreliable instrument contains too much error to be a valid indicator of the target variable. An instrument can, however, be reliable without being valid. Suppose we had the idea to assess patients' anxiety by measuring their height. We could obtain highly accurate, consistent measurements of their height, but such measures would not be valid indicators of anxiety. Thus, the high reliability of an instrument provides no evidence of its validity; low reliability *is* evidence of low validity.

Like reliability, validity has different aspects and assessment approaches, but unlike reliability, an instrument's validity is difficult to evaluate. There are no equations that can easily be applied to the scores of a hopelessness scale to estimate how good a job the scale is doing in measuring the critical variable. Validation is an evidence-building enterprise, in which the goal is to assemble sufficient evidence from which validity can be inferred. The greater the amount of evidence supporting validity, the more sound the inference.

TIP: Instrument developers usually gather evidence of the validity and reliability of their instrument in a **psychometric assessment** before making the instrument available for general use. If you use an existing instrument, choose one with demonstrated high reliability and validity.

Face Validity

Face validity refers to whether the instrument *looks* like it is measuring the target construct. Although face validity is not considered strong evidence of validity, it is helpful for a measure to have face validity if other types of validity have also been demonstrated. It might be easier to persuade people to participate in a study if the instruments have face validity, for example.

Example of face validity: Jones and colleagues (2008) developed the Stroke Self-Efficacy Questionnaire for use by practitioners working in stroke care. Face validity was addressed through consultation with experts in stroke rehabilitation and self-efficacy theory, as well as with stroke survivors.

Content Validity

Content validity concerns the degree to which an instrument has an appropriate sample of items for the construct being measured and adequately covers the construct domain. Content validity is relevant for both affective measures (i.e., measures of psychological traits) and cognitive measures.

For cognitive measures, the content validity question is, how representative are the test questions of the universe of questions on this topic? For example, suppose we were testing students' knowledge about major nursing theories. The test would not be content valid if it omitted questions about, for example, Orem's Self-Care Theory.

Content validity is also relevant in developing affective measures. Researchers designing a new instrument should begin with a thorough conceptualization of the construct so the instrument can capture the full content domain. Such a conceptualization might come from a variety of sources, including rich first-hand knowledge, an exhaustive literature review, consultation with experts, or findings from a qualitative inquiry.

Example of using qualitative data to enhance content validity: Williams and Kristjanson (2009) developed a scale to measure hospitalized patients' perceptions of the emotional care they experienced. The items were based on the themes identified in a grounded theory study, which explored characteristics of interpersonal interactions patients perceived to be therapeutic.

An instrument's content validity is necessarily based on judgment. There are no completely objective methods of ensuring adequate content coverage on an instrument, but it is common to use a panel of experts to evaluate the content validity of new instruments.

There are various approaches to assessing content validity using an expert panel, but nurse researchers have been in the forefront in developing approaches that involve the calculation of a **content validity index (CVI)**. The experts are asked to evaluate individual items on the new measure as well as the overall instrument. Two key issues in such an evaluation are whether individual items are relevant and appropriate in terms of the construct, and whether the items taken together adequately measure all dimensions of the construct.

At the item level, a common procedure is to have experts rate items on a four-point scale of relevance. There are several variations of labeling the 4 points, but the scale used most often is as follows: 1 = *not relevant,* 2 = *somewhat relevant,* 3 = *quite relevant,* 4 = *highly relevant.* Then, for each item, the **item CVI (I-CVI)** is computed as the number of experts giving a rating of 3 or 4, divided by the number of experts—that is, the proportion in agreement about relevance. For example, an item rated as "quite" or "highly" relevant by 4 out of 5 judges would have an I-CVI of .80, which is considered an acceptable value.

There are two approaches to calculating **scale CVIs (S-CVIs)**, and unfortunately, instrument development papers seldom indicate which approach was used (Polit & Beck, 2006). One approach is to calculate the percentage of items on the scale for which *all* judges agreed on content validity. In other words, if a 10-item scale had 6 items for which the I-CVIs were 1.00, then the S-CVI would be .60. We call this the S-CVI/UA (universal agreement) approach. Because disagreements (as well as agreements) can occur by chance, and because disagreements could reflect bias or misunderstanding, we find this approach too stringent.

A second method is to compute the S-CVI by averaging I-CVIs. We recommend the averaging approach, which we refer to as S-CVI/Ave, and suggest a value of .90 as the standard for establishing excellent content validity (Polit & Beck, 2006). Content validation should be done with at least 3 experts, but a larger group is preferable. Further guidance is offered in Chapter 15.

Example of using a content validity index: Chien and Chan (2009) tested the Chinese version of the Level of Expressed Emotion Scale, a scale used with families of people with schizophrenia. The item-level CVIs ranged from .86 to 1.00 and the scale-level CVI, using the averaging approach, was .993.

Criterion-Related Validity

An instrument is said to have **criterion-related validity** if its scores correlate highly with scores on an external criterion. For example, if scores on a scale of attitudes toward premarital sex correlate highly with subsequent loss of virginity in a sample of teenagers, then the attitude scale would have good validity. For criterion-related validity, the key issue is whether the instrument is a useful predictor of other behaviors, experiences, or conditions.

A requirement of this approach is the availability of a reliable and valid criterion with which measures on the instrument can be compared. This is, unfortunately, seldom easy. If we were developing an instrument to measure nursing students' clinical skills, we might use supervisory ratings as our criterion—but can we be sure that these ratings are valid and reliable? The ratings might themselves need validation. Criterion-related validity is most appropriate when there is a concrete, reliable criterion. For example, a scale to measure smokers' motivation to quit smoking has a clear-cut, objective criterion: subsequent smoking.

Once a criterion is selected, a criterion-related **validity coefficient** can be computed by correlating scores on the instrument and the criterion. The magnitude of the coefficient is a direct estimate of how valid the instrument is, according to this validation method. To illustrate, suppose we developed a scale to measure nurses' professionalism. We administer the instrument to a sample of nurses and also ask the nurses to indicate how many professional conferences they have attended. The conference variable was chosen as one of many potential objective criteria of professionalism. Fictitious data are presented in Table 14.2. The correlation coefficient of .83 indicates that the professionalism scale correlates fairly well with the number of conferences attended. Whether the scale is really measuring professionalism is a different issue—an issue that is a construct validation concern discussed in the next section.

A distinction is sometimes made between two types of criterion-related validity. **Predictive validity** refers to the adequacy of an instrument in differentiating between people's performance on a future criterion. When a school of nursing correlates incoming students' high school grades with subsequent grade-point averages, the predictive validity of the high school grades for nursing school performance is being evaluated.

Example of predictive validity: Chang and colleagues (2009) developed and tested the Chinese version of the Positive and Negative Suicide Ideation Inventory. To assess predictive validity, a subsample of students used in the original instrument development study was recruited 1 year later to see if scores on the scale were predictive of recent suicide attempts.

Concurrent validity reflects an instrument's ability to distinguish individuals who differ on a present criterion. For example, a psychological test to differentiate between patients in a mental institution who can and cannot be released could be correlated with current behavioral ratings of healthcare

TABLE 14.2 Fictitious Data for Criterion-Related Validity Example

PARTICIPANT	SCORE ON PROFESSIONALISM SCALE	NUMBER OF NURSING CONFERENCES	
1	25	2	
2	30	4	
3	17	0	
4	20	1	
5	22	0	
6	27	2	
7	29	5	
8	19	1	
9	28	3	
10	15	1	$r = .83$

personnel. The difference between predictive and concurrent validity, then, is the difference in the timing of obtaining measurements on a criterion.

Example of concurrent validity: Cha and colleagues (2008) assessed the concurrent validity of a condom self-efficacy scale in Korean college students by correlating scores on the scale with actual condom use.

Criterion-related validation is most often used in practically oriented research. Criterion-related validity is helpful in assisting decision makers by giving them some assurance that their decisions will be effective, fair, and, in short, valid.

Construct Validity

Construct validity is a key criterion for assessing the quality of a study. As noted in Chapter 10, construct validity concerns inferences from study particulars (such as measures used to operationalize variables) to higher-order constructs. The key construct validity question in measurement is: What is this instrument *really* measuring? Unfortunately, the more abstract the concept, the more difficult it is to establish construct validity; at the same time, the more abstract the concept, the less suitable it is to rely on criterion-related validity. It is really not just a question of suitability, but feasibility. What objective criterion is there for such concepts as empathy or separation anxiety?

Construct validation of an instrument is a challenging but vital task. Construct validation is a hypothesis-testing endeavor, typically linked to a theoretical perspective about the construct. In validating a measure of death anxiety, its relationship to a criterion would be less informative than its correspondence to a cogent conceptualization of death anxiety. Construct validation can be approached in several ways, but it always involves logical analysis and hypothesis tests. Constructs are explicated in terms of other abstract concepts; researchers develop hypotheses about the manner in which the target construct functions in relation to other constructs.

There are a number of ways to gather evidence about construct validity, which we discuss in this section. It should also be noted, however, that if an instrument developer has taken strong steps to ensure the content validity of the instrument, construct validity will also be strengthened.

Known Groups

One construct validation approach is the **known-groups technique**, which yields evidence of **contrast validity**. In this procedure, the instrument is administered to groups hypothesized to differ on the critical attribute because of a known characteristic. For instance, in validating a measure of fear of childbirth, we could contrast the scores of primiparas and multiparas. We would expect that women who had never given birth would be more anxious than women who had done so, and so we might question the instrument's validity if such differences did not emerge. We would not necessarily expect large differences; some primiparas would feel little anxiety, and some multiparas would express fears. We would, however, hypothesize differences in *average* group scores.

Example of the known-groups technique: Gozum and Hacihasanoglu (2009) did a psychometric assessment of the Turkish version of the Medication Adherence Self-Efficacy Scale with a sample of hypertensive patients. Using the known-groups approach, they compared scale scores for those with controlled versus uncontrolled blood pressure.

Hypothesized Relationships

A similar method of construct validation involves testing hypothesized relationships, often on the basis of theory or prior research. This is really a variant of the known-groups approach, which involves hypotheses about the relationship between the measure of the construct and a variable representing group membership. A researcher might reason as follows:

- According to theory, construct X is positively related to construct Y.
- Instrument A is a measure of construct X; instrument B is a measure of construct Y.

- Scores on A and B are correlated positively, as predicted.
- Therefore, it is inferred that A and B are valid measures of X and Y.

This logical analysis does not constitute proof of construct validity, but yields important evidence. Construct validation is essentially an ongoing evidence-building enterprise.

Example of testing relationships: Simmons and colleagues (2009) developed and tested a scale to measure psychological adjustment in patients with an ostomy. In the construct validation efforts, they hypothesized that adjustment scores would be positively correlated with time elapsed since surgery and with scores on an acceptance of illness scale, and their hypotheses were supported.

Convergent and Discriminant Validity

The **multitrait–multimethod matrix method** (**MTMM**) is a significant construct validation tool (Campbell & Fiske, 1959). This procedure involves the concepts of convergence and discriminability. **Convergence** is evidence that different methods of measuring a construct yield similar results. Different measurement approaches should converge on the construct. **Discriminability** is the ability to differentiate the construct from other similar constructs. Campbell and Fiske argued that evidence of both convergence and discriminability should be brought to bear in construct validation.

To help explain the MTMM approach, fictitious data from a study to validate a "need for autonomy" measure are presented in Table 14.3. In using this approach, researchers must measure the critical concept by two or more methods. Suppose we measured need for autonomy in nursing home residents by (1) giving a sample of residents a self-report scale (the measure we are attempting to validate), (2) asking nurses to rate residents after observing them in a task designed to elicit autonomy or dependence, and (3) having residents react to a pictorial stimulus depicting an autonomy-relevant situation (a so-called *projective* measure).

A second requirement of the full MTMM is to measure a differentiating construct, using the same measuring methods. In the current example, suppose we wanted to differentiate "need for autonomy" from "need for affiliation." The discriminant concept must be similar to the focal concept, as in our example: We would expect that people with high need for autonomy would tend to be relatively low on need for affiliation. The point of including both concepts in a single validation study is to gather evidence

TABLE 14.3 Multitrait–Multimethod Matrix

		SELF-REPORT (1)		OBSERVATION (2)		PROJECTIVE (3)	
METHOD	**TRAITS**	**AUT$_1$**	**AFF$_1$**	**AUT$_2$**	**AFF$_2$**	**AUT$_3$**	**AFF$_3$**
Self-report (1)	AUT$_1$	(.88)					
	AFF$_1$	−.38	(.86)				
Observation (2)	AUT$_2$	.60	−.19	(.79)			
	AFF$_2$	−.21	.58	−.39	(.80)		
Projective (3)	AUT$_3$	.51	−.18	.55	−.12	(.74)	
	AFF$_3$	−.14	.49	−.17	.54	−.32	(.72)

AUT = need for autonomy trait; AFF = need for affiliation trait.

that the two concepts are distinct, rather than two different labels for the same underlying attribute.

The numbers in Table 14.3 represent correlation coefficients between scores on six measures (two traits × three methods). For instance, the coefficient of –.38 at the intersection of AUT_1–AFF_1 is the correlation between self-report scores on the need for autonomy and need for affiliation measures. Recall that a minus sign before the correlation coefficient signifies an inverse relationship. In this case, the –.38 tells us that there was a slight tendency for people scoring high on the need for autonomy scale to score low on the need for affiliation scale. (The numbers in parentheses along the diagonal of this matrix are the reliability coefficients.)

Various parts of the MTMM matrix have a bearing on construct validity. The most direct evidence (**convergent validity**) comes from the correlations between two different methods measuring the same trait. In the case of AUT_1–AUT_2, the coefficient is .60, which is reasonably high. Convergent validity should be large enough to encourage further scrutiny of the matrix. Second, the convergent validity entries should be higher, in absolute magnitude,* than correlations between measures that have neither method nor trait in common. That is, AUT_1–AUT_2 (.60) should be greater than AUT_2–AFF_1 (–.21) or AUT_1–AFF_2 (–.19), as it is here. This requirement is a minimum one that, if failed, should cause researchers to have serious doubts about the measures. Third, convergent validity coefficients should be greater than coefficients between measures of different traits by a single method. Once again, the matrix in Table 14.3 fulfills this criterion: AUT_1–AUT_2 (.60) and AUT_2–AUT_3 (.55) are higher in absolute value than AUT_1–AFF_1 (–.38), AUT_2–AFF_2 (–.39), and AUT_3–AFF_3 (–.32). The last two requirements provide evidence for **discriminant validity**.

The evidence is seldom as clear-cut as in this contrived example. Indeed, a common problem with MTMM is interpreting the pattern of coefficients. Another issue is that there are no clear-cut criteria for deciding whether MTMM requirements have been met—that is, there are no objective means of assessing the magnitude of similarities and differences within the matrix. The MTMM is nevertheless a valuable tool for exploring construct validity. Researchers sometimes decide to use MMTM concepts even when the full model is not feasible, as in focusing only on convergent validity.

Example of convergent and discriminant validity: Morea and colleagues (2008) developed and tested the Illness Self-Concept Scale, an instrument designed to predict adjustment in fibromyalgia. Their analyses provided some evidence that their construct, illness self-concept, is distinct from other similar constructs like depression (discriminant validity) and various analyses also supported evidence of convergent validity.

Factor Analysis

Another approach to construct validation uses a statistical procedure called factor analysis. Although factor analysis, which is discussed in Chapter 15, is computationally complex, it is conceptually rather simple. **Factor analysis** is a method for identifying clusters of related variables—that is, dimensions underlying a broad construct. Each dimension, or **factor**, represents a relatively unitary attribute. The procedure is used to identify and group together different items measuring an underlying attribute. In effect, factor analysis constitutes another means of testing hypotheses about the interrelationships among variables, and for looking at the convergent and discriminant validity of a large set of items. Indeed, a procedure known as **confirmatory factor analysis** (CFA) is sometimes used as a method for analyzing MTMM data (Ferketich, et al., 1991; Lowe & Ryan-Wenger, 1992).

Example of factor analysis in construct validation: Zheng and colleagues (2010) developed and tested the Dialysis Patient-Perceived Exercise Benefits and Barriers Scale. Responses to the scale's 24 items by a sample of 269 hemodialysis patients in China were factor analyzed to assess construct validity. Confirmatory factor analysis confirmed a 6-factor structure.

*[**Absolute value** refers to the value without a plus or minus sign. A value of −.80 is of a higher absolute magnitude than +.40.

Interpretation of Validity

Like reliability, validity is not an all-or-nothing characteristic of an instrument. An instrument does not possess or lack validity; it is a question of degree. An instrument's validity is not proved, established, or verified but rather is supported to a greater or lesser extent by evidence.

Strictly speaking, researchers do not validate an instrument but rather an application of it. A measure of anxiety may be valid for presurgical patients on the day of an operation but may not be valid for nursing students on the day of a test. Of course, some instruments may be valid for a wide range of uses with different types of samples, but each use requires new supporting evidence. The more evidence that can be gathered that an instrument is measuring what it is supposed to be measuring, the more confidence researchers will have in its validity.

TIP: When you select an instrument, you should seek evidence of the scale's psychometric soundness by examining the instrument developers' report. However, you also should consider evidence from others who have used the scale. Each time the scale "performs" as hypothesized, this constitutes supplementary evidence for its validity. Conversely, if hypotheses involving the use of the scale are not supported, this suggests potential validity problems (although, of course, other factors may account for nonsupported hypotheses, such as a small sample).

SENSITIVITY, SPECIFICITY, AND LIKELIHOOD RATIOS

Reliability and validity are the two most important criteria for evaluating quantitative instruments, but researchers sometimes need to consider other qualities of an instrument. In particular, sensitivity and specificity are criteria that are important in evaluating instruments used as screening or diagnostic tools (e.g., a scale to measure risk of osteoporosis). Screening/diagnostic instruments can be self-report, observational, or biophysiologic measures.

Sensitivity is the ability of a measure to identify a "case" correctly, that is, to screen in or diagnosis a condition correctly. A measure's sensitivity is its rate of yielding "true positives." **Specificity** is the measure's ability to identify noncases correctly, that is, to screen *out* those without the condition. Specificity is an instrument's rate of yielding "true negatives." To evaluate an instrument's sensitivity and specificity, researchers need a reliable and valid criterion of "caseness" against which scores on the instrument can be assessed.

Calculating Sensitivity, Specificity, and Related Indicators

Suppose we wanted to evaluate whether adolescents' self-reports about their smoking were accurate, and we asked 100 teenagers about whether they had smoked a cigarette in the previous 24 hours. The "gold standard" for nicotine consumption is cotinine levels in a body fluid, so assume that we did a urinary cotinine assay. Some fictitious data are shown in Table 14.4.

Sensitivity, in this example, is calculated as the proportion of teenagers who said they smoked *and* who had high concentrations of cotinine, divided by all real smokers as indicated by the urine test. Put another way, it is the true positives divided by all positives. In this case, there was considerable under-reporting of smoking and so the sensitivity of the self-report was only .50. Specificity is the proportion of teenagers who accurately reported they did not smoke, or the true negatives divided by all negatives. In our example, specificity is .83. There was considerably less over-reporting of smoking ("faking bad") than under-reporting ("faking good"). Sensitivity and specificity are often reported as percentages rather than proportions, by multiplying the proportions by 100.

Often, other related indicators are calculated with such data. **Predictive values** are posterior probabilities—the probability of an outcome after the results are known. A **positive predictive value** (or PPV) is the proportion of people with a positive result who have the target outcome or disease. In our example, the PPV is the proportion of teens who

TABLE 14.4 Example Illustrating Sensitivity, Specificity, and Likelihood Ratios

	URINARY COTININE LEVEL		
SELF-REPORTED SMOKING	**Positive (Cotinine > 200 ng/mL)**	**Negative (Cotinine ≤ 200 ng/mL)**	**Total**
Yes, smoked	A (true positive) 20	B (false positive) 10	A + B 30
No, did not smoke	C (false negative) 20	D (true negative) 50	C + D 70
Total	A + C 40	B + D 60	A + B + C + D 100

Sensitivity = A/(A + C) = .50
Specificity = D/(B + D) = .83
Positive predictive value (PPV) = A/(A + B) = .67
Negative predictive value (NPV) = D/(C = D) = .71
Likelihood ratio—positive (LR+) = sensitivity/(1 − specificity) = 2.99
Likelihood ratio—negative (LR−) = (1 − sensitivity)/specificity = .60

said they smoke who actually *do* smoke, according to the cotinine test results. Two out of three of those who reported smoking had high concentrations of cotinine, and so PPV = .67. A **negative predictive value** (NPV) is the proportion of people who have a negative test result who do not have the target outcome or disease. As shown in Table 14.4, 50 out of the 70 teenagers who reported not smoking actually were nonsmokers, and so NPV in our example is .71.

Example of sensitivity, specificity, and predictive values: Chichero and colleagues (2009) developed a dysphagia screening tool to triage patients at risk of dysphagia on admission to acute hospital wards. Sensitivity was 95% and specificity was 97%. Positive predictive value was 92% and negative predictive value was 98%.

In the medical community, reporting **likelihood ratios** has come into favor because it summarizes the relationship between specificity and sensitivity in a single number. The likelihood ratio addresses the question, "How much more likely are we to find that an indicator is positive among those *with* the outcome of concern compared to those for whom the indicator is negative?" For a positive test result, then, the likelihood ratio (LR+) is the ratio of true-positive results to false-positive results. The formula for LR+ is sensitivity divided by 1 minus specificity. For the data in Table 14.4, LR+ is 2.99: We are about three times as likely to find that a self-report of smoking really *is* for a true smoker than it is for a nonsmoker. For a negative test result, the likelihood ratio (LR−) is the ratio of false-negative results to true-negative results. For the data in Table 14.4, the LR− is .60. In our example, we are about half as likely to find that a self-report of nonsmoking is false than we are to find that it reflects a true nonsmoker. When a test is high on both sensitivity and specificity (which is not especially true in our example), the likelihood ratio is high and discrimination is good.

Example of likelihood ratios: Novotny and Anderson (2008) tested an algorithm for predicting the probability of readmission (Pra) of medical inpatients within 41 days of discharge from the hospital, using hospital records data. Pra score values ranged from .16 to .75. With a Pra value of .45, the likelihood ratio was 1.6.

Receiver Operating Characteristic (ROC) Curves

All of the indicators that we calculated for the data in Table 14.4 are contingent upon the critical value that we established for cotinine concentration. Sensitivity and specificity would be quite different if we had used 100 ng/mL as indicative of smoking status, rather than 200 ng/mL. There is almost invariably a trade-off between the sensitivity and specificity of a measure. When sensitivity is increased to include more true positives, the proportion of true negatives declines. Therefore, a critical task in developing new diagnostic or screening measures is to develop the appropriate **cutoff point** (or *cutpoint*), that is, a score to distinguish cases and noncases.

To identify the best cutoff point, researchers often are guided by a **receiver operating characteristic curve** (**ROC curve**) (Fletcher, et al., 2005). To construct an ROC curve, the sensitivity of an instrument (i.e., the rate of correctly identifying a case vis-à-vis a well-established criterion) is plotted against the false-positive rate (i.e., the rate of incorrectly diagnosing someone as a case, which is the inverse of its specificity) over a range of different scores. The score (cutoff point) that yields the best balance between sensitivity and specificity can then be determined. The optimum cutoff is at or near the shoulder of the ROC curve.

ROC curves can best be explained with an illustration. Figure 14.2 presents an ROC curve from a study in which a goal was to establish cutoff points for scores on the Braden Q scale for predicting pressure ulcer risk in children (Curley et al., 2003). In this figure, sensitivity and one minus specificity are plotted for each possible score of the Braden Q scale. The upper left corner represents sensitivity at its highest possible value (1.0) and false positives at its lowest possible value (.00). Screening instruments that do an excellent job of discriminating

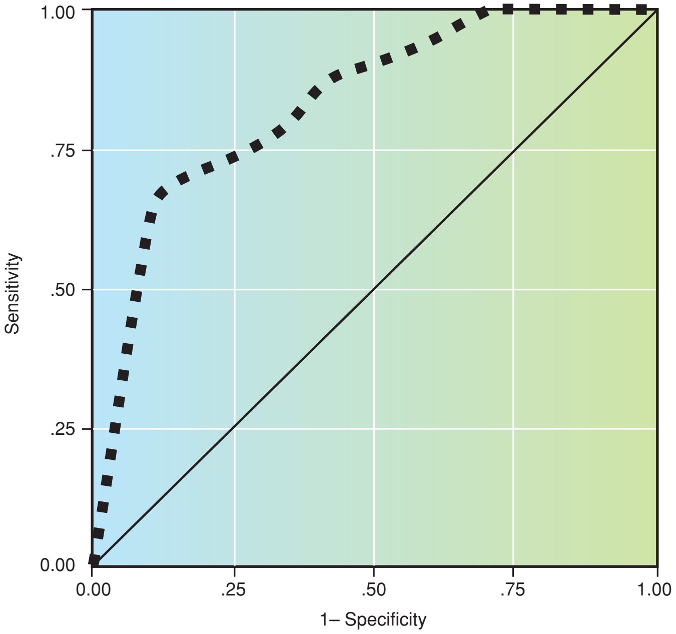

FIGURE 14.2 Receiver operating characteristic (ROC) curve for Braden Q Scale. From Curley, M. A. Q., Razmus, I. S., Roberts, K. E., & Wypij, D. (2003). Predicting pressure ulcer risk in pediatric patients: The Braden Q Scale. *Nursing Research, 52*, p. 27.

have points that crowd close to the upper left corner, which indicates that as sensitivity increases there is relatively little loss in specificity. ROC curves that are closer to a diagonal, from lower left to upper right, are indicative of an instrument with poor discriminatory power.

The overall accuracy of an instrument can be calculated as the proportion of the area under the ROC curve, an index referred to as **area under the curve**, or **AUC**. The larger the area, the more accurate the instrument. The AUC for the data portrayed in Figure 14.2 is .83. The cutoff score in this example was established at 16. At this cutoff value, the sensitivity was .88 and the specificity was .58. The researchers used these preliminary analyses to improve on the Braden Q scale and achieved even better results.

In selecting an appropriate cutoff point, the final decision is likely to be driven by clinical or economic factors and not just statistical ones. The financial and emotional costs of misclassifying people may be greater for false positives than false negatives, or vice versa.

OTHER CRITERIA FOR ASSESSING QUANTITATIVE MEASURES

Although we have already discussed the major criteria that are used to evaluate the quality of measuring instruments, we briefly mention a few others.

Efficiency

Instruments of comparable reliability and validity may differ in their efficiency. A depression scale that requires 5 minutes of people's time is efficient compared with a depression scale that requires 20 minutes to complete. In most studies, efficient instruments are desirable because they reduce participant burden.

One aspect of efficiency is the number of items on the instrument. Long instruments tend to be more reliable than shorter ones, but there is a point of diminishing returns. As an example, consider a 40-item scale to measure social support that has an internal consistency reliability of .94. We can use a formula, known as the **Spearman-Brown formula**, to estimate how reliable the scale would be with fewer items. As an example, if we wanted to shorten the scale to 30 items, the formula would result in an estimated reliability of .92.** Thus, a 25% reduction in the instrument's length resulted in a negligible decrease in reliability, from .94 to .92. Most researchers likely would sacrifice a modest amount of reliability in exchange for reducing response burden and data collection costs. Other things being equal, it is desirable to select as efficient an instrument as possible.

Other Criteria

A few remaining qualities that sometimes are considered in assessing a quantitative instrument can be noted. Most of the following criteria are actually aspects of the reliability and validity:

1. *Comprehensibility*. Participants and researchers should be able to comprehend the behaviors required to secure accurate and valid measures.
2. *Precision*. An instrument should discriminate between people with different amounts of an attribute as precisely as possible.
3. *Range*. The instrument should be capable of achieving a meaningful measure from the smallest expected value of the variable to the largest.
4. *Linearity*. A researcher normally strives to construct measures that are equally accurate and sensitive over the entire range of values.
5. *Reactivity*. The instrument should, insofar as possible, avoid affecting the attribute being measured.

**The equation (and the worked-out example) for this situation is as follows:

$$r^1 = \frac{kr}{1 + [(k - 1)r]} = \frac{.75(.94)}{1 + [(-.25)(.94)]} = .92$$

where k = the factor by which the instrument is being decreased, in this case, $k = 30 \div 40 = .75$; r = reliability for the full scale, here, .94; and r^1 = reliability estimate for the shorter scale.

DATA QUALITY WITH SINGLE INDICATORS

The discussion in this chapter has primarily focused on methods of evaluating data quality for multi-item scales, which are widely used by nurse researchers. Textbooks on research methods or measurement rarely say much about reliability or validity for single questions (e.g., "What is your date of birth?") or single-item scales, such as visual analog scales.

The truth of the matter is that it is not easy to evaluate data quality in such situations. This is of great concern in large national surveys, such as the National Longitudinal Study of Adolescent Health. Population estimates of, say, average number of times adolescents have been hospitalized, or the percentage who have ever used marijuana, are based on reports in response to individual (nonscaled) questions, so the accuracy of the responses is vital. We touch briefly here on data quality assessment for single indicators.

The two basic strategies for estimating measurement error in such situations are a test–retest approach and external verification. In the former, the questions that are of interest are asked on two separate occasions. When this happens for the express purpose of assessing consistency (in what is called a *response variance reinterview*), the second administration typically involves a subsample of respondents and an abbreviated instrument with key questions. Survey researchers compute various statistical indexes (e.g., an *index of inconsistency*) to help them understand and interpret response differences—that is, measurement error—in the two administrations (Subcommittee on Measuring and Reporting the Quality of Survey Data, 2001). Although few nurse researchers would have the resources to undertake such an enterprise, there may be opportunities to use the underlying principle for critical pieces of information. For example, in a self-report instrument, it might be possible to ask the same question twice, early and later, for example, or to ask the question in slightly different ways in the same questionnaire or interview. Also, if a study is longitudinal, factual information (e.g., date of birth) could be gathered twice to assess any discrepancies.

The second approach is to verify information provided in the primary data gathering method against an external source—a form of criterion-related validation. For example, information from a question about birth date could be checked against birth records. Responses to questions about health status, diagnosis, or healthcare could be checked against medical records. Measurement errors are then estimated based on a comparison of the two types of information. It should not necessarily be assumed that records are free of error, but they may be less prone to certain types of bias. Other forms of external verification may be available. In particular, **proxy reports** (obtaining data from another person, such as a family member) might be an option. Patrician (2004) has offered additional guidance regarding single-item scales.

Researchers using biophysiologic measures should also give data quality some thought rather than assuming they will be error free. Instruments may not be properly calibrated, the person doing the tests may not follow laboratory protocols, and laboratory procedures can vary from one lab to the next. Measurement errors can also occur because of patient circumstances, such as insufficient sleep. Moreover, if physiologic measures are taken from charts, the possibility of error should be considered.

CRITIQUING DATA QUALITY IN QUANTITATIVE STUDIES

If data are seriously flawed, the study cannot contribute useful evidence. Therefore, in drawing conclusions about a study's evidence, it is important to consider whether researchers have taken appropriate steps to collect data that accurately reflect reality. Research consumers have the right—indeed, the obligation—to ask: Can I trust the data? Do the data accurately and validly reflect key constructs?

Information about data quality should be provided in every quantitative research report because it is not possible to come to conclusions about the quality of study evidence without such information. Reliability estimates are usually reported because they are

BOX 14.1 Guidelines for Critiquing Data Quality in Quantitative Studies

1. Is there congruence between the research variables as conceptualized (i.e., as discussed in the introduction of the report) and as operationalized (i.e., as described in the method section)?
2. If operational definitions (or scoring procedures) are specified, do they clearly indicate the rules of measurement? Do the rules seem sensible? Were data collected in such a way that measurement errors were minimized?
3. Does the report offer evidence of the reliability of measures? Does the evidence come from the research sample itself, or is it based on other studies? If the latter, is it reasonable to conclude that data quality would be similar for the research sample as for the reliability sample (e.g., are sample characteristics similar)?
4. If reliability is reported, which estimation method was used? Was this method appropriate? Should an alternative or additional method of reliability appraisal have been used? Is the reliability sufficiently high?
5. Does the report offer evidence of the validity of the measures? Does the evidence come from the research sample itself, or is it based on other studies? If the latter, is it reasonable to believe that data quality would be similar for the research sample as for the validity sample (e.g., are the sample characteristics similar)?
6. If validity information is reported, which validity approach was used? Was this method appropriate? Does the validity of the instrument appear to be adequate?
7. If there is no reliability or validity information, what conclusion can you reach about the quality of the data in the study?
8. If a diagnostic or screening tool was used, is information provided about its sensitivity and specificity, and were these qualities adequate?
9. Were the research hypotheses supported? If not, might data quality play a role in the failure to confirm the hypotheses?

easy to communicate. Ideally—especially for composite scales—the report should provide reliability coefficients based on data from the study itself, not just from previous research. Interrater or interobserver reliability is especially crucial for coming to conclusions about data quality in observational studies. The values of the reliability coefficients should be sufficiently high to support confidence in the findings. It is especially important to scrutinize reliability information in studies with nonsignificant findings because the unreliability of measures can undermine statistical conclusion validity.

Validity is more difficult to document in a report than reliability. At a minimum, researchers should defend their choice of existing measures based on validity information from the developers, and they should cite the relevant publication. If a study used a screening or diagnostic measure, information should also be provided about its sensitivity and specificity.

Box 14.1 provides some guidelines for critiquing aspects of data quality of quantitative measures. The guidelines are available in the Toolkit of the accompanying *Resource Manual* for your use and adaptation.

RESEARCH EXAMPLE

In this section, we describe a study that used both self-report and observational measures. We focus on the researchers' excellent documentation of data quality in their study.

Study: Communication and outcomes of visits between older patients and nurse practitioners (Gilbert and Hayes, 2009)

Statement of Purpose: The purpose of this study was to examine relationships among patient–clinician communication, background characteristics of the patients and the clinicians (nurse practitioners or NPs), and both proximal outcomes (e.g., patient satisfaction) and longer-term outcomes (e.g., changes in patients' physical and mental health).

Design: Visits between 31 NPs and 155 patients were video recorded and various aspects of patient and NP behaviors were coded. Proximal outcomes were measured by self-report after the visits. Four weeks later, changes in patients' health outcomes were assessed using self-report measures.

Instruments and Data Quality: Communications during the visits were measured using the Roter Interaction Analysis System (RIAS) for verbal interaction and a checklist for nonverbal behaviors. The Roter system involves coding for both the content of the communication and relationship aspects, using a system of 69 categories for all utterances (only 43 were used in this study). The researchers noted that the predictive validity of the RIAS had considerable support. The average interrater reliability in the present study for the 43 coded behavior categories was .95. For the nonverbal behavior checklist, various actions (e.g., gazes, nods, smiles) were coded in 1-second segments over a 30-second sample. Two coders independently coded all segments and any discrepancies in coding were resolved by a third party. Several variables were measured by patients' self-report, including both 1-item measures (e.g., satisfaction with the visit) and multi-item scales (e.g., physical and mental health). For example, patient satisfaction with the NP visit was measured using one item, previously used in a large national survey, which asked for ratings of perceived quality of care on a 10-point scale from 1 (*worst care possible*) to 10 (*best care possible*). The authors noted that a correlation of .72 between the ratings and the average of several other satisfaction items provided some evidence for the reliability of the single item. Physical and mental health were measured with a 12-item scale called the SF-12 Health Survey, a widely used and well-validated instrument. The test developer had reported results indicating Cronbach alpha values of .89 for physical health and .82 for mental health among people 65 years and older. In the present study, the researchers computed the internal consistency reliability to be .87 and .72 for physical and mental health, respectively.

Key Findings: Among the many findings reported in this study, the researchers found that better patient outcomes were associated with a higher amount of communication content involving seeking and giving biomedical and psychosocial information, and with a relationships component of more positive talk and greater trust and receptivity.

SUMMARY POINTS

- **Measurement** involves assigning numbers to objects to represent the amount of an attribute, using a specified set of rules. Researchers strive to develop or use measurements whose rules are *isomorphic* with reality.
- Few quantitative measuring instruments are infallible. Sources of measurement error include situational contaminants, response-set biases, and transitory personal factors, such as fatigue.
- **Obtained scores** from an instrument consist of a **true score** component (the value that would be obtained for a hypothetical perfect measure of the attribute) and an error component, or **error of measurement**, that represents measurement inaccuracies.
- **Reliability**, one of two primary criteria for assessing an instrument, is the degree of consistency or accuracy with which an instrument measures an attribute. The higher an instrument's reliability, the lower the amount of error in obtained scores.
- There are different methods for assessing an instrument's reliability and for computing a **reliability coefficient**. A reliability coefficient typically is based on the computation of a **correlation coefficient** that indicates the magnitude and direction of a relationship between two variables.
- Correlation coefficients can range from –1.00 (a **perfect negative relationship**) through zero to +1.00 (a **perfect positive relationship**). Reliability coefficients usually range from .00 to 1.00, with higher values reflecting greater reliability.
- The **stability** aspect of reliability, which concerns the extent to which an instrument yields the same results on repeated administrations, is evaluated as **test–retest reliability**.
- The **internal consistency** aspect of reliability—the extent to which all the instrument's items are measuring the same attribute—is usually assessed by **Cronbach's alpha**.
- When the reliability assessment focuses on **equivalence** between observers in rating or coding behaviors, estimates of **interrater** (or

interobserver) **reliability** are obtained. When a consensus measure capturing interrater agreement within a small number of categories is desired, the **kappa** statistic is often used.

- **Reliability coefficients** reflect the proportion of true variability in a set of scores to the total obtained variability.
- **Validity** is the degree to which an instrument measures what it is supposed to measure.
- **Face validity** refers to whether the instrument appears, on the face of it, to be measuring the appropriate construct.
- **Content validity** concerns the sampling adequacy of the content being measured. Expert ratings on the relevance of items can be used to compute **content validity index** (**CVI**) information. **Item CVIs** (**I-CVIs**) represent the proportion of experts rating each item as relevant. A **scale CVI** using the averaging calculation method (S-CVI/Ave) is the average of all I-CVI values.
- **Criterion-related validity** (which includes both **predictive validity** and **concurrent validity**) focuses on the correlation between the instrument and an outside criterion.
- **Construct validity**, an instrument's adequacy in measuring the focal construct, is a hypothesis-testing endeavor. One approach assesses **contrast validity**, using the **known-groups technique** to contrast scores of groups hypothesized to differ on the attribute; another approach is **factor analysis**, a statistical procedure for identifying unitary clusters of items or measures.
- Another construct validity approach is the **multitrait–multimethod (MTMM) matrix technique**, which is based on the concepts of convergence and discriminability. **Convergence** refers to evidence that different methods of measuring the same attribute yield similar results. **Discriminability** refers to the ability to differentiate the construct being measured from other, similar concepts.
- A **psychometric assessment** of a new instrument is usually undertaken to gather evidence about validity, reliability, and other assessment criteria.
- Sensitivity and specificity are important criteria for screening and diagnostic instruments. **Sensitivity** is the instrument's ability to identify a case correctly (i.e., its rate of yielding true positives). **Specificity** is the instrument's ability to identify noncases correctly (i.e., its rate of yielding true negatives). Other related indexes include the measure's **positive predictive value (PPV), negative predictive value (NPV)**, and **likelihood ratios**.
- Sensitivity is sometimes plotted against specificity in a **receiver operating characteristic curve** (**ROC curve**) to determine the optimum **cutoff point** for caseness.

STUDY ACTIVITIES

Chapter 14 of the *Resource Manual for Nursing Research*: *Generating and Assessing Evidence for Nursing Practice*, *9th edition*, offers exercises and study suggestions for reinforcing concepts presented in this chapter. In addition, the following study questions can be addressed:

1. Explain in your own words the meaning of the following correlation coefficients:
   a. The relationship between intelligence and grade-point average was found to be .72.
   b. The correlation coefficient between age and gregariousness was –.20.
   c. It was revealed that patients' compliance with nursing instructions was related to their length of stay in the hospital (r = –.50).
2. Use the critiquing guidelines in Box 14.1 to evaluate data quality in the study by Gilbert and Hayes (2009), referring to the original study if possible.

STUDIES CITED IN CHAPTER 14

Cha, E., Kim, K., & Burke, L. (2008). Psychometric validation of a condom self-efficacy scale in Korean. *Nursing Research, 57*, 245–251.

Chang, H., Lin, C., Chou, K., Ma, W., & Yang, C. (2009). Chinese version of the positive and negative suicide ideation: Instrument development. *Journal of Advanced Nursing, 65*, 1485–1496.

Chichero, J., Heaton, S., & Bassett, L. (2009). Triaging dysphagia: Nurses screening for dysphagia in an acute hospital. *Journal of Clinical Nursing, 18*, 1649–1659.

Chien, W. T., & Chan, S. (2009). Testing the psychometric properties of a Chinese version of the Level of Expressed Emotion Scale. *Research in Nursing & Health, 32*, 59–70.

Curley, M. A. Q., Razmus, I. S., Roberts, K. E., & Wypij, D. (2003). Predicting pressure ulcer risk in pediatric patients. *Nursing Research, 52*, 22–33.

Gilbert, D., & Hayes, E. (2009). Communication and outcomes of visits between older patients and nurse practitioners. *Nursing Research, 58*, 283–293.

Gozum, S., & Hacihasanoglu, R. (2009). Reliability and validity of the Turkish adaptation of Medication Adherence Self-Efficacy Scale in hypertensive patients. *European Journal of Cardiovascular Nursing, 8*, 129–136.

Jones, F., Partridge, C., & Reid, F. (2008). The Stroke Self-Efficacy Questionnaire: Measuring individual confidence in functional performance after stroke. *Journal of Clinical Nursing, 17*, 244–252.

Kao, H., & Lynn, M. (2009). Use of the measurement of medication administration hassles with Mexican American family caregivers. *Journal of Clinical Nursing, 18*, 2596–2603.

Morea, J., Friend, R., & Bennett, R. (2008). Conceptualizing and measuring illness self-concept. A comparison with self-esteem and optimism in predicting fibromyalgia adjustment. *Research in Nursing & Health, 31*, 563–575.

Novotny, N., & Anderson, M. A. (2008). Prediction of early readmission in medical inpatients using the probability of repeated admission instrument. *Nursing Research, 57*, 406–415.

Schilling, L., Dixon, J., Knafl, K., Lynn, M., Murphy, K., Dumser, S., & Grey, M. (2009). A new self-report measure of self-management of type I diabetes for adolescents. *Nursing Research, 58*, 228–236.

Simmons, K., Smith, J., & Maekawa, A. (2009). Development and psychometric evaluation of the Ostomy Adjustment Inventory-23. *Journal of Wound, Ostomy & Continence Nursing, 36*, 69–75.

Williams, A., & Kristjanson, L. (2009). Emotional care experienced by hospitalized patients: Development and testing of a measurement instrument. *Journal of Clinical Nursing, 18*, 1069–1077.

Villanueva, C., Scott, S., Guzzetta, C., & Foster, B. (2009). Development and psychometric testing of the Attitudes toward Mental Illness in Pediatric Patients Scale. *Journal of Child & Adolescent Psychiatric Nursing, 22*, 220–227.

Voepel-Lewis, T., Zanotti, J., Dammeyer, J., & Merkel, S. (2010). Reliability and validity of the Face, Legs, Activity, Cry, Consolability Behavioral Tool in assessing acute pain in critically ill patients. *American Journal of Critical Care, 19*, 55–61.

Zheng, J., You, L., Lou, T., Chen, N., Lai, D., Liang, Y., Li, Y. N., Gu, Y. M., Lv, S. F., & Zhai, C. Q. (2010). Development and psychometric evaluation of the Dialysis Patient-Perceived Exercise Benefits and Barriers Scale. *International Journal of Nursing Studies, 47*, 166–180.

Methodologic and nonresearch references cited in this chapter can be found in a separate section at the end of the book.