Unit 3

Answering the Research Question: Quantitative Designs

6

Measurement

Goals

- *Analyze the levels of measurement.*
- *Describe the key concepts related to measurement.*
- *Describe knowledge necessary to evaluate adequately the use of instruments.*

Introduction

The goal of research is to provide accurate answers to questions of interest. Questions of interest within nursing are developed from nursing concepts, namely, those thoughts, notions, or ideas that relate to nursing or nursing practice. Nursing concepts are then translated into observable facts or events. Translating a nursing concept into an observable fact or event permits the investigator to measure the event(s) of interest. The key word in relation to measurement is *accuracy*. The selection of a method of sampling influences the degree to which findings can be generalized; the measurement process influences the degree of accuracy of the results.

Concerns related to measurement include the following:

1. What level of measurement is used (**nominal, ordinal, interval, ratio**)?
2. What strategy is used to measure the variables under study (paper-and-pencil questionnaire, interview, etc.)?
3. To what degree can a measurement strategy provide accurate information (**reliability, validity, normative data**)?
4. What factors related to data collection enhance or diminish the accuracy of the results?
5. Given the measurement strategies used in a particular study, what methods of data analysis are most likely to provide accurate information?

This chapter deals with each of these concerns and discusses the importance of appraising the literature from a measurement perspective.

Measuring Variables

The number of variables appropriate for examination by nurses is almost limitless. Any of the physiological and psychosocial concepts that are a part of nursing practice are open to systematic examination. Specific variables that may be identified in nursing research can include age, sex, measures of pain, blood pressure readings, a psychological attribute such as perception of control, and others too numerous to mention (see Figure 6.1). Although a major focus of nursing research at present is the development of a scientific body of knowledge as the underpinnings for practice, in nursing as a whole there is tremendous diversity among the variables appropriate for study.

Regardless of the variables under study, in order to make sense out of the data collected, each variable must be measured in such a way that its magnitude or quantity can be clearly identified. A variety of measurement methods are available for use in nursing research. The specific strategies chosen depend

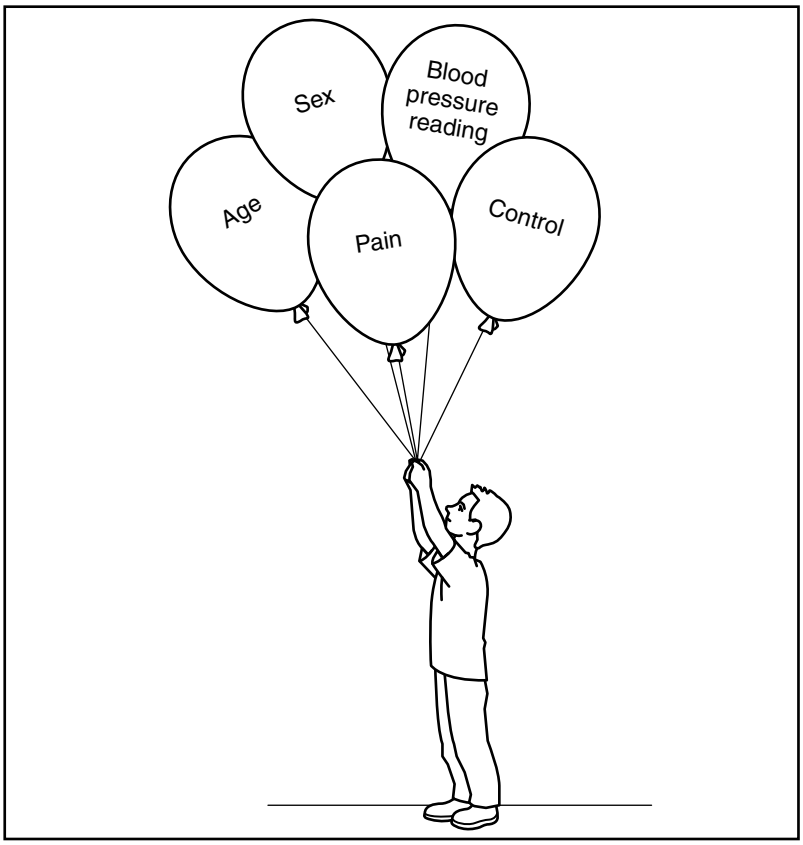

FIGURE 6.1 Variables related to nursing

on the particular research question, the sample under study, the availability of instruments, and the general feasibility of the project. One measurement strategy is not necessarily better than another. Four scales or levels of measurement have been identified: nominal, ordinal, interval, or ratio. Each level of measurement is classified in relation to certain characteristics. When investigators' data fall within the first level of measurement, the range of choice relative to statistical analysis is limited. Data that fall within the fourth level of measurement can be analyzed using a broad array of statistical techniques.

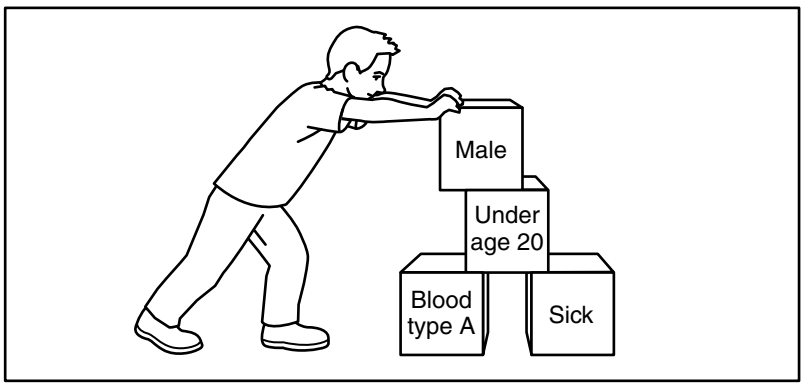

FIGURE 6.2 The nominal level of measurement: Discrete categories

The Nominal Level of Measurement

The first or **nominal** level of measurement is characterized by variables that are discrete and noncontinuous. These variables are categorical and include such examples as sex (male, female), marital status (married, unmarried), blood types (0, A, B, AB), and health state (sick, well). Examples are shown in Figure 6.2.

Expressions commonly used to denote this level of measurement include *nominal scale*, *categorical data*, *nominal data*, and *nominal measurement*. In each instance the expression describes categories that are discrete, named, and therefore mutually exclusive.

WORKING DEFINITION

The Nominal Level of Measurement

The nominal level of measurement is the most primitive method of classifying information. Nominal replies that categories of people, events, and other phenomena are named, are exhaustive in nature, and are mutually exclusive. These categories are discrete and noncontinuous.

EXAMPLE

The Nominal Level of Measurement

A nurse researcher investigated the effects of an educational package dealing with diabetes on the teenage diabetic's compliance with dietary restrictions. The following data were collected three months after the educational presentation:

Exposure to educational package: 65

Compliance. Males, *N* = 38
Good, *N* = 15
Poor, *N* = 23

Females, *N* = 27
Good, *N* = 18
Poor, *N* = 9

To analyze the data, the researcher used a method of statistical analysis appropriate to the nominal level of measurement (chi-square).

In this example compliance is measured in a discrete fashion; that is, the participants were either compliers or noncompliers. A focus of the research was on sexual differences, males versus females. The nominal level of measurement implies that there are named categories–for example, male, female; that the categories are mutually exclusive—for example, male–female, compliers–noncompliers; and that the categories are not ordered in any fashion. There are not, for example, degrees of maleness or femaleness, nor are there degrees of compliance, as defined in this study.

The choice of the appropriate statistical method with which to analyze data is a crucial factor in the eventual worth of the research project. When a nominal level of measurement has been used, the choice of statistical methods that will provide meaningful results is limited.

PRACTICE

Identifying Nominal Variables

From the following research questions, identify those variables that are most likely to be measured using a nominal scale:

- What is the relationship between gender and the incidence of diabetes?
- *Variables:* Gender, incidence of diabetes
- How does paying a shift differential affect absenteeism among nurses on a critical care unit?
- *Variables:* Shift differential, absenteeism
- How do age and sex affect compliance in a cardiac rehabilitation program?
- *Variables:* Age, sex, compliance

CRITICAL APPRAISAL

The Nominal Level of Measurement

If a nominal level of measurement is used, does it seem appropriate given the topic under study?

The Ordinal Level of Measurement

The second or **ordinal** level of measurement is characterized by variables that are assessed incrementally. For example, pain can be measured as slight, moderate, or intense. Exercise can be measured in terms of frequency—that is, often, sometimes, or never. In each case, there are increments or intervals in the scale, but these intervals cannot be considered equal (see Figure 6.3).

As with the nominal variable, it is not some innate value of the specific variable that causes it to be ordinal, but the manner in which it is measured. Pain or exercise could be meas-

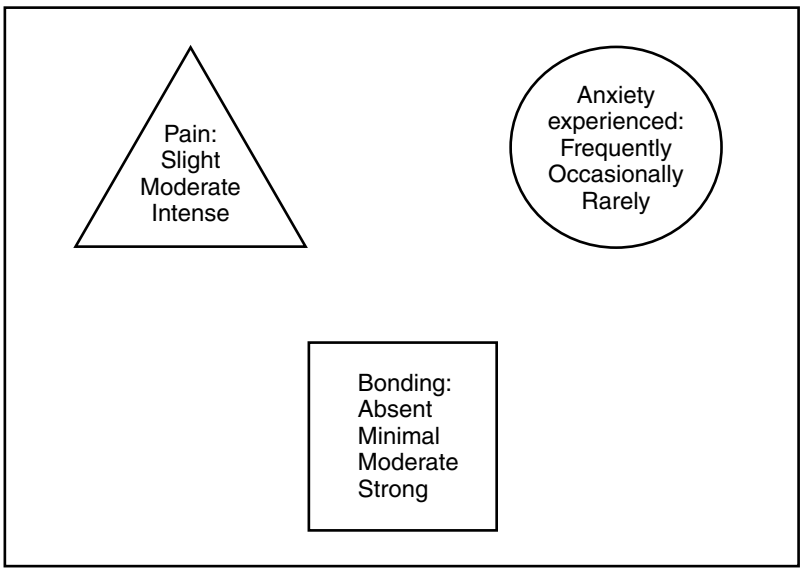

FIGURE 6.3 The ordinal level of measurement: Continuous variables, increments not of equal value; variables can be rank-ordered from highest to lowest

ured using a scale that denotes equal intervals. Exercise could be assessed in terms of the number of miles run or the amount of weight lifted.

Expressions commonly used to denote this kind of measurement include *ordinal scale*, *ordinal variables*, *ordinal data*, and *ordinal measurement*. Unlike the nominal level of measurement, the ordinal level of measurement suggests an ordering of variables. Within the ordinal level of measurement, an event is assigned to a category based on the amount of a particular attribute. This level of measurement can be described as the ranking of events in terms of the relative amounts of a specified characteristic. Pain can be ordered or ranked in terms of the intensity of the experience—namely, intense, moderate, or mild—thus reflecting an ordinal approach. Sex is usually perceived as male or female without an ordering of sexual characteristics, thus reflecting the nominal level of measurement.

WORKING DEFINITION

The Ordinal Level of Measurement

The ordinal level of measurement is second in terms of its refinement as a means of classifying information. *Ordinal* implies that the values of variables can be rank-ordered from highest to lowest.

EXAMPLE

The Ordinal Level of Measurement

A team of nurse researchers designed a research project to compare the degree of social support (SS) perceived by 100 teenage girls diagnosed as bulimics with a group of nonbulimic girls of the same age, educational level, family size, and placement within the family. Each participant responded to a paper-and-pencil test. The following data were collected:

	Considerable SS	**Moderate SS**	**Little SS**
Bulimics	13	30	57
Nonbulimics	35	55	10

To analyze the data, the researcher used a method of statistical analysis appropriate to the ordinal level of measurement (chi-square).

In this example, the variable of social support is rank-ordered. Some participants felt that they had considerable social support; others had moderate social support; and still others had little social support. Even though social support is measured in an incremental fashion (considerable, moderate, little), a consistent number cannot be associated with the distance between considerable, moderate, and little. The data col-

lected, therefore, are ordinal and are somewhat limited in relation to the statistical techniques available for data analysis.

PRACTICE

Identifying Ordinal Variables

How could the following variables be measured at the ordinal level?

Example: Pain—using a scale that measures pain as absent, moderate, strong, or intense.

Variables: Depression, skin turgor, helplessness, grief, stress, confusion, nausea.

CRITICAL APPRAISAL

The Ordinal Level of Measurement

If an ordinal level of measurement is used, does it seem appropriate given the topic under study?

The Interval Level of Measurement

The third or **interval** level of measurement is characterized by a scale that is quantitative in nature. Increments on the scale can be measured, and they are equidistant. An individual's temperature is measured in terms of numbers of degrees. Between a temperature of 97 and 100 degrees Fahrenheit, there are three equal increments of one degree each (Figure 6.4).

An interval scale, though quantitative in nature, does not have an absolute or actual zero. Temperature is a good example of an interval scale in that zero changes depending on the scale used—Fahrenheit or Celsius. Expressions commonly used

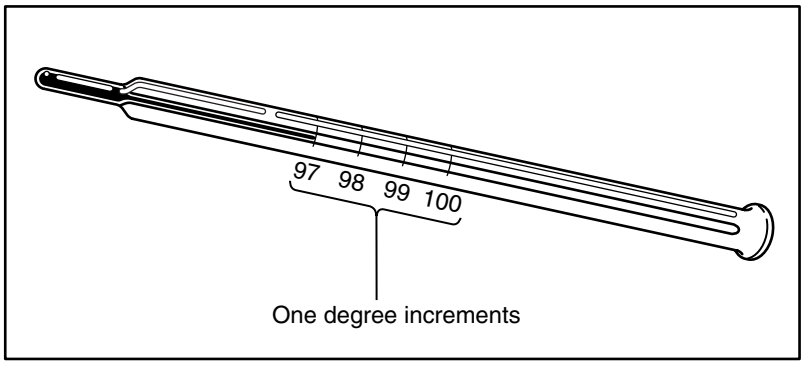

FIGURE 6.4 The interval level of measurement: Continuous variables, increments of equal value

to denote this kind of measurement include *interval scale*, *interval variables*, *interval data*, and *interval measurement*.

WORKING DEFINITION

The Interval Level of Measurement

Interval measurement refers to the third level of measurement in relation to the complexity of statistical techniques that can be used to analyze data. Variables within this level of measurement are assessed incrementally, and the increments are equal. Many statistical techniques can be used to analyze interval variables.

EXAMPLE

The Interval Level of Measurement

A nurse researcher investigated the psychological status of 60 abused women before and after an intervention designed to enhance their feelings of independence. A psychological test was given before and after the intervention, and test scores

were then compared using a paired *t*-test. Examples of the data collected are given below:

Abused woman	Before	After	Difference
1	36	42	6
2	25	40	15
3	39	45	6
4	40	40	0
5	41	44	3
6	35	40	5

The scores on the psychological test represent interval data. To analyze this data the researcher could choose any one of many statistical methods appropriate for use with interval data.

PRACTICE

Identifying Interval Variables

Identify those variables that are most likely to be measured using a nominal scale, an ordinal scale, or an interval scale:

- How does assertiveness training affect the lifestyle patterns of abused women?

  Variables: Assertiveness training, lifestyle patterns

- What effect does health education have on compliance with a low-fat diet on men diagnosed with hypertension?

  Variables: Health education, compliance

- How does extended visitation on the neonatal unit affect the bonding process between father and infant?

  Variables: Extended visitation, bonding process

CRITICAL APPRAISAL

The Interval Level of Measurement

If an interval level of measurement is used, does it seem appropriate given the topic under study?

The Ratio Level of Measurement

The fourth or **ratio** level of measurement is characterized by variables that are assessed incrementally with equal distances between the increments and a scale that has an absolute zero. Obvious ratio scales include time, length, and weight (Figure 6.5).

Even though the variables of time, length, and weight are obvious examples of ratio scales, they can also be measured using a nominal or ordinal scale. For example, the nurse researcher who chooses to measure the time it takes nursing students to perform a given task could divide time into two categories: less than five minutes and more than five minutes. These two categories could then be represented by Time 1 and Time 2, nominal variables necessitating statistical analysis appropriate to that method of measurement.

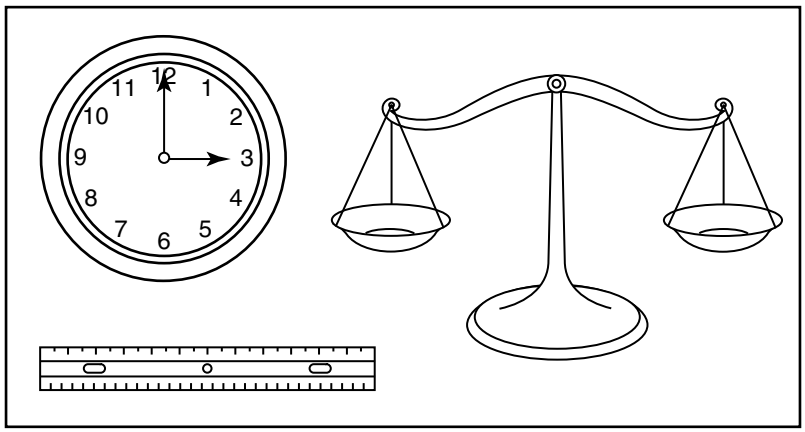

FIGURE 6.5 The ratio level of measurement: Continuous variables, increments of equal value, absolute zero exists

WORKING DEFINITION

The Ratio Level of Measurement

The ratio level of measurement is the fourth and least primitive method of classifying information. *Ratio* implies that the variables reflect the characteristics of ordinal and interval measurement and can also be compared by ratios; that is, the number representing a given variable can be compared by describing it as two or three times another number or as one-third, one-quarter, and so on. The ratio level of measurement, unlike the other three levels, has an absolute zero.

EXAMPLE

The Ratio Level of Measurement

A team of nurse investigators designed a research project to determine the effect of a support group on the weight loss of 35 morbidly obese women motivated to comply with a specific diet. Both the experimental and control groups were weighed one week before the initiation of the support group, and participants in both groups agreed to try the specified diet. The following data were collected and data analysis completed using statistical procedures appropriate to the ratio level of measurement.

	Experimental group	**Control group**
Sample size	35	35
Mean weight loss	26 lbs.	19 lbs.
Standard deviation	2 lbs.	4.5 lbs.

The advantage of interval and ratio levels of measurement compared with nominal and ordinal levels of measurement is

that more complex statistical techniques can be used to analyze the data collected. The result of having an increased number of options for the analysis of data is that the description of what has actually happened in the study is more precise.

PRACTICE

Identifying Ratio Variables

Formulate four research questions and incorporate the four levels of measurement in the methods of measurement proposed.

Example: What is the relationship of teenage anorexia to weight, sex, emotional status, and family structure?

Variables and proposed measurement level:

Weight—ratio

Sex—nominal

Emotional status—interval

Family structure—ordinal

CRITICAL APPRAISAL

The Ratio Level of Measurement

If a ratio level of measurement is used, does it seem appropriate given the topic under study?

Methods of Measurement

The Instrument

The term **instrument** is used in research to describe a particular method of collecting data. Instruments used in research include paper-and-pencil tests, structured interviews, puzzles, mechanical equipment, direct observations, and the like. Instruments may provide a total score or may consist of a

series of items that must be analyzed independently. The instrument is the device that is used to measure the concept of interest. The construction of an instrument is a complex and time-consuming undertaking. The development of an effective instrument requires an in-depth knowledge of the content area under study and considerable skill in the area of measurement theory. Beginning researchers are advised to use instruments that have already been developed and have had their effectiveness established wherever possible. Consumers of research are advised to carefully appraise the instruments used when reviewing a research report.

Whether the instrument is a paper-and-pencil test, a piece of equipment, or a direct observation, there are characteristics of the instrument that are important in terms of the accuracy and meaning of the results. The two major characteristics that are essential in relation to the meaning and accuracy produced by a given instrument are **validity** and **reliability**. When appraising the literature or designing a study, it is important to evaluate the validity and reliability of the methods of measurement used in the research.

WORKING DEFINITION

An Instrument

An instrument is a device used to measure the concept of interest in a research project. An instrument may be a paper-and-pencil test, a structured interview, or a piece of equipment.

When measuring attributes of interest, a norm-referenced or a criterion-referenced approach may be used. The norm-referenced approach is one in which a participant's performance can be evaluated against the performance of others in similar circumstances. For example, many of the achievement tests taken by high school students have been taken by thousands of students, and statistical information has been gathered that describes the distribution of scores among that group of individuals. Any individual taking the test can then be measured against similar individuals.

The criterion-referenced-approach involves the identification of certain attributes of interest—for example, a set of standards for a nursing intervention—and then the construction of an instrument to measure how well participants meet the standards. In this instance, comparing participants' performance without standards would be irrelevant (Cozby, 2000).

EXAMPLE

An Instrument

A team of nurse investigators plan to examine the association between attitudes toward health and the performance of routine breast self-examination. The investigators choose a paper-and-pencil test (the instrument) that requires individuals to answer "agree" or "disagree" to 25 items designed to measure attitudes toward health. A total score on attitudes toward health is calculated. Participants are also asked to identify whether they perform routine breast self-examination.

PRACTICE

Identifying instruments

Identify five instruments used in various nursing studies, and describe the differences in their approach to measurement by answering the following questions:

1. Does the investigator want to compare the results with a set standard?
2. Does the investigator want to compare the results of one group with another group?
3. Does the investigator design the instrument specifically for a particular study?
4. Is the instrument one that has been used many times in a variety of studies?

5. Does the instrument provide a total score, or is it a series of items that must be analyzed independently?

Errors of Measurement

Whenever a variable is measured, there is the potential for errors to occur. Some of the factors that can influence the outcome can be controlled, others cannot. The score obtained from using a particular instrument in a particular setting consists of two parts—the true score and the error. A concerted effort should be made to limit the error portion of the score.

The following list consists of potential sources of error when measures of specific attributes are taken.

1. *Instrument clarity:* Frequently, participants will respond to an instrument inappropriately—for example, placing a check-mark in a box when "yes" or "no" was required. If the instructions for taking the instrument are not clear, participants cannot respond appropriately and the information received is not accurate. Similarly, if the items themselves are not readily understood, responses may not reflect the participants' perceptions and the resultant information is of limited value.
2. *Variations in administration:* If some participants are allowed to respond to an instrument at their leisure while others are given time constraints, the information received is not comparable. If some participants are assisted in responding to an instrument while others are not, again, the information received may be different.
3. *Situational variations:* If an instrument is administered under differing environmental conditions—unpleasant versus pleasant surroundings, or threatening versus nonthreatening conditions—responses may vary according to the situation.
4. *Response set biases:* Frequently, participants will give an answer that is socially desirable. For example, questions about an individual's sexual practices, views on religion, or politics may produce responses that are untrue but are chosen because they are considered acceptable by the

majority of individuals in society. Another problem is the tendency of some individuals to consistently respond in an extreme fashion. On a scale of I to 5, some people will consistently respond at one end of the scale, regardless of the topic.

5. *Transitory personal factors:* Participants' mood, state of mind, and level of stress at the time of responding to the instrument may affect either their answers or their desire to participate in the project.
6. *Response sampling:* The content of the instrument—that is, the sampling of items—may affect the participant's score. Depending on the items selected, a nurse might perform well on a pain knowledge questionnaire or poorly on the same questionnaire.
7. *Instrument format:* The order of items on a questionnaire and the kind of questions asked (open-ended or closed) can affect the responses given.

Validity

The concept of **validity** in relation to research is a judgment regarding the degree to which the components of research reflect the theory, concept, or variable under study (Streiner & Norman, 1996). The validity of the instrument used (how well it measures what it is supposed to measure) and the validity of the research design as a whole are important criteria in evaluating the worth of the results of the research conducted.

Instruments may not be designed in such a way that the concept under study is reflected in the items. For example, a group of nurse researchers may want to design an instrument that measures stress. In examining the literature related to stress, it may become apparent that the items are really assessing anxiety rather than stress. In this instance the instrument would not be valid. It does not measure what the investigators want it to measure.

In relation to the overall research design, the term validity can refer to the likelihood that the experimental manipulation was indeed responsible for the differences observed. This kind of validity is termed *internal validity*. Another kind of validity

used in relation to the research design is termed *external validity* and refers to the extent to which the results of a study can be generalized to the larger population (Polit & Hungler, 1999).

The Validity of an Instrument

Four types of validity are used to judge the accuracy of an instrument: (1) content validity, (2) predictive validity, (3) concurrent validity, and (4) construct validity. Using these categories, it is important for the consumer of research and the investigator designing a study to evaluate the instruments used.

Content validity is a judgment regarding how well the instrument represents the characteristics to be assessed. Instruments with a high degree of content validity are as representative as possible of all the items that could be included to measure the concept under study. For example, to design a questionnaire on individuals' attitudes toward eating but forget to ask anything about the importance of food in their lives would be an oversight. The content validity of the instrument would be in question.

Judgments about content validity are subjective. Objective methods for measuring the areas of concern that must be reflected on an instrument are not available. Judgments are generally based on prior research in the field and on the opinions of experts.

Predictive validity is a judgment as to the degree to which an instrument can accurately forecast the future. For example, certain achievement tests can predict the academic futures of students. Some personality tests can predict behavior patterns. The assessment of predictive validity is an objective task and can be accomplished through comparing one instrument with another of known predictive validity, or through examining the outcome in terms of available data. For example, a lifestyle questionnaire aimed at identifying women who are likely to develop osteoporosis could be evaluated in terms of its predictive validity by following identified individuals to see whether they indeed develop the disease.

Concurrent validity is a judgment as to the degree to which an instrument can accurately identify a difference in the present. Such questions as "Can the instrument accurately determine the difference between the learning disabled child and the

normal child?" or "Can the piece of equipment identify those individuals who are hypertensive?" reflect the degree of concurrent validity of a particular instrument. Concurrent validity suggests that the instrument in question can indicate a specific behavior or characteristic in the present. Judgments can be made through assessing the literature and using multiple criteria to examine results.

Construct validity refers to the extent to which a participant actually possesses the characteristic under study. For example, nurse investigators may choose to examine the concept of pain among children diagnosed with leukemia, and they may want to design an instrument to measure pain in this setting. They will want to know if their instrument is actually measuring pain (construct validity) or if it is in fact assessing something else (e.g., anxiety). In order to determine whether their instrument is measuring pain, these investigators can use multiple measures to assess the same construct, namely, a questionnaire, a structured interview, and direct observation. In judging construct validity, theoretical considerations come into play. The investigator needs to be able to make predictions about the construct in terms of other related constructs. For example, the degree of pain expressed may be a function of the kind of social support perceived.

The validity of an instrument (how well it measures what it is supposed to measure) is essential to the success of any research endeavor. If the investigator designs a study to examine parenting skills but uses an instrument that is actually measuring general coping skills, the results of the study are of little value. Similarly, if an investigator decides to examine the relationship between effective nursing care and a particular kind of nursing education, and yet the instrument used actually measures attitudes toward nursing as a profession, the results can be either misleading or meaningless.

WORKING DEFINITION

The Validity of an instrument

Validity describes the usefulness of an instrument given the context in which it is applied. It reflects how well an instru-

ment has measured what it was supposed to measure given a particular set of circumstances. The following kinds of validity are frequently assessed in trying to make a judgment about a given instrument.

Content: All important areas of concern are reflected.

Predictive: Events are accurately predicted.

Concurrent: Accurate differences are shown in the present.

Construct: The attribute of interest is actually being measured.

The characteristic of validity is one that frequently takes years to ascertain. As instruments are used and tested repeatedly, information about exactly what they are measuring can be gathered and assessed. Unfortunately, nursing does not have a long history of instrument development, and therefore it is frequently difficult to find an instrument measuring a nursing concept of interest that has adequate data on its validity.

EXAMPLE

Assessing the Validity of an Instrument

A team of nurse investigators design a lifestyle questionnaire to identify those women who are at risk of developing osteoporosis. In order to validate the instrument, they take the following steps:

1. Obtain from the literature, as well as from experts in osteoporosis, those lifestyle factors that are believed to predispose an individual to osteoporosis (content validity).
2. Incorporate all lifestyle factors believed to predispose individuals to osteoporosis within the instrument (content validity).
3. Involve experts in the field of osteoporosis in the formulation of the instrument (content validity).
4. Give the instrument to a randomly selected sample of individuals and follow them to see if those identified as at risk do, in fact, develop osteoporosis (predictive validity).

5. Give the instrument to a randomly selected group of individuals along with another accepted measure to see if this instrument identifies the same individuals as being at risk (concurrent validity).
6. Give the instrument to a group of women who have osteoporosis and a group of women who do not have osteoporosis (construct validity).

Designing an instrument that is capable of accurately measuring a specific concept takes a considerable amount of effort and time. When an instrument does not accurately measure what it is supposed to measure, the results of a study can be misleading or meaningless. An important process in appraising the literature or in designing a study is the assessment of the instruments used in relation to their validity.

PRACTICE

Assessing the Validity of an instrument

Select three published nursing studies and identify the information given in relation to the validity of the instruments used. List possible concerns about the instruments used based on the information provided.

CRITICAL APPRAISAL

Assessing the Validity of an Instrument

- Is sufficient information available in regard to the instrument's validity; that is, does it seem to measure what it purports to measure?
- Does the instrument cover all of the important factors related to the topic?
- Is there evidence that using this instrument to collect data can help the investigator to predict an occurrence in the future?

- Is there evidence that the instrument can identify the desired characteristic(s) in the present?
- Does the instrument measure the construct that it claims to measure?

Reliability

The **reliability** of an instrument reflects its stability and consistency within a given context. For example, a scale developed to measure attitudes toward pain among the elderly might not be reliable if used with young adults. Because the consistency and stability of responses to questions asked is such an important concept, instruments must be evaluated for their reliability.

Three characteristics of reliability are commonly evaluated: (1) stability, (2) internal consistency, and (3) equivalence (Figure 6.6). *Stability* refers to the degree to which research participants' responses change overtime. Ideally, we would like participants to respond to an instrument measuring their self-esteem in a similar fashion on any number of occasions unless intervening events occur to change their perceptions.

Stability is measured by giving the same individuals an instrument on two occasions within a relatively short period of time (two to three weeks apart is often suggested) and then examining their responses for similarities. This method for determining reliability is termed test–retest. A correlation coefficient is calculated to determine how closely participants' responses on the second occasion matched their responses on the first occasion. Reliability coefficients range from −1 to +1.00. A coefficient of .60 shows a moderately strong relationship, .20 a weak relationship, and 0 no relationship. We rarely find perfect relationships (1.00).

If a group of adolescents respond to a questionnaire that asks them to examine their level of confidence on two occasions and the resulting correlation coefficient is .15, we know that the construct of level of confidence is not stable over time as measured by that instrument. Problems with the notion of test-retest as a measure of reliability include the fact that some may respond to the instrument the second time on the basis of

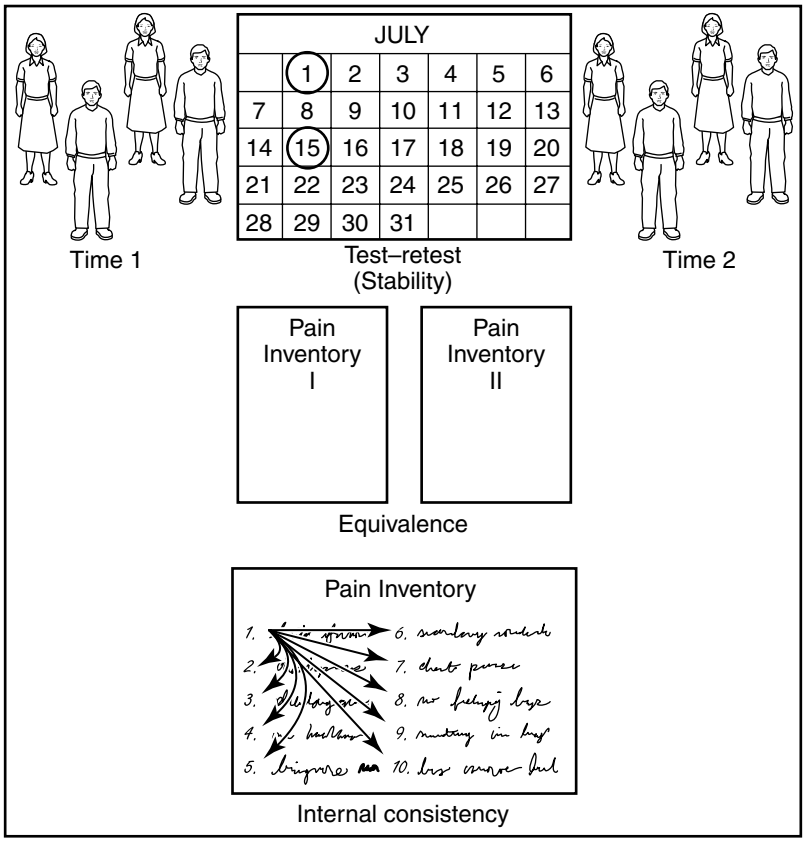

FIGURE 6.6 Three types of reliability

their memory of their first exposure to the instrument; participants may change as a result of responding to the instrument the first time; and, knowing that they have already responded to the instrument, participants may not answer questions carefully. In addition, an assessment of test–retest reliability is not useful for constructs that we know change over time, such as pain, anxiety, and anger.

Internal consistency is a measure of reliability that is frequently used with scales designed to assess psychosocial characteristics. An instrument is describe as internally consistent to the extent that all of its subparts measure the same characteris-

tic. This method of reliability assessment deals with error made in relation to the sampling of items. If an instrument is designed to assess self-esteem and a few of the items actually measure depression, individuals are likely to answer those questions differently. The instrument as a whole will not be internally consistent.

Instruments can be assessed for internal consistency using a split-half technique (i.e., answers to one half of the items are compared with answers to the other half of the items) or by calculating an alpha coefficient or using the Kuder-Richardson formula. In the case of the alpha coefficient and the Kuder-Richardson formula, a coefficient that ranges from 0 to 1.00 generally results.

The notion of *equivalence* is often a concern when different observers are using the same instrument to collect data at the same time. For example, three observers may be using a checklist to identify the mood states of preschool children. Each observer needs to understand what constitutes the characteristics listed—happy, sad, angry, and so forth. In this instance, interrater reliability is calculated on information gathered by the various observers. When appropriate, a coefficient can be calculated or other statistical and nonstatistical procedures can be used.

An important practical note regarding the assessment of reliability is that the testing of an instrument is done before the study is initiated and on individuals not participating in the study. If a few individuals participated in a test–retest of an instrument and were then included in the study they would have responded to the instrument on three occasions, and other participants would have only one chance to respond.

WORKING DEFINITION

Reliability

Reliability is a characteristic of an instrument that reflects the degree to which the instrument provokes consistent responses.

There are three characteristics of reliability that are commonly evaluated:

Test–retest: Degree of consistency when individuals respond on two separate occasions

Equivalence: Degree of consistency in providing measurements of same attributes

Internal consistency: Degree of consistency among responses to items

EXAMPLE

Establishing Reliability

A team of nurse investigators designs an instrument to identify the population at risk for developing eating disorders. They review the literature and work closely with experts in the field while formulating the items. Their questionnaire contains 10 items, and participants are asked to respond using a 5-point scale.

Sample Item:

(Scale: frequently = 5, often = 4, sometimes = 3, rarely = 2, never = 1)

I am critical of my appearance 5 4 3 2 1

The research team takes the following steps to establish the reliability of the instrument:

- Ten volunteers from a support group for bulimics respond to the items at the same time and place and then repeat the experience two weeks later. A correlation coefficient is calculated, and the test–retest reliability is considered adequate.
- Ten volunteers from a different support group for bulimics respond to the items at the same time and place. An alpha

coefficient is calculated, and the internal consistency of the instrument is considered adequate.

PRACTICE

Reliability

- Identify a nursing concept of interest and develop three related items. Administer the three items to a small group of people on two separate occasions and examine the results for consistency (visually examine the responses).
- Find two studies published in nursing journals that give an acceptable description of the reliability of the instrument used.

CRITICAL APPRAISAL

Reliability

Is sufficient information available in regard to the instrument's reliability; that is, is there consistency over time, is there consistency relative to a parallel form of the instrument, and is there internal consistency?

Critical Overview of Measurement

The manner in which variables are measured can make the difference between useful and meaningless results. When instruments used are of questionable validity and reliability, the data collected cannot be considered accurate. In addition to the concepts of validity and reliability, a knowledge of the various levels of measurement is important because appropriate statistical analysis is dependent on such an understanding.

CRITICAL OVERVIEW MEASUREMENT

Two nurse investigators have identified the following research question: What are the predisposing characteristics of osteoporosis for women 19 years of age? They identify three factors as important to their study: diet, exercise, and family incidence.

Activity 1

Given the variables of diet, exercise, and family incidence, identify which levels of measurement are possible.

Activity 2

Formulate four items related to each of the following: diet, exercise, and family incidence. Divide the items in half (two on each topic on separate sheets of paper) and give the two forms of the questionnaire to two different groups of students. Scan the responses within each category to see if they are similar.

References

Cozby, P. C. (2000). *Methods in Behavioral Research* (7th Edition). Toronto: Mayfield Publishing Co.

Polit, D. F. & Hungler, B. P. (1999). *Essentials of nursing research* (6th Edition). Philadelphia, J. B. Lippincott.

Streiner, D. & Norman, G. (1996). *PDQ Epidemiology* (2nd Edition). St. Louis: Mosby.