



## Assessing Data Quality

**A**n ideal data collection procedure is one that captures a construct in a way that is relevant, credible, accurate, truthful, and sensitive. For most concepts of interest to nurse researchers, there are few data collection procedures that match this ideal. 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 in a study. We begin by discussing principles of measurement and assessments of quantitative data. Later in this chapter, we discuss assessments of qualitative data.

### MEASUREMENT

Quantitative studies derive data through the measurement of variables. **Measurement** involves the assignment of numbers to represent the amount of an attribute present in an object or person, using a specified set of rules. As this definition implies, quantification and measurement go hand in hand. An often-quoted statement by early American psychologist L. L. Thurstone advances a fundamental position: “Whatever exists, exists in some amount and can be measured.” Attributes are not constant: They vary from day to day, from situation to situation, or from one person to another. This variability is presumed to be capable of a numeric expression that signifies *how much* of an attribute is present.

The purpose of assigning numbers is to differentiate between people or objects that possess varying degrees of the critical attribute.

### Rules and Measurement

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

As an example, suppose we were studying attitudes toward distributing condoms in school-based clinics and asked parents to express their extent of agreement with the following statement:

Teenagers should have access to contraceptives in school clinics.

- { } Strongly agree
- { } Agree
- { } Slightly agree
- { } Neither agree nor disagree
- { } Slightly disagree
- { } Disagree
- { } Strongly disagree

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,” 27 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 any *a priori* information about the “distance” between the seven options, the most defensible procedure is to assign a 1 to “strongly agree” and a 7 to “strongly disagree.” This rule would quantitatively differentiate, in increments of one point, among people with seven different reactions to the statement. With a new instrument, researchers seldom know in advance if their rules are the best possible. New measurement rules reflect researchers’ hypotheses about how attributes function and vary. The adequacy of the hypotheses—that is, the worth of the instruments—needs to be assessed empirically.

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

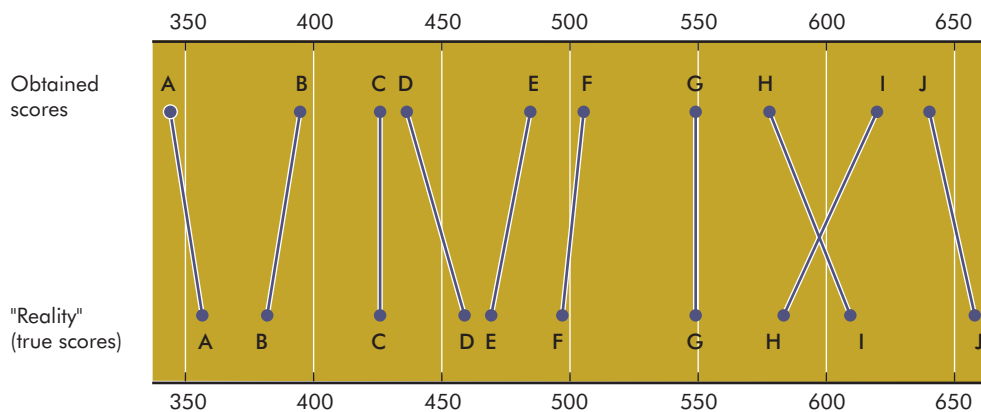
To illustrate the concept of isomorphism, suppose the Scholastic Assessment Test (SAT) were

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 18-1. Now suppose that in reality the true scores of these same students on a hypothetically perfect test were as follows: 360, 375, 430, 465, 470, 500, 550, 610, 590, and 670, as shown at the bottom of Figure 18-1. This figure shows that, although not perfect, the test came fairly close to representing true scores; only two people (H and I) were improperly ordered in the actual test. This example illustrates a measure whose isomorphism with reality is high, but improvable.

Researchers almost always work with fallible measures. Instruments that measure psychological 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 health care professionals—and researchers—would be in the absence of measurement. What would happen, for example, if there were no measures of body temperature or blood pressure? 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



**FIGURE 18.1** Relationship between obtained and true scores for a hypothetical set of test scores.

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. Not all measures are completely objective, but most incorporate mechanisms for minimizing subjectivity.

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 2 inches tall. If we chose, we could obtain even greater precision. With precise measures, researchers can more readily differentiate among people with different degrees of an attribute.

Finally, measurement is a language of communication. Numbers are less vague than words and therefore 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 develop different conceptions 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

Both the procedures involved in applying measurements and 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 always be made to do so, but such efforts are rarely completely successful.

Instruments that are not perfectly accurate yield measurements containing some error. Conceptually, an **observed** (or **obtained**) **score** can be decomposed into 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 systolic blood pressure reading or 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 in the equation 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 fashion 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 subjects step on the scale, the researcher places a hand on their shoulders and applies some pressure. The resulting measures (the  $X_O$ s) will be biased upward because the scores reflect both actual weight ( $X_T$ ) and the researcher’s pressure ( $X_E$ ). Errors of measurement are problematic because their value is unknown and also because they are variable. In this example, the amount of pressure applied likely would vary from one subject 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. The most common are 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. The anonymity of the response situation, the friendliness of researchers, or the location of the data gathering can affect subjects’ responses. Other environmental factors, such as temperature, lighting, and time of day, can represent sources of measurement error.
2. *Transitory personal factors.* A person’s score can be influenced by such temporary personal states as fatigue, hunger, anxiety, or mood. In some cases, such factors directly affect the measurement, as when anxiety affects a pulse

rate measurement. In other cases, personal factors can alter scores by influencing people's motivation to cooperate, act naturally, or do their best.

3. *Response-set biases.* Relatively enduring characteristics of respondents can interfere with accurate measures. Response sets such as social desirability, acquiescence, and extreme responses are potential problems in self-report measures, particularly in psychological scales (see Chapter 15).
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. If observers alter their coding categories, if interviewers improvise question wording, if test administrators change the test instructions, or 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 for obtaining measures are poorly understood, then scores may be affected by misunderstanding. For example, questions in a self-report instrument may be interpreted differently by different respondents, leading to a distorted measure of the variable. Observers may miscategorize observations if the classification scheme is unclear.
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 nursing knowledge will be influenced somewhat 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. Open-ended questions may yield different information than closed-ended ones. Oral responses to a question may be at odds with written responses to the same question. The ordering of questions in an instrument may also influence responses.

## RELIABILITY OF MEASURING INSTRUMENTS

The reliability\* of a quantitative instrument is a major criterion for assessing its quality and adequacy. 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, we would consider it 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 a measure's accuracy. An instrument is reliable to the extent that its measures reflect true scores—that is, to the extent that errors of measurement are absent from obtained scores. A reliable measure maximizes the true score component and minimizes the error component.

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. The method chosen 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.

---

\*The discussion of reliability presented here is based on classic measurement theory. Readers concerned with assessing the reliability of instructional measures that can be classified as mastery-type or criterion-referenced should consult Thorndike (1996).

## Stability

The **stability** of an instrument is the extent to which similar results are obtained on two separate administrations. The reliability estimate focuses on the instrument's susceptibility to extraneous factors over time, such as subject fatigue or environmental conditions.

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

To explain a reliability coefficient, we must briefly discuss a statistic known as the **correlation coefficient**.<sup>\*</sup> We have pointed out repeatedly that researchers strive to detect and explain relationships among phenomena: Is there a relationship between patients' gastric acidity levels and exposure to stress? Is there a relationship between body temperature and physical exertion? 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 relationships are. 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 does exist but is not perfect, the correlation coefficient is typically 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 wherein older people tend to sleep fewer hours and younger people tend to sleep more, but a "crossing of lines" is common. That is, many young people sleep few hours, and many 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 sample. 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 3 weeks apart to a sample of 10 people. Fictitious data for this example are presented in Table 18-1. 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 computed reliability coefficient is  $.95$ , which is high.

<sup>\*</sup>Computational procedures and additional information concerning correlation coefficients (Pearson's  $r$ ) are presented in Chapter 19.

TABLE 18.1    Fictitious Data for Test–Retest Reliability of Self-Esteem Scale			
SUBJECT 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	$r = .95$

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. In practice, reliability coefficients normally range between  $.00$  and  $1.00$ . The higher the coefficient, the more stable the measure. Reliability coefficients above  $.70$  usually are considered satisfactory. In some situations, a higher coefficient may be required, or a lower one may be acceptable.

The test–retest method is a relatively easy approach to estimating reliability, and can be used with self-report, observational, and physiologic measures.\* The test–retest approach has certain disadvantages, however. One issue is that many traits *do* change over time, independently of the measure’s stability. Attitudes, behaviors, knowledge, physical condition, and so forth can be

\*There are more sophisticated methods of assessing test–retest reliability, as described by Yen and Lo (2002).

modified by experiences between testings. Test–retest procedures confound changes from measurement error and those from true changes in the attribute being measured. 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 subjects’ responses or observers’ coding on the second administration will be influenced by their memory of initial responses or coding, regardless of the actual values the second day. Such memory interference results in spuriously high reliability coefficients. Another difficulty is that subjects 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 (i.e., those greater than 1 or 2 months) because of actual changes in the attribute being measured. Stability indexes are most appropriate for relatively enduring characteristics such as personality, abilities, or certain physical attributes such as adult height.



**Example of test–retest reliability:**

Gauthier and Froman (2001) developed an instrument called the Preferences for Care near the End of Life (PCEOL) scale. The scale’s reliability assessment included administering the scale to 38 adults 2 weeks apart. The test–retest reliability coefficients for subscales of the PCEOL ranged from  $.80$  to  $.94$ .


**Internal Consistency**

Scales and tests that involve summing items are often 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 among nurse researchers. Its popularity reflects the fact that it is economical (it requires only one test 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 and tests contain multiple **subscales** or subtests, each of which tap distinct, but related, concepts (e.g., a measure of independent functioning might include subscales for motor activities, communication, and socializing). The internal consistency of the subscales is typically assessed and, if subscale scores are summed for an overall score, the scale's internal consistency would also be assessed.

One of the oldest methods for assessing internal consistency is the **split-half technique**. For this approach, items on a scale are split into two groups and scored independently. Scores on the two half-tests then are used to compute a correlation coefficient. To illustrate, the 10 fictitious scores from the

first administration of the self-esteem scale are reproduced in the second column of Table 18-2. Let us say that the total instrument consists of 20 questions, and so the items must be divided into two groups of 10. Although many splits are possible, the usual procedure is to use odd items versus even items. One half-test, therefore, consists of items 1, 3, 5, 7, 9, 11, 13, 15, 17, and 19, and the even-numbered items compose the second half-test. Scores on the two halves are shown in the third and fourth columns of Table 18-2. The correlation coefficient for scores on the two half-tests gives an estimate of the scale's internal consistency. If the odd items are measuring the same attribute as the even items, then the reliability coefficient should be high. The correlation coefficient computed on these fictitious data is .67.

The correlation coefficient computed on split-halves tends to underestimate the reliability of the entire scale. Other things being equal, longer scales are more reliable than shorter ones. The correlation coefficient for the data in Table 18-2 is the estimated reliability for a 10-item, not a 20-item, instrument. A correction formula has been developed to give a reliability estimate for the entire test. The equation,

**TABLE 18.2** Fictitious Data for Split-Half Reliability of the Self-Esteem Scale

SUBJECT NUMBER	TOTAL SCORE	ODD-NUMBERS SCORE	EVEN-NUMBERS SCORE
1	55	28	27
2	49	26	23
3	78	36	42
4	37	18	19
5	44	23	21
6	50	30	20
7	58	30	28
8	62	33	29
9	48	23	25
10	67	28	39
$r = .80$			

known as the **Spearman-Brown prophecy formula**, is as follows for this situation:

$$r^1 = \frac{2r}{1 + r}$$

where  $r$  = the correlation coefficient computed on the split halves

$r^1$  = the estimated reliability of the entire test

Using the formula, the reliability for our hypothetical 20-item measure of self-esteem would be:

$$r^1 = \frac{(2)(.67)}{1 + .67} = .80$$

The split-half technique is easy to use, but is handicapped by the fact that different reliability estimates can be obtained with different splits. That is, it makes a difference whether one uses an odd—even split, a first-half—second-half split, or some other method of dividing items into two groups.

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 described here; the normal range of values is between .00 and +1.00, and higher values reflect a higher internal consistency. Coefficient alpha is preferable to the split-half procedure because it gives an estimate of the split-half correlation for *all possible* ways of dividing the measure into two halves. It is beyond the scope of this text to explain this method in detail, but more information is available in textbooks on psychometrics (e.g., Cronbach, 1990; Nunnally & Bernstein, 1994).\*

---

\*The coefficient alpha equation, for the advanced student, is as follows:

$$r = \frac{k}{k - 1} \left[ 1 - \frac{\Sigma \sigma_i^2}{\sigma_y^2} \right]$$

where  $r$  = the estimated reliability

$k$  = the total number of items in the test

$\sigma_i^2$  = the variance of each individual item

$\sigma_y^2$  = the variance of the total test scores

$\Sigma$  = the sum of

In summary, indices of homogeneity or internal consistency estimate the extent to which different subparts of an instrument are equivalent in measuring the critical attribute. The split-half technique has been used to estimate homogeneity, but coefficient alpha is preferable. Neither approach considers fluctuations over time as a source of unreliability.



### Example of internal consistency reliability:

Brown, Becker, Garcia, Barton, and Hanis (2002) adapted a measure of health beliefs for use with Spanish-speaking Mexican Americans with type 2 diabetes. The adapted instrument, administered to 326 Mexican Americans, was found to have 5 subscales, with alpha coefficients ranging from .56 to .90.

## Equivalence

Nurse researchers estimate a measure's reliability by way of the **equivalence** approach primarily with observational measures. In Chapter 16, we pointed out that a potential weakness of observational methods is observer error. The accuracy of observer ratings and classifications can be enhanced by careful training, the specification of clearly defined, nonoverlapping categories, and the use of a small number of categories. Even when such care is taken, researchers should assess the reliability of observational instruments. In this case, "instrument" includes both the category or rating system *and* the observers making the measurements.

**Interrater (or interobserver) reliability** is estimated by having two or more trained observers watching an event simultaneously, and independently recording data according to the instrument's instructions. The data can then be used to compute an index of equivalence or agreement between observers. For certain types of observational data (e.g., ratings), correlation techniques are suitable. That is, a correlation coefficient is computed to demonstrate the strength of the relationship between one observer's ratings and another's.

Another procedure is to compute reliability as a function of agreements, using the following equation:

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



This simple formula unfortunately tends to overestimate observer agreements. If the behavior under investigation is one that observers code for absence or presence every, say, 10 seconds, the observers will agree 50% of the time by chance alone. Other approaches to estimating interrater reliability may be of interest to advanced students. Techniques such as Cohen's kappa, analysis of variance, intraclass correlations, and rank-order correlations have been used to assess interobserver reliability.



#### Example of interrater reliability:

Kovach and Wells (2002) observed the behaviors of older people with dementia over 30-minute observation sessions. Interrater reliability, calculated as percentage agreement between two raters, was .74 for the variable *activity*, .92 for *noxiousness*, and .84 for *agitation*.

### Interpretation of Reliability Coefficients

Reliability coefficients are important indicators of an instrument's quality. Unreliable measures do not provide adequate tests of researchers' hypotheses. If data fail to confirm a prediction, one possibility is that the instruments were unreliable—not necessarily that the expected relationships do not exist. Knowledge about an instrument's reliability thus is critical in interpreting research results, especially if research hypotheses are not supported.

For group-level comparisons, coefficients in the vicinity of .70 are usually adequate, although 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 subjects. If measures are used for making decisions about individuals, then reliability coefficients ideally should be .90 or better. For instance, if a test score was used as a criterion for admission to a graduate nursing program, then the accuracy of the test would be of critical importance to both individual applicants and the school of nursing.

Reliability coefficients have a special interpretation that should be briefly explained without elaborating on technical details. This interpretation relates

to the earlier discussion of decomposing observed scores into error components and true components. Suppose we administered a scale that measures hope to 50 cancer patients. It would be expected that 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,

$$V_O = V_T + V_E$$

where  $V_O$  = observed total variability in scores

$V_T$  = true variability

$V_E$  = variability owing to random errors

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

$$r = \frac{V_T}{V_O}$$

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 random, extraneous fluctuations. Looked at in this way, it should be clearer why instruments with reliability lower than .70 are risky to use.

### Factors Affecting Reliability

Researchers who develop or adapt instruments for their own use or for use by others must undertake reliability assessments. The availability of computer programs to calculate coefficient alphas has made this task convenient and economical. There are also things that instrument developers should keep in mind during the development process that can enhance reliability.

First, as previously noted, 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. Scale

developers can assess whether items are tapping the same construct and are sufficiently discriminating by doing an **item analysis**. In general, items that elicit a 50/50 split (e.g., agree/disagree or correct/incorrect) have the best discriminating power. As a general guideline, if the split is 80/20 or worse, the item should probably be replaced. Another aspect of an item analysis is an inspection of the correlations between individual items and the overall scale score. Item-to-total correlations below .30 are usually considered unacceptably low.

With observational scales, reliability can usually be improved by greater precision in defining categories, or greater clarity in explaining the underlying dimension for rating scales. The most effective means of enhancing reliability in observational studies, however, is thorough training of observers.

The reliability of an instrument 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 being measured. For example, a depression scale will be less reliable when administered to a homeless sample than when it is used with a general population.

Choosing an instrument previously demonstrated to be reliable is no guarantee of its high quality in a new study. 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 a certain sample 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 whom it was developed. If the group is similar to the population for a new study, then the reliability estimate provided by the scale developer is probably a reasonably good index of the instrument's accuracy in the new study.



**TIP:** You should not be satisfied with an instrument that will *probably* be reliable in your study. The recommended procedure is to compute estimates of reliability whenever research data are collected. For physiologic measures that are relatively impervious to fluctuations from personal or situational factors, this procedure may be unnecessary. However, observational tools, self-report measures, tests of knowledge or ability, and projective tests—all of which are highly susceptible to errors of measurement—should be subjected to a reliability check as a routine step in the research process.

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 most relevant.



**Example of different reliability estimates:** Chaiyawat and Brown (2000) did a psychometric assessment of the Thai version of the State-Trait Anxiety Inventory for Children. The test–retest reliability was .68 for the Trait subscale and .63 for the State subscale. Cronbach alphas for both subscales at both administrations exceeded .80.

## VALIDITY

The second important criterion for evaluating a quantitative 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, how can they 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 possibly 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 the

circumference of their wrists. We could obtain highly accurate, consistent, and precise measurements of wrist circumferences, 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 of a measure *is* evidence of low validity.

Like reliability, validity has different aspects and assessment approaches. Unlike reliability, however, an instrument's validity is difficult to establish. 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.

### Face Validity

**Face validity** refers to whether the instrument *looks* as though it is measuring the appropriate construct. Although face validity should not be considered primary evidence for an instrument's validity, it is helpful for a measure to have face validity if other types of validity have also been demonstrated. For example, it might be easier to persuade people to participate in an evaluation if the instruments being used have face validity.



#### Example of face validity:

Shin and Colling (2000) undertook a cultural verification of the Profile of Mood States (POMS) scale for Korean elders. One part of the study involved an assessment of the translated scale's face validity, using a panel of Korean experts.

### Content Validity

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

For cognitive measures, the content validity question is, How representative are the questions on this test 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 the development of affective measures. Researchers designing a new instrument should begin with a thorough conceptualization of the construct so the instrument can capture the entire content domain. Such a conceptualization might come from rich first-hand knowledge, an exhaustive literature review, or findings from a qualitative inquiry.



#### Example of using qualitative data for content validity:

Holley (2000) developed a self-report scale to measure distress from fatigue in cancer patients. Items for the scale were drawn from 23 in-depth interviews with patients experiencing cancer-related fatigue.

An instrument's content validity is necessarily based on judgment. There are no completely objective methods of ensuring the adequate content coverage of an instrument. However, it is becoming increasingly common to use a panel of substantive experts to evaluate and document the content validity of new instruments. The panel typically consists of at least three experts, but a larger number may be advisable if the construct is complex. The experts are asked to evaluate individual items on the new measure as well as the entire 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 adequately measure all dimensions of the construct. With regard to item relevance, some researchers compute interrater agreement indexes and a formal **content validity index (CVI)** across the experts' ratings of each item's relevance. One procedure is to have experts rate items on a four-point scale (from 1 = not relevant to 4 = very relevant). The CVI for the total instrument is the proportion of items rated as either 3 or 4. A CVI score of .80 or better indicates good content validity.



#### Example of using a content validity index:

Rew (2000) developed a scale to tap nurses' acknowledgment of using intuition in clinical

decision making. In the first phase of the study, scale items were generated from published literature and a CVI of .96 was computed on responses from a panel of five experts.

**Criterion-Related Validity**

Establishing **criterion-related validity** involves determining the relationship between an instrument and an external criterion. The instrument is said to be valid if its scores correlate highly with scores on the criterion. For example, if a measure of attitudes toward premarital sex correlates 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.

One 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 the nursing effectiveness of nursing students, 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, well-accepted criterion. For example, a scale to measure smokers’ motivation to quit smoking has a clearcut, objective criterion (subsequent smoking).

Once a criterion is selected, criterion validity can be assessed easily. A correlation coefficient is computed between 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 researchers developed a scale to measure nurses’ professionalism. They administer the instrument to a sample of nurses and also ask the nurses to indicate how many articles they have published. The publications variable was chosen as one of many potential objective criteria of professionalism. Fictitious data are presented in Table 18-3. The correlation coefficient of .83 indicates that the professionalism scale correlates fairly well with the number of published articles. Whether the scale is really measuring professionalism is a different issue—an issue that is the concern of construct validation discussed in the next section.

**TABLE 18.3**    Fictitious Data for Criterion-Related Validity Example

SUBJECT	SCORE ON PROFESSIONALISM SCALE	NUMBER OF PUBLICATIONS	
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	$r = .83$

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 some future criterion. When a school of nursing correlates incoming students' SAT scores with subsequent grade-point averages, the predictive validity of the SATs for nursing school performance is being evaluated.

**Example of predictive validity:**

Marsh, Prochada, Pritchett, and Vojir (2000) used a predictive validity approach in their assessment of their Alzheimer's Hospice Placement Evaluation Scale (AHOPE), a scale designed to measure the appropriateness of hospice care. The criterion used was survival status 6 months after the scale was administered.

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

Resnick and Jenkins (2000) developed the Self-Efficacy for Exercise Scale (SEE). As one of their methods of validating the scale, they correlated SEE scores with whether participants engaged in regular exercise activity, defined as 20 minutes of aerobic activity three times a week.

Validation by means of the criterion-related approach is most often used in applied or 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

Validating an instrument in terms of **construct validity** is a challenging task. The key construct

validity questions are: What is this instrument really measuring? Does it adequately measure the abstract concept of interest? 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. Actually, it is really not just a question of suitability: What objective criterion is there for such concepts as empathy, role conflict, or separation anxiety? Despite the difficulty of construct validation, it is an activity vital to the development of a strong evidence base. The constructs in which nurse researchers are interested must be validly measured.

Construct validity is inextricably linked with theoretical factors. In validating a measure of death anxiety, we would be less concerned with the adequate sampling of items or with its relationship to a criterion than with its correspondence to a cogent conceptualization of death anxiety. Construct validation can be approached in several ways, but it always involves logical analysis and tests predicted by theoretical considerations. Constructs are explicated in terms of other abstract concepts; researchers make predictions about the manner in which the target construct will function in relation to other constructs.

One construct validation approach is the **known-groups technique**. In this procedure, the instrument is administered to groups expected to differ on the critical attribute because of some known characteristic. For instance, in validating a measure of fear of the labor experience, 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 some fears. On the whole, however, we would anticipate differences in average group scores.

**Example of the known-groups technique:**

Davies and Hodnett (2002) developed a scale to measure nurses' self-efficacy in providing



support to women in labor. To validate the scale, they compared the scores of labor and delivery nurses with those of nurses who worked in postpartum care and found significantly higher scores among the first group, as predicted.

Another method of construct validation involves an examination of relationships based on theoretical predictions, which is really a variant of the known-groups approach. 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 by theory.
- Therefore, it is inferred that A and B are valid measures of X and Y.

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



**Example of testing relationships:**

Ryden and her colleagues (2000) created a Satisfaction with the Nursing Home Instrument (SNHI). Their approach to construct validation included a scrutiny of the correlation between SNHI

scores with scores on two affective measures. As predicted, SNHI scores were negatively associated with depression and positively associated with morale.

A significant construct validation tool is a procedure known as the **multitrait–multimethod matrix method (MTMM)** (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 the construct validity question.

To help explain the MTMM approach, fictitious data from a study to validate a “need for autonomy” measure are presented in Table 18-4. 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 summated rating 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 respond to a pictorial (projective) stimulus depicting an autonomy-relevant situation.

**TABLE 18.4** Multitrait–Multimethod Matrix

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

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



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

The numbers in Table 18-4 represent the correlation coefficients between the scores on six different measures (two traits  $\times$  three methods). For instance, the coefficient of  $-.38$  at the intersection of  $AUT_1$ – $AFF_1$  expresses the relationship 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 aspects 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 in fact. This requirement is a minimum one that, if failed, should cause re-

searchers 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 18-4 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 clearcut as in this contrived example. Indeed, a common problem with the MTMM is interpreting the pattern of coefficients. Another issue is that there are no clearcut criteria for determining 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. Executing any part of the model is better than no effort at construct validation.



#### Example of convergent validity:

Garvin and Kim (2000) evaluated the reliability and validity of three instruments measuring patients' preference for information. Convergent validity between two of the three measures was fair in both a Korean and U.S. sample ( $.30$  and  $.51$ , respectively), but the third instrument had inadequate convergent validity.



#### Example of divergent validity:

Cacchione (2002) compared four acute confusion assessment instruments. She found that only one of them (the Visual Analog Scale for Acute Confusion) demonstrated divergent validity with a measure of geriatric depression.

Another approach to construct validation uses a statistical procedure known as factor analysis.\*

\*The **absolute magnitude** refers to the value without a plus or minus sign. A value of  $-.50$  is of a higher absolute magnitude than  $+.40$ .

\*Another sophisticated and complex approach to construct validation is based on **item response theory**. For an explanation, see Chapter 10 of Nunnally and Bernstein (1994). See also Hambleton, Swaminathan, and Rogers (1991).

Although factor analysis, which is discussed in Chapter 21, is computationally complex, it is conceptually rather simple. **Factor analysis** is a method for identifying clusters of related variables. Each cluster, called a **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 looking at the convergent and discriminant validity of a large set of items. Indeed, a procedure known as **confirmatory factor analysis** is sometimes used as a method for analyzing MTMM data (Ferketich, Figueredo, & Knapp, 1991; Lowe & Ryan-Wenger, 1992).

Construct validation is the most important type of validity for a quantitative instrument. Instrument developers should use one or more of the techniques described here in their effort to assess the instrument's worth.

### 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:** Instrument developers usually gather evidence of the validity and reliability of their instruments in a thorough **psychometric assessment** before making them available for general use. If you use an existing instrument, choose one with demonstrated high reliability and validity. If you select an instrument for which there is no

published psychometric information, try contacting the instrument developer and ask about evidence of data quality.

## OTHER CRITERIA FOR ASSESSING QUANTITATIVE MEASURES

Reliability and validity are the two most important criteria for evaluating quantitative instruments. High reliability and validity are a necessary, although not sufficient, condition for good quantitative research. Researchers sometimes need to consider other qualities of an instrument, as discussed in this section.

### Sensitivity and Specificity

Sensitivity and specificity are criteria that are important in evaluating instruments designed as screening instruments or diagnostic aids. For example, a researcher might develop a new scale to measure risk of osteoporosis. Such screening/diagnostic instruments could be self-report, observational, or biophysiologic measures.

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

There is, unfortunately, a tradeoff between the sensitivity and specificity of an instrument. When sensitivity is increased to include more true positives, the number of true negatives declines. Therefore, a critical task is to develop the appropriate **cutoff point**, that is, the score value used to distinguish cases and noncases. To determine the best cutoff point, researchers often use what is called a **receiver operating characteristic curve (ROC curve)**. To construct an ROC curve, the sensitivity

of an instrument (i.e., the rate of correctly identifying a case vis-à-vis a 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 cutoff points. The 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. The example at the end of this chapter illustrates the use of ROC curves. Fletcher, Fletcher, and Wagner (1996) is a good source for further information about these procedures.



#### Example of sensitivity and specificity:

Bergquist and Frantz (2001) studied the use of the Braden scale for predicting pressure ulcers in a sample of community-based adults receiving home health care. The sensitivity and specificity of the scale at cutoff scores between 16 and 22 were assessed against actual pressure ulcer development. The cutoff score of 19 yielded the best balance between sensitivity (61%) and specificity (68%) for stage I to IV pressure ulcers.

### Efficiency

Instruments of comparable reliability and validity may differ in their efficiency. A depression scale that requires 10 minutes of people's time is efficient compared with a depression scale that requires 30 minutes to complete. One aspect of efficiency is the number of items incorporated in an instrument. Long instruments tend to be more reliable than shorter ones. There is, however, 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. Using the Spearman-Brown formula, we can estimate how reliable the scale would be with only 30 items:

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

where  $k$  = the factor by which the instrument is being incremented or decreased; in this case,  $k = 30 \div 40 = .75$

$r^1$  = reliability estimate for shorter (longer) scale

As this calculation shows, 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 subjects' response burden and data collection costs.

Efficiency is more characteristic of certain types of data collection procedures than others. In self-reports, closed-ended questions are more efficient than open-ended ones. Self-report scales tend to be less time-consuming than projective instruments for a comparable amount of information. Of course, a researcher may decide that other advantages (such as depth of information) offset a certain degree of inefficiency. Other things being equal, however, 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 six criteria are actually aspects of the reliability and validity issues:

1. *Comprehensibility.* Subjects 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. *Speededness.* For most instruments, researchers should allow adequate time to obtain complete measurements without rushing the measuring process.
4. *Range.* The instrument should be capable of achieving a meaningful measure from the smallest expected value of the variable to the largest.
5. *Linearity.* A researcher normally strives to construct measures that are equally accurate and sensitive over the entire range of values.
6. *Reactivity.* The instrument should, insofar as possible, avoid affecting the attribute being measured.

## ASSESSMENT OF QUALITATIVE DATA AND THEIR INTERPRETATION

The criteria and methods of assessment described thus far apply to quantitative data collection instruments. The procedures cannot be meaningfully applied to such qualitative materials as narrative interview data or descriptions from a participant observer's field notes, but qualitative researchers are also concerned with data quality. The central question underlying the concepts of reliability and validity is: Do the data reflect the truth? Qualitative researchers are as eager as quantitative researchers to have data reflecting the true state of human experience.

Nevertheless, there has been considerable controversy about the criteria to use for assessing the "truth value" of qualitative research. Whittemore, Chase, and Mandle (2001), who listed different criteria recommended by 10 influential authorities, noted that the difficulty in achieving universally accepted criteria (or even universally accepted labels for those criteria) stems in part from various tensions, such as the tension between the desire for rigor and the desire for creativity.

The criteria currently thought of as the gold standard for qualitative researchers are those outlined by Lincoln and Guba (1985). As noted in Chapter 2, these researchers have suggested four criteria for establishing the **trustworthiness** of qualitative data: credibility, dependability, confirmability, and transferability. These criteria go beyond an assessment of qualitative *data* alone, but rather are concerned with evaluations of interpretations and conclusions as well. These standards are often used by qualitative researchers in all major traditions, but some exceptions are noted.

### Credibility

Credibility is viewed by Lincoln and Guba as an overriding goal of qualitative research, and is considered in the Whittemore et al. (2001) synthesis as a primary validity criterion. **Credibility** refers to confidence in the truth of the data and interpreta-

tions of them. Lincoln and Guba point out that credibility involves two aspects: first, carrying out the study in a way that enhances the believability of the findings, and second, taking steps to *demonstrate* credibility to consumers. They suggest a variety of techniques for improving and documenting the credibility of qualitative research.

### Prolonged Engagement and Persistent Observation

Lincoln and Guba recommend several activities that make it more likely to produce credible data and interpretations. A first and very important step is **prolonged engagement**—the investment of sufficient time collecting data to have an in-depth understanding of the culture, language, or views of the group under study and to test for misinformation and distortions. Prolonged engagement is also essential for building trust and rapport with informants, which in turn makes it more likely that useful, accurate, and rich information will be obtained.



#### Example of prolonged engagement:

Albertín-Carbó, Domingo-Salvany, and Hartnoll (2001) studied the meaning that injecting drug users attribute to risk behaviors linked to HIV transmission. They gathered participant observation data (e.g., accompanying people to look for drugs, walking the streets) during 10 months of fieldwork in a district of Barcelona that had a high prevalence of opiate use.

Credible data collection in naturalistic inquiries also involves **persistent observation**, which concerns the salience of the data being gathered and recorded. Persistent observation refers to the researchers' focus on the characteristics or aspects of a situation or a conversation that are relevant to the phenomena being studied. As Lincoln and Guba (1985) note, "If prolonged engagement provides scope, persistent observation provides depth" (p. 304).



#### Example of persistent observation:

Beck (2002) conducted a grounded theory study of mothering twins during the first year of

life. In addition to prolonged engagement for 10 months of fieldwork, she engaged in persistent observation. After interviewing mothers in their homes, Beck often stayed and helped them with their twins, using the time for persistent observation of the mothers caretaking (e.g., details of what and how the mothers talked to their twins).

### Triangulation

Triangulation can also enhance credibility. As previously noted, triangulation refers to the use of multiple referents to draw conclusions about what constitutes truth, and has been compared with convergent validation. The aim of triangulation is to “overcome the intrinsic bias that comes from single-method, single-observer, and single-theory studies” (Denzin, 1989, p. 313). It has also been argued that triangulation helps to capture a more complete and contextualized portrait of the phenomenon under study—a goal shared by researchers in all qualitative traditions. Denzin (1989) identified four types of triangulation: data triangulation, investigator triangulation, method triangulation, and theory triangulation.

**Data triangulation** involves the use of multiple data sources for the purpose of validating conclusions. There are three basic types of data triangulation: time, space, and person. **Time triangulation** involves collecting data on the same phenomenon or about the same people at different points in time. Time triangulation can involve gathering data at different times of the day, or at different times in the year. This concept is similar to test–retest reliability assessment; that is, the point is not to study the phenomenon longitudinally to determine how it changes, but to determine the congruence of the phenomenon across time. **Space triangulation** involves collecting data on the same phenomenon in multiple sites. The aim is to validate the data by testing for cross-site consistency. Finally, **person triangulation** involves collecting data from different levels of persons: individuals, groups (e.g., dyads, triads, families), and collectives (e.g., organizations, communities, institutions), with the aim of validating data through multiple perspectives on the phenomenon.



#### Example of data (person/space) triangulation:

Lipson (2001) studied the experience of living with multiple chemical sensitivity (MCS). She collected data in four sites (Dallas, San Francisco, Vancouver, and Halifax). Participant observation included two treatment centers, a support organization, and an Internet chat room; Lipson conducted interviews with MCS sufferers, activists, and educators.

The second major type of triangulation is **investigator triangulation**, which refers to the use of two or more researchers to analyze and interpret a data set. Through collaboration, investigators can reduce the possibility of a biased interpretation of the data. Moreover, if the investigators bring to the analysis task a complementary blend of skills and expertise, the analysis and interpretation can benefit from divergent perspectives. Blending diverse methodologic, disciplinary, and clinical skills also can contribute to other types of triangulation. Investigator triangulation is conceptually somewhat similar to interrater reliability in quantitative studies.



#### Example of investigator triangulation:

Woodhouse, Sayre, and Livingood (2001), who evaluated youth tobacco prevention policies in Florida, did an ethnography of tobacco possession enforcement. Interview data were analyzed simultaneously and separately by two researchers. The two analyses were compared and contrasted “to perfect the definition of themes and provide triangulation of findings” (p. 687).

With **theory triangulation**, researchers use competing theories or hypotheses in the analysis and interpretation of their data. Qualitative researchers who develop alternative hypotheses while still in the field can test the validity of each because the flexible design of qualitative studies provides ongoing opportunities to direct the inquiry. Theory triangulation can help researchers to rule out rival hypotheses and to prevent premature conceptualizations. The quantitative analogue for theory triangulation is construct validation.



**Method triangulation** involves the use of multiple methods of data collection about the same phenomenon. In qualitative studies,\* researchers often use a rich blend of unstructured data collection methods (e.g., interviews, observations, documents) to develop a comprehensive understanding of a phenomenon. Multiple data collection methods provide an opportunity to evaluate the extent to which an internally consistent picture of the phenomenon emerges.



#### Example of method triangulation:

Carter (2002) studied the experiences of chronic pain in children and their families. Data were collected by means of journals (in which study participants reflected on what it was like to live with chronic pain) and in-depth interviews.

Although Denzin's (1989) seminal work discussed these four types of triangulation as a method of converging on valid understandings about a phenomenon, other types have been suggested. For example, Kimchi, Polivka, and Stephenson (1991) have described **analysis triangulation** (i.e., using two or more analytic techniques to analyze the same set of data). This approach offers another opportunity to validate the meanings inherent in a qualitative data set. Analysis triangulation can also involve using multiple units of analysis (e.g., individuals, dyads, families). Finally, **multiple triangulation** is used when more than one of these types of triangulation is used in the collection and analysis of the same data set.

In summary, the purpose of using triangulation is to provide a basis for convergence on the truth. By using multiple methods and perspectives, researchers strive to sort out "true" information from "error" information, thereby enhancing the credibility of the findings.

### Peer Debriefing

Another technique for establishing credibility involves external validation. **Peer debriefing** involves

sessions with peers to review and explore various aspects of the inquiry. Peer debriefing exposes researchers to the searching questions of others who are experienced in either the methods of naturalistic inquiry, the phenomenon being studied, or both.

In a peer debriefing session, researchers might present written or oral summaries of the data that have been gathered, categories and themes that are emerging, and researchers' interpretations of the data. In some cases, taped interviews might be played. Among the questions that peer debriefers might address are the following:

- Is there evidence of researcher bias?
- Have the researchers been sufficiently reflexive?
- Do the gathered data adequately portray the phenomenon?
- If there are important omissions, what strategies might remedy this problem?
- Are there any apparent errors of fact?
- Are there possible errors of interpretation?
- Are there competing interpretations? More comprehensive or parsimonious interpretations?
- Have all important themes been identified?
- Are the themes and interpretations knit together into a cogent, useful, and creative conceptualization of the phenomenon?



#### Example of peer debriefing:

Phillips, Cohen, and Tarzian (2001) conducted a phenomenological study of the experience of breast cancer screening for African-American women. The researchers interviewed 23 low- and middle-income women; then another objective researcher reviewed each interview, including the questions and techniques used. This peer reviewer debriefed with the three researchers before the next interview was conducted.

### Member Checking

Lincoln and Guba consider member checking the most important technique for establishing the credibility of qualitative data. In a **member check**, researchers provide feedback to study participants regarding the emerging data and interpretations, and obtain participants' reactions. If researchers purport

---

\*We have already discussed method triangulation involving a combination of qualitative and quantitative approaches (see Chapter 12).



that their interpretations are good representations of participants' realities, participants should be given an opportunity to react to them. Member checking with participants can be carried out both informally in an ongoing way as data are being collected, and more formally after data have been fully analyzed.



**TIP:** Not all qualitative researchers use member checking to ensure credibility. For example, member checking is not a component of Giorgi's method of descriptive phenomenology. Giorgi (1989) argued that asking participants to evaluate the researcher's psychological interpretation of their own descriptions exceeds the role of participants.

Member checking is sometimes done in writing. For example, researchers can ask participants to review and comment on case summaries, interpretive notes, thematic summaries, or drafts of the research report. Member checks are more typically done in face-to-face discussions with individual participants or small groups of participants. Many of the questions relevant for peer debriefings are also appropriate in the context of member checks.

Despite the role that member checking can play in enhancing credibility and demonstrating it to consumers, several issues need to be kept in mind. One is that some participants may be unwilling to participate in this process. Some—especially if the research topic is emotionally charged—may feel they have attained closure once they have shared their concerns, feelings, and experiences. Further discussion might not be welcomed. Others may decline being involved in member checking because they are afraid it might arouse suspicions of their families. Choudhry (2001) encountered this in her study of the challenges faced by elderly women from India who had immigrated to Canada. When Choudhry asked participants for a second interview to examine the transcripts of their first interviews, the participants refused. They feared that a second visit to their homes might arouse suspicions among their family members and increase their sense of loss and regret.

A second issue is that member checks can lead to misleading conclusions of credibility if partici-

pants “share some common myth or front, or conspire to mislead or cover up” (Lincoln & Guba, p. 315). At the other extreme, some participants might express agreement (or fail to express disagreement) with researchers' interpretations either out of politeness or in the belief that researchers are “smarter” or more knowledgeable than they themselves are.



**TIP:** It is important to explain to participants the helpful role that member checking plays in establishing trustworthiness. Participants should be given every encouragement to provide critical feedback about factual or interpretive errors or inadequacies.



#### Example of member checking:

King, Cathers, Polgar, MacKinnon, and Havens (2000) interviewed 10 adolescents with cerebral palsy to determine how they defined success in life. After the researchers' original thematic analysis, major themes and text segments representing themes were presented to a subgroup of participants in a member-checking focus group. Participants were asked to appraise critically the researchers' interpretations, and the importance of honest feedback was emphasized. Major themes were confirmed, but focus group members provided additional information elaborating on the themes.

### Searching for Disconfirming Evidence

The credibility of a data set can be enhanced by the researcher's systematic search for data that will challenge an emerging categorization or descriptive theory. The search for disconfirming evidence occurs through purposive sampling methods but is facilitated through other processes already described here, such as prolonged engagement and peer debriefings. As noted in Chapter 13, the purposive sampling of individuals who can offer conflicting accounts or points of view can greatly strengthen a comprehensive description of a phenomenon.

Lincoln and Guba (1985) refer to a similar activity of **negative case analysis**—a process by which researchers revise their interpretations by

including cases that appear to disconfirm earlier hypotheses. The goal of this procedure is to continuously refine a hypothesis or theory until it accounts for *all* cases.



#### Example of negative case analysis:

Beck (1995) developed three different versions describing the experience of burnout in nursing students, with refinements stemming from negative cases. The original formulation was: “Burnout occurs in nursing students as they become engulfed with competing demands from school, work, and family. Fatigue is all encompassing. Students gain weight from overeating due to stress. Social activities and physical exercise are outlets for this stress.” This description had to be revised when other nursing students revealed that they had lost weight because of a lack of appetite and not having enough time to eat. The second version was: “Burnout occurs in nursing students as they become engulfed with competing demands from school, work, and family. Fatigue is all encompassing. Stress affects students’ weight but not in a uniform direction. Some nursing students gain weight while others lose weight. Social activities and physical exercise are outlets for this stress.” A third revision became necessary as subsequent interviews revealed that other students noted their lack of outlets for their stress because they did not have free time to socialize or exercise. Based on these negative cases, the latest version was as follows: “Burnout occurs in nursing students as they become engulfed with competing demands from school, work, and family. Fatigue is all encompassing. Stress affects students’ weight but not in a uniform direction. Some nursing students gain weight while others lose weight. Outlets such as physical exercise and social activities are life savers but not for all nursing students. Due to time pressures, some could not indulge in these stress relieving activities.”

#### Researcher Credibility

Another aspect of credibility discussed by Patton (2002) is **researcher credibility**, that is, the faith that can be put in the researcher. In qualitative studies, researchers *are* the data collecting instruments—as well as creators of the analytic process.

Therefore, researcher qualifications, experience, and reflexivity are important in establishing confidence in the data.

It is sometimes argued that, for readers to have confidence in the validity of a qualitative study’s findings, the research report should contain information about the researchers, including information about credentials. In addition, the report may need to make clear the personal connections they had to the people, topic, or community under study. For example, it is relevant for a reader of a report on the coping mechanisms of AIDS patients to know that the researcher is HIV positive. Patton argues that researchers should report “any personal and professional information that may have affected data collection, analysis and interpretation—either negatively or positively...” (p. 566).



#### Example of researcher credibility:

Mohr (2000) studied how families with children under care in mental health care settings experience that care. In a section of her report labeled “Reflexive Notes,” Mohr presented her credentials as a nurse (various nursing roles in psychiatric hospitals), her personal background (her mother was schizophrenic), and her strong advocacy activities (involvement with the National Alliance for the Mentally Ill). In addition, a brief biography establishing her credentials was included at the end of the report, a common feature in the journal *Qualitative Health Research*.

### Dependability

The second criterion used to assess trustworthiness in qualitative research is dependability. The **dependability** of qualitative data refers to the stability of data over time and over conditions. This is similar conceptually to the stability and equivalence aspects of reliability assessments in quantitative studies (and similar also to time triangulation).

One approach to assessing the dependability of data is to undertake a procedure referred to as **step-wise replication**. This approach involves having a research team that can be divided into two groups. These groups deal with data sources separately and conduct, essentially, independent inquiries through

which data can be compared. Ongoing, regular communication between the groups is essential for the success of this procedure.

Another technique relating to dependability is the **inquiry audit**. An inquiry audit involves a scrutiny of the data and relevant supporting documents by an external reviewer, an approach that also has a bearing on the confirmability of the data, a topic we discuss next.



#### Example of dependability:

Williams, Schutte, Evers, and Holkup (2000) used a stepwise replication and inquiry audit in their study of coping with normal results from predictive gene testing for neurodegenerative disorders. Ten participants were interviewed three times. Three researchers read through transcripts of the first set of interviews and made marginal notes for coding. The three researchers compared their codes and revised them until agreement was reached. All transcripts from the remaining interviews were coded independently, and the researchers met periodically to compare codes and reach a consensus. Once coding was completed, a qualitative nurse researcher reviewed the entire set of transcribed interviews and validated the findings with the three researchers.

## Confirmability

**Confirmability** refers to the objectivity or neutrality of the data, that is, the potential for congruence between two or more independent people about the data's accuracy, relevance, or meaning. Bracketing (in phenomenological studies) and maintaining a reflexive journal are methods that can enhance confirmability, although these strategies do not actually document that it has been achieved.

Inquiry audits can be used to establish both the dependability and confirmability of the data. For an inquiry audit, researchers develop an **audit trail**, that is, a systematic collection of materials and documentation that allows an independent auditor to come to conclusions about the data. There are six classes of records that are of special interest in creating an adequate audit trail: (1) the raw data

(e.g., field notes, interview transcripts); (2) data reduction and analysis products (e.g., theoretical notes, documentation on working hypotheses); (3) process notes (e.g., methodologic notes, notes from member check sessions); (4) materials relating to researchers' intentions and dispositions (e.g., reflexive notes); (5) instrument development information (e.g., pilot forms); and (6) data reconstruction products (e.g., drafts of the final report).

Once the audit trail materials are assembled, the inquiry auditor proceeds to audit, in a fashion analogous to a financial audit, the trustworthiness of the data and the meanings attached to them. Although the auditing task is complex, it can serve as an invaluable tool for persuading others that qualitative data are worthy of confidence. Relatively few comprehensive inquiry audits have been reported in the literature, but some studies report partial audits or the assembling of auditable materials. Rodgers and Cowles (1993) present useful information about inquiry audits.



#### Example of confirmability:

In her research on mothering twins, Beck (2002) developed a four-phased grounded theory entitled, "life on hold: releasing the pause button." In her report, Beck provided a partial audit trail for phase three, which she called "striving to reset."

## Transferability

In Lincoln and Guba's (1985) framework, **transferability** refers essentially to the generalizability of the data, that is, the extent to which the findings can be transferred to other settings or groups. This is, to some extent, a sampling and design issue rather than an issue relating to the soundness of the data per se. However, as Lincoln and Guba note, the responsibility of the investigator is to provide sufficient descriptive data in the research report so that consumers can evaluate the applicability of the data to other contexts: "Thus the naturalist cannot specify the external validity of an inquiry; he or she can provide only the thick description necessary to enable someone interested in making a transfer to reach a conclusion about whether transfer can be

contemplated as a possibility” (p. 316). Thick description, as noted earlier, refers to a rich and thorough description of the research setting or context and of the transactions and processes observed during the inquiry. Thus, if there is to be transferability, the burden of proof rests with the investigator to provide sufficient information to permit judgments about contextual similarity.



#### Example of transferability:

In their phenomenological study of homeless patients’ experience of satisfaction with health care, McCabe, Macnee, and Anderson (2001) interviewed 17 homeless people at a nurse-managed primary health care clinic for the homeless, at three shelters, and at a night-time soup kitchen. To assess the transferability of their themes, the researchers checked their findings with a group of homeless people who had not participated in the study and who lived in a shelter in a neighboring city.



**TIP:** Establishing the trustworthiness of focus group data poses special challenges to researchers. Morrison-Beedy, Côté-Arsenault, and Feinstein (2001) include many excellent suggestions.

### Other Criteria for Assessing Quality in Qualitative Research

Qualitative researchers who take steps to enhance, assess, and document quality are most likely to use Lincoln and Guba’s criteria. However, as noted previously, other criteria have been proposed, and new ways of thinking about quality assessments for qualitative studies are emerging.

Whittemore, Chase, and Mandle (2001), in their synthesis of qualitative criteria, use the term *validity* as the overarching goal. Although this term has been eschewed by many qualitative researchers as a “translation” from quantitative perspectives, Whittemore and her colleagues argue that validity is the most appropriate term. According to their view, the dictionary definition of validity as “the state or quality of being sound, just, and well-founded” lends itself equally to qualitative and quantitative research.

In their synthesis of criteria that can be used to develop evidence of validity in qualitative studies, Whittemore and associates proposed four primary criteria and six secondary criteria. In their view, the primary criteria are essential to all qualitative inquiry, whereas secondary criteria provide supplementary benchmarks of validity and are not relevant to every study. They argue that judgment is needed to determine the optimal weight given to each of the 10 criteria in specific studies. The primary criteria include credibility (as discussed earlier), authenticity, criticality, and integrity. The six secondary criteria include explicitness, vividness, creativity, thoroughness, and congruence. Table 18-5 lists these 10 criteria and the assessment questions relevant to achieving each. The questions are ones that can be used by qualitative researchers in their efforts to enhance the rigor of their studies and by consumers to evaluate the quality of the evidence studies yield.

A scrutiny of Table 18-5 reveals that the list contains many of the same concerns as those encompassed in Guba and Lincoln’s four criteria. This overlap is further illustrated by considering techniques that can be used to contribute evidence of study validity according to these 10 criteria. As shown in Table 18-6, many of the techniques previously described in this chapter, as well as some methods discussed in earlier chapters, are important strategies for developing evidence of validity. These techniques can be used throughout the data collection and analysis process, and in preparing research reports.

Meadows and Morse (2001) discuss the components of rigor in qualitative studies and, similar to Whittemore and colleagues, conclude that the traditional terms of validity and reliability are appropriate in qualitative studies. Meadows and Morse argued that by not using traditional quantitative terminology, qualitative research has not yet taken its rightful place in the world of evidence and science. They call for the use of three components of rigor: verification, validation, and validity. Verification refers to strategies researchers use to enhance validity in the process of conducting a high-quality study. Verification strategies include the

TABLE 18.5 Primary and Secondary Qualitative Validity Criteria\*

CRITERIA	ASSESSMENT QUESTIONS
<b>Primary Criteria</b>	
Credibility	Do the research results reflect participants' experiences and their context in a believable way?
Authenticity	Has the researcher adequately represented the multiple realities of those being studied? Has an emic perspective been accurately portrayed?
Criticality	Has the inquiry involved critical appraisal and reflexivity?
Integrity	Does the research reflect ongoing checks on the many aspects of validity? Are the findings humbly presented?
<b>Secondary Criteria</b>	
Explicitness	Have methodologic decisions been explained and justified? Have biases been identified? Is evidence presented in support of conclusions and interpretations?
Vividness	Have rich, evocative, and compelling descriptions been presented?
Creativity	Do the findings illuminate in an insightful and original way? Are new perspectives and rich imagination brought to bear on the inquiry?
Thoroughness	Has sufficient attention been paid to sampling adequacy, information richness, data saturation, and contextual completeness?
Congruence	Is there congruity between the questions and the methods, the methods and the participants, the data and categories? Do themes fit together in a coherent way?
Sensitivity	Has the research been undertaken in a way that is sensitive to the cultural, social, and political contexts of those being studied?

\*Criteria are from Whittemore and colleagues' (2001) synthesis of qualitative validity criteria. The assessment questions are adapted from Whittemore and colleagues and other sources.

conduct of a thorough literature review, bracketing, theoretical sampling, and data saturation. Validation deals with the researcher's efforts to *assess* validity, apart from efforts to enhance it. Validation techniques include those discussed earlier, such as member checking, inquiry audits, triangulation, and so on. The final step in achieving validity involves the use of external judges to assess whether the project as a whole is trustworthy and valid.



**TIP:** Unfortunately, most qualitative reports do *not* provide information about validity

efforts and assessments. With increasing emphasis on developing evidence for practice, nurses should expect such information when they read reports, and include it in the reports they prepare.

## RESEARCH EXAMPLES

In this section, we describe the efforts of researchers to develop and evaluate a structured self-report instrument, and of another researcher to evaluate her qualitative data.

(Text continue on page 440)

TABLE 18.6 Techniques for Addressing Criteria for Qualitative Validity

TECHNIQUE	CRITERIA									
	Credibility	Authenticity	Criticality	Integrity	Explicitness	Vividness	Creativity	Thoroughness	Congruence	Sensitivity
Data Generation										
Persistent observation	X	X		X						X
Prolonged engagement	X	X								X
Bracketing			X							
Reflexive journaling		X	X							
Comprehensive field notes				X		X				
Theoretical sampling								X		
Disconfirming evidence	X		X					X		
Triangulation (data, method)	X							X		
Verbatim recording	X	X			X	X				
Stepwise replication					X			X		
Data saturation								X		
Data Management										
Transcription rigor	X		X							
Maintenance of audit trail				X	X				X	



Data Analysis					
Investigator triangulation	X	X			X
Theory triangulation	X			X	X
Analysis triangulation	X			X	
Peer debriefing	X	X			X
Member checking	X	X			
Negative case analysis	X	X		X	
Report Preparation					
Thick description			X		X
Researcher credibility	X				
Evidence supporting interpretation	X		X	X	
Demonstrating reflexivity		X			
Humble presentation				X	

## Research Example Involving Assessment of a Structured Scale

Beck studied the phenomenon of postpartum depression (PPD) in a series of qualitative studies, using both a phenomenological approach (1992, 1996) and a grounded theory approach (1993). Based on her in-depth understanding of PPD, she began in the late 1990s to develop a scale that could be used to screen for PPD, the Postpartum Depression Screening Scale (PDSS). Working with an expert psychometrician, Beck refined and evaluated the PDSS, as documented in two reports (Beck & Gable, 2000, 2001).

The PDSS is a Likert scale designed to tap seven dimensions, such as sleeping/eating disturbances and cognitive impairment. A 56-item pilot form of the PDSS was developed (8 per dimension), using direct quotes from women interviewed in Beck's studies of PPD (e.g., "I felt like I was losing my mind"). The reading level of the PDSS was assessed as at the seventh grade level. The pilot form was subjected to two content validation procedures, including ratings by a panel of five content experts. Feedback from these procedures led to several changes (e.g., removal and addition of some items).

The revised PDSS was then subjected to rigorous psychometric testing. Eight health care facilities in 6 states and a PPD support group distributed the PDSS to a sample of 525 new mothers. Item analysis procedures were used to streamline the scale while maintaining adequate reliability. That is, Beck and Gable determined that three items could be deleted from each subscale without sacrificing internal consistency; they examined the correlations between individual items and subscale scores to select which items to delete.

Figure 18-2 shows a portion of the computer printout for the reliability analysis of the five selected items for the cognitive impairment subscale. Some information in this table is difficult to explain to those without statistical backgrounds, but we will point out a few salient pieces of information. In the panel labeled "Item-total Statistics," the first column identifies the items in the subscale by number: I11, I18, and so on (I11 is the "I felt like I was losing my mind" item). In the fourth column are correlation coefficients indicating the strength of the relationship between a woman's score on an item and her score on the total five-item subscale. I11 has an item-total correlation of nearly .80, which is very high; all five items have excellent correlations with the total score. The sixth and

final column indicates what the internal consistency reliability would be if an item were deleted. If I11 were removed from the subscale and only the four other items remained, the reliability coefficient would be .8876; in the panel below, we see that the reliability for the 5-item scale is even higher:  $\alpha = .9120$ . Deleting any of the five items on the scale would reduce the internal consistency of the scale, albeit by a rather small amount. On the seven subscales of the PDSS, each now with five items, the reliability coefficients ranged from .83 to .94, demonstrating high internal consistency across all dimensions.

Beck and Gable (2000) used confirmatory factor analysis to evaluate the construct validity of the PDSS. Essentially, this procedure involves a validation of Beck's hypotheses about how individual items map onto underlying constructs, like cognitive impairment. Item response theory was also used, and both techniques provided evidence of the scale's construct validity.

Beck and Gable (2001) administered the PDSS to a second sample of 150 new mothers in a further effort to validate the scale, and to establish the best possible cutoff points for diagnostic assessment. In this new study, reliability coefficients were also strong, ranging from .80 to .91 for the seven subscales.

A convergent validity approach was used to examine construct validity of the overall scale. Women in the sample completed two measures of PPD and one measure of general depression. Correlations among the three were high. Validity was further established by having each study participant interviewed by a nurse psychotherapist, who used a rigorous interviewing process to confirm and document a suspected diagnosis for major or minor depression with postpartum onset. The correlation between this expert diagnosis and scores on the PDSS was .70, which was higher than the correlations between the diagnosis and scores on the other two depression scales, indicating its superiority as a screening instrument.

ROC curves were then constructed to examine the sensitivity and specificity of the PDSS at different cutoff points, using the expert diagnosis to establish true positives and true negatives. Figure 18-3 presents the ROC curve for a diagnosis of major or minor depression (46 of the 150 mothers had this diagnosis). Sensitivity, the rate of true positives, is plotted on the vertical axis. The rate of false positives (the inverse of the true-negative rate, or specificity) is plotted on the horizontal axis. To illustrate how to read ROC information, let us suppose we used a cutoff score of 95 on

\*\*\*\*\* Method 2 (covariance matrix) will be used for this analysis \*\*\*\*\*

RELIABILITY ANALYSIS - SCALE (ALPHA)

1.	I11					
2.	I18					
3.	I25					
4.	I39					
5.	I53					

		Mean	Std Dev	Cases		
1.	I11	2.3640	1.4243	522.0		
2.	I18	2.2146	1.2697	522.0		
3.	I25	2.2050	1.3736	522.0		
4.	I39	2.3985	1.3511	522.0		
5.	I53	2.2759	1.3491	522.0		

Correlation Matrix

	I11	I18	I25	I39	I53	
I11	1.0000					
I18	.6540	1.0000				
I25	.8143	.6031	1.0000			
I39	.6456	.6594	.6519	1.0000		
I53	.6469	.7508	.6012	.7240	1.0000	

N of Cases = 522.0

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	2.2916	2.2050	2.3985	.1935	1.0877	.0076

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	1.8347	1.6122	2.0285	.4163	1.2582	.0225

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
I11	9.0939	21.3712	.7991	.7145	.8876
I18	9.2433	23.0060	.7639	.6231	.8951
I25	9.2529	22.0972	.7696	.6912	.8937
I39	9.0594	22.2901	.7687	.6097	.8938
I53	9.1820	22.1760	.7812	.6662	.8912

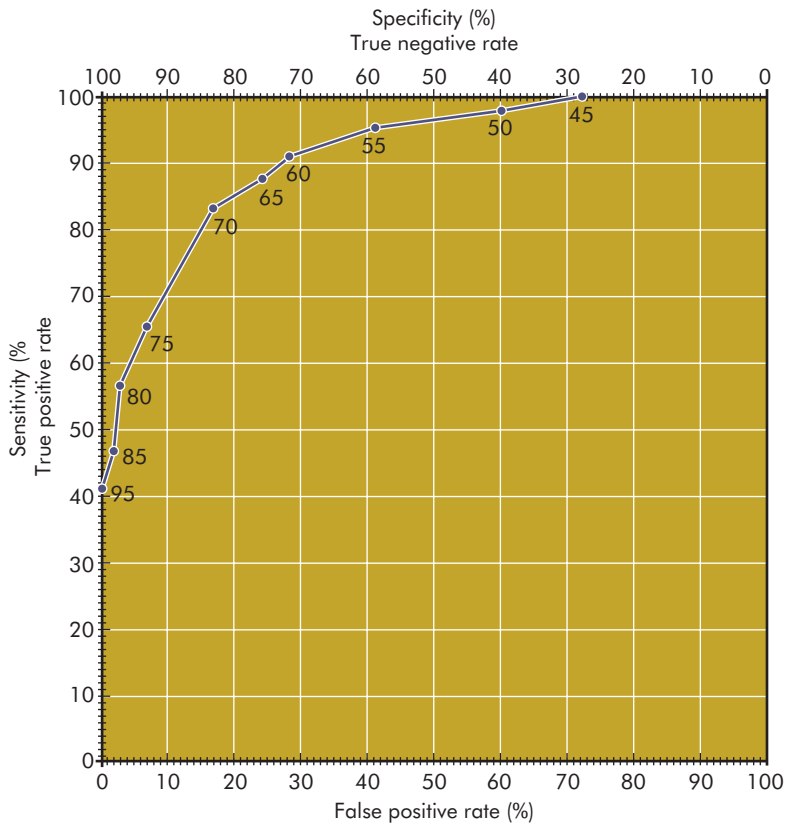
  

RELIABILITY ANALYSIS - SCALE (ALPHA)

Reliability Coefficients 5 items

Alpha = .9120 Standardized item alpha = .9122

**FIGURE 18.2** Reliability analysis for the Postpartum Depression Screening Scale.



**FIGURE 18.3** Receiver operating characteristic (ROC) curve for Postpartum Depression Screening Scale (PDSS): major or minor postpartum depression. Area = 0.91 (SD = 0.03). Used with permission from Beck, C. T., & Gable, R. K. (2001). Further validation of the Postpartum Depression Screening Scale. *Nursing Research*, 50, 161.

the PDSS to screen in PPD cases. With this score, the sensitivity is 41%, meaning that only 41% of the women actually diagnosed with PPD would be identified. A score of 95 has a specificity of 100%, meaning that all cases without an actual PPD diagnosis would be accurately screened out as based on the scale. At the other extreme, a cutoff score of 45 would have 100% sensitivity but only 28% specificity (i.e., 62% false positive), an unacceptable rate of overdiagnosis. Screening instruments that perform well have ROC curves that crowd into the upper left corner, and whose area under the curve is a high proportion of overall space. The area under the curve in Figure 18-3 is .91, which is excellent. Based on these results, Beck and Gable recommended a cutoff score of 60,

which would accurately screen in 91% of PPD cases, and would mistakenly screen in 28% who do not have this mood disorder. In a further analysis, they determined that using this cutoff point would have correctly classified 85% of their sample.

### Research Example Involving Assessment of Qualitative Data

Stubblefield and Murray (2001) examined how parents whose children have undergone lung transplantation perceive their relationships with others before, during, and after the transplantation. Fifteen parents of 12 children who had undergone lung transplantation

were interviewed. Interviews ranged in length from 45 minutes to 2½ hours. The researchers used several strategies to help ensure credibility, including prolonged engagement. The researchers were in contact with the parents over a 9-month period, facilitating their understanding of what it was like to live with a lung transplant.

Investigator triangulation and peer debriefings were additional techniques used to increase credibility. The investigators discussed their descriptive categories as they were developing them each other and with a peer who was an expert in qualitative research. The peer debriefings had two purposes: (1) data analysis decisions were supported and credibility was thereby enhanced, and (2) discussions with the peer provided the researchers with an opportunity for much-needed catharsis after conducting the emotionally difficult interviews. Member checks were done by returning to the parents for a final validating step. Parents were asked how the findings compared with their experiences with the transplantation situation. Data saturation (i.e., obtaining redundant information) was yet another method used to enhance credibility.

With respect to dependability and confirmability, Stubblefield and Murray maintained an audit trail that identified the analytical decisions made during data collection and data analysis. Transferability was facilitated through rich description of the parents' experiences with their children's lung transplantation. Also, the report included information about the demographic characteristics of the parents and children, which can be used to determine transferability of the study findings.

According to Stubblefield and Murray's report, some parents perceived a sense of diminished support from family and friends as they coped with living with the transplant. Parents felt misunderstood and labeled.

## SUMMARY POINTS

- **Measurement** involves the assignment of 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 a quantitative instrument, is the degree of consistency or accuracy with which an instrument measures an attribute. The higher the reliability of an instrument, 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 by **test-retest procedures**.
- The **internal consistency** aspect of reliability, which refers to the extent to which all the instrument's items are measuring the same attribute, is assessed using either the **split-half reliability technique** or, more likely, **Cronbach's alpha method**.
- When the reliability assessment focuses on **equivalence** between observers in rating or coding behaviors, estimates of **interrater** (or **inter-observer**) **reliability** are obtained.
- 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 be measuring.
- **Face validity** refers to whether the instrument appears, on the face of it, to be measuring the appropriate construct.

- **Content validity** is concerned with the sampling adequacy of the content being measured. Expert judgments can be used to compute a **content validity index (CVI)**, which provides ratings across experts of the relevance of items on a scale.
- **Criterion-related validity** (which includes both **predictive validity** and **concurrent validity**) focuses on the correlation between the instrument and an outside criterion.
- **Construct validity** is an instrument's adequacy in measuring the focal construct. One construct validation method is the **known-groups technique**, which contrasts scores of groups presumed to differ on the attribute; another is **factor analysis**, a statistical procedure for identifying unitary clusters of items or measures.
- Another construct validity approach is the **multitrait-multimethod 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.
- 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).
- Sensitivity is sometimes plotted against specificity in a **receiver operating characteristic curve (ROC curve)** to determine the optimum **cutoff point** for caseness.
- A **psychometric assessment** of a new instrument is usually undertaken to gather evidence about validity, reliability, and other assessment criteria.
- There is less agreement among qualitative researchers about criteria to use in enhancing and documenting data quality. The most widely used approach is Lincoln and Guba's method of evaluating the **trustworthiness** of data and interpretations, using the criteria of credibility, dependability, confirmability, and transferability.
- **Credibility** refers to the believability of the data. Techniques to improve the credibility include **prolonged engagement**, which strives for adequate scope of data coverage, and **persistent observation**, which is aimed at achieving adequate depth.
- **Triangulation** is the process of using multiple referents to draw conclusions about what constitutes the truth. The four major forms are **data triangulation**, **investigator triangulation**, **theoretical triangulation**, and **method triangulation**.
- Two important tools for establishing credibility are **peer debriefings**, wherein the researcher obtains feedback about data quality and interpretive issues from peers, and **member checks**, wherein informants are asked to comment on the data and interpretations.
- Credibility can also be enhanced through a systematic search for disconfirming evidence (including a **negative case analysis**), and by having investigators whose credibility is evident through their training and experience.
- **Dependability** of qualitative data refers to the stability of data over time and over conditions, and is somewhat analogous to the concept of reliability in quantitative studies.
- **Confirmability** refers to the objectivity or neutrality of the data. Independent **inquiry audits** by external auditors can be used to assess and document dependability and confirmability.
- **Transferability** is the extent to which findings from the data can be transferred to other settings or groups. Transferability can be enhanced through thick descriptions of the context of the data collection.
- Criteria that have been proposed in a recent synthesis of qualitative validity approaches include credibility, authenticity, criticality, and integrity (primary criteria), and explicitness, vividness, creativity, thoroughness, and congruence (secondary criteria).

## STUDY ACTIVITIES

Chapter 18 of the accompanying *Study Guide for Nursing Research: Principles and Methods, 7th ed.*, offers various exercises and study suggestions for



reinforcing the 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. Suppose the split-half reliability of an instrument to measure attitudes toward contraception was .70. Calculate the reliability of the full scale by using the Spearman-Brown formula.
3. If a researcher had a 20-item scale whose reliability was .60, about how many items would have to be added to achieve a reliability of .80?
4. An instructor has developed an instrument to measure knowledge of research terminology. Would you say that more reliable measurements would be yielded before or after a year of instruction on research methodology, using the exact same test, or would there be no difference? Why?
5. In Figure 18-2, if Beck had wanted a four-item subscale rather than a five-item subscale, which item would she have eliminated?
6. What types of groups do you feel might be useful for a known-groups approach to validating a measures of the following: emotional maturity, attitudes toward alcoholics, territorial aggressiveness, job motivation, and subjective pain?
7. Suppose you were interested in doing an in-depth study of people's struggles with obesity. Outline a data collection plan that would include opportunities for various types of triangulation.
8. Suppose you were going to conduct a grounded theory study investigating parenting of children with attention-deficit/hyperactivity disorder (ADHD). What measures could you take to enhance the credibility of your study?
9. You have been asked to be a peer debriefer for two nurse researchers who are conducting a

phenomenological study on the experiences of women who have been physically abused during pregnancy. What are some questions you could ask these researchers during the debriefing sessions?

## SUGGESTED READINGS

### Methodologic References

- Banik, B. J. (1993). Applying triangulation in nursing research. *Applied Nursing Research*, 6, 47–52.
- Beck, C. T. (1993). Qualitative research: The evaluation of its credibility, fittingness, and auditability. *Western Journal of Nursing Research*, 15, 263–266.
- Beck, C. T. (1994). Reliability and validity issues in phenomenological research. *Western Journal of Nursing Research*, 16, 254–267.
- Berk, R. A. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, 12, 659–670.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Davis, L. L., & Grant, J. S. (1993). Guidelines for using psychometric consultants in nursing studies. *Research in Nursing & Health*, 16, 151–155.
- DeKeyser, F. G., & Pugh, L. C. (1990). Assessment of the reliability and validity of biochemical measures. *Nursing Research*, 39, 314–317.
- Denzin, N. K. (1989). *The research act* (3rd ed.). New York: McGraw-Hill.
- Ferketich, S. L., Figueredo, A., & Knapp, T. R. (1991). The multitrait-multimethod approach to construct validity. *Research in Nursing & Health*, 14, 315–319.
- Fletcher, R. H., Fletcher, S. W., & Wagner, E. H. (1996). *Clinical epidemiology: The essentials*. Baltimore: Williams & Wilkins.
- Giorgi, A. (1989). Some theoretical and practical issues regarding the psychological and phenomenological method. *Saybrook Review*, 7, 71–85.
- Goodwin, L. D., & Goodwin, W. L. (1991). Estimating construct validity. *Research in Nursing & Health*, 14, 235–243.
- Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health*, 20, 269–274.

- Hambleton, R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Hoffart, N. (1991). A member check procedure to enhance rigor in naturalistic research. *Western Journal of Nursing Research*, 13, 522–534.
- Hopkins, K. D. (1997). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn Bacon.
- Hutchinson, S., & Wilson, H. S. (1992). Validity threats in scheduled semistructured research interviews. *Nursing Research*, 41, 117–119.
- Kimchi, J., Polivka, B., & Stevenson, J. S. (1991). Triangulation: Operational definitions. *Nursing Research*, 40, 364–366.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
- Lowe, N. K., & Ryan-Wenger, N. M. (1992). Beyond Campbell and Fiske: Assessment of convergent and discriminant validity. *Research in Nursing & Health*, 15, 67–75.
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35, 382–385.
- Meadows, L. M., & Morse, J. M. (2001). Constructing evidence within the qualitative project. In J. M. Morse, J. M. Swanson, & A. J. Kuzel (Eds.), *The nature of qualitative evidence* (pp. 187–200). Thousand Oaks, CA: Sage.
- Morrison-Beedy, D., Côté-Arsenault, D., & Feinstein, N. F. (2001). Maximizing results with focus groups: Moderator and analysis issues. *Applied Nursing Research*, 14, 48–53.
- Morse, J. M. (1999). Myth # 93: Reliability and validity are not relevant to qualitative inquiry. *Qualitative Health Research*, 9, 717–718.
- Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Patton, M. Q. (2002). *Qualitative evaluation and research methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Rodgers, B. L., & Cowles, K. V. (1993). The qualitative research audit trail: A complex collection of documentation. *Research in Nursing and Health*, 16, 219–226.
- Sim J., & Sharp, K. (1998). A critical appraisal of the role of triangulation in nursing research. *International Journal of Nursing Studies*, 35, 23–31.
- Slocumb, E. M., & Cole, F. L. (1991). A practical approach to content validation. *Applied Nursing Research*, 4, 192–195.
- Thorndike, R. L. (1996). *Measurement and evaluation in psychology and education* (6th ed.). Columbus, OH: Prentice-Hall.
- Tilden, V. P., Nelson, C. A., & May, B. A. (1990). Use of qualitative methods to enhance content validity. *Nursing Research*, 39, 172–175.
- Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative Health Research*, 11, 522–537.
- Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago: Mesa Press.
- Yen, M., & L. L. (2002). Examining test–retest reliability: An intraclass correlation approach. *Nursing Research*, 51, 59–62.

## Studies Cited in Chapter 18

- Albertín-Carbó, P., Domingo-Salvany, A., & Hartnoll, R. L. (2001). Psychosocial consideration for the prevention of HIV infection of injecting drug users. *Qualitative Health Research*, 11, 26–39.
- Beck, C. T. (1992). The lived experience of postpartum depression: A phenomenological study. *Nursing Research*, 41, 166–170.
- Beck, C. T. (1993). Teetering on the edge: A substantive theory of postpartum depression. *Nursing Research*, 42, 42–48.
- Beck, C. T. (1995). Burnout in undergraduate nursing students. *Nurse Educator*, 20, 19–23.
- Beck, C. T. (1996). Postpartum depressed mothers interacting with their children. *Nursing Research*, 45, 98–104.
- Beck, C. T. (2002). Releasing the pause button: Mothering twins during the first year of life. *Qualitative Health Research*, 12, 593–608.
- Beck, C. T., & Gable, R. K. (2000). Postpartum Depression Screening Scale: Development and psychometric testing. *Nursing Research*, 49, 272–282.
- Beck, C. T., & Gable, R. K. (2001). Further validation of the Postpartum Depression Screening Scale. *Nursing Research*, 50, 155–164.
- Bergquist, S., & Frantz, R. (2001). Braden Scale: Validity in community-based older adults receiving home health care. *Applied Nursing Research*, 14, 36–43.
- Brown, S. A., Becker, H. A., Garcia, A. A., Barton, S. A., & Hanis, C. L. (2002). Measuring health beliefs in Spanish-speaking Mexican Americans with type 2 diabetes: Adapting an existing instrument. *Research in Nursing & Health*, 25, 145–158.
- Cacchione, P. Z. (2002). Four acute confusion assessment instruments: Reliability and validity for use in long-term care facilities. *Journal of Gerontological Nursing*, 28, 12–19.

- Carter, B. (2002). Chronic pain in childhood and the medical encounter: Professional ventriloquism and hidden voices. *Qualitative Health Research*, 12, 28–41.
- Chaiyawat, W., & Brown, J. K. (2000). Psychometric properties of the Thai versions of State-Trait Anxiety Inventory for Children and Child Medical Fear Scale. *Research in Nursing & Health*, 23, 406–414.
- Choudhry, U. K. (2001). Uprooting and resettlement experiences of South Asian immigrant women. *Western Journal of Nursing Research*, 23, 376–393.
- Davies, B. L., & Hodnett, E. (2002). Labor support: Nurses' self-efficacy and views about factors influencing implementation. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, 31, 48–56.
- Garvin, B., & Kim, C. (2000). Measurement of preference for information in U.S. and Korean cardiac catheterization patients. *Research in Nursing & Health*, 23, 310–318.
- Gauthier, D. M., & Froman, R. D. (2001). Preferences for care near the end of life: Scale development and validation. *Research in Nursing & Health*, 24, 298–306.
- Holley, S. K. (2000). Evaluating patient distress from cancer-related fatigue: An instrument development study. *Oncology Nursing Forum*, 27, 1425–1431.
- King, G. A., Cathers, T., Polgar, J. M., MacKinnon, E., & Havens, S. (2000). Success in life for older adolescents with cerebral palsy. *Qualitative Health Research*, 10, 734–749.
- Kovach, C. R., & Wells, T. (2002). Pacing of activity as a predictor of agitation. *Journal of Gerontological Nursing*, 28, 28–35.
- Lipson, J. (2001). We are the canaries: Self-care in multiple chemical sensitivity sufferers. *Qualitative Health Research*, 11, 103–116.
- Marsh, G., Prochada, K., Pritchett, E., & Vojir, C. (2000). Predicting hospice appropriateness for patients with dementia of the Alzheimer's type. *Applied Nursing Research*, 13, 187–196.
- McCabe, S., Macnee, C. L., & Anderson, M. K. (2001). Homeless patients' experience of satisfaction with care. *Archives of Psychiatric Nursing*, 15, 78–85.
- Mohr, W. K. (2000). Rethinking professional attitudes in mental health settings. *Qualitative Health Research*, 10, 595–611.
- Phillips, J. M., Cohen, M. Z., & Tarzian, A. J. (2001). African American women's experiences with breast cancer screening. *Journal of Nursing Scholarship*, 33, 135–140.
- Resnick, B., & Jenkins, L. S. (2000). Testing the reliability and validity of the Self-Efficacy for Exercise Scale. *Nursing Research*, 49, 154–159.
- Rew, L. (2000). Acknowledging intuition in clinical decision making. *Journal of Holistic Nursing*, 18, 94–113.
- Ryden, M., Gross, C., Savik, K., Snyder, M., Oh, H., Jang, Y., Wang, J., & Krichbaum, K. (2000). Development of a measure of resident satisfaction with the nursing home. *Research in Nursing & Health*, 23, 237–245.
- Shin, Y., & Colling, K. B. (2000). Cultural verification and application of the Profile of Mood States (POMS) with Korean elders. *Western Journal of Nursing Research*, 22, 68–83.
- Stubblefield, C., & Murray, R. L. (2001). Pediatric lung transplantation: Families' need for understanding. *Qualitative Health Research*, 11, 58–68.
- Williams, J. K., Schutte, D. L., Evers, C., & Holkup, P. A. (2000). Redefinition: Coping with normal results from predictive gene testing for neurodegenerative disorders. *Research in Nursing & Health*, 23, 260–269.
- Woodhouse, L. D., Sayre, J. J., & Livingood, W. C. (2001). Tobacco policy and the role of law enforcement in prevention. *Qualitative Health Research*, 11, 682–692.

