CHAPTER 4

Conceptualization and Measurement

From Concepts to Observations 67
- *Concepts and Variables* 68
- *Operationalization* 69
- *Using Scales to Measure Variables* 70
- *Treatment as a Variable* 72
- *From Observations to Concepts* 72
- *Gathering Data* 73
- *Combining Measurement Operations* 73

Levels of Measurement 74
- *Nominal Level of Measurement* 74
- *Ordinal Level of Measurement* 75
- *Interval Level of Measurement* 77
- *Ratio Level of Measurement* 77
- *The Case of Dichotomies* 78
- *Mathematical Comparisons* 78

Measurement Error 79

How to Assess Measurement Accuracy 80
- *Measurement Reliability* 80
  - Test-Retest Reliability 81
  - Internal Consistency 81
  - Alternate-Forms Reliability 81
  - Interrater Reliability 81
  - Intrarater Reliability 82
- *Measurement Validity* 82
  - Face Validity 82
  - Content Validity 82
  - Criterion Validity 83
  - Construct Validity 83
- *Reliability and Validity of Existing Measures* 84

Using Scales to Identify a Clinical Status 85

Measurement in a Diverse Society 86

Measurement Implications for Evidence-Based Practice 87

Conclusion 89

Key Terms 89

Highlights 90

Discussion Questions 90

Critiquing Research 90

Making Research Ethical 91

Developing a Research Proposal 91

Web Exercises 91

Measurement is a crucial component of social work practice and research. When you think of measurement in social work practice, you typically think of assessment whereby you are collecting information about a client system; the assessment often includes key concepts and measures of these concepts about which you are collecting information. When evaluating a program's outcomes, broadly stated goals and objectives are translated into something that can be measured. What you learn from the assessment helps guide intervention decisions with clients; what you learn about a program's outcomes influences the design or continuation of the program. Therefore, the decisions you make about how to measure a client's status or a program's outcomes are critical.

Similarly, in reviewing or designing a research study, how key concepts are defined and measured is important in order to evaluate the validity of the research. Judgments about the evidence to support a particular intervention are not just about the demonstration of successful outcomes but also entail considerations about the quality of the measures of these outcomes.

Whether for practice or for research, you will have to answer three questions: (1) What do the main concepts mean? (2) How are the main concepts measured? (3) Is the measurement method accurate and valid? In this chapter, we review each of these questions. We first address the issue of conceptualization, or how you define key terms. We then discuss the levels of measurement reflected in different measures. This section is followed by a discussion of measurement error. Next, we discuss different methods to assess the quality of measures. Finally, we consider the implications of measurement for diverse population groups and evidence-based practice. By the chapter's end, you should have a good understanding of measurement and the crucial role it plays in social work research.

2 From Concepts to Observations

In 2011, 46.2 million people, 15.0% of the U. S. population, lived in poverty (DeNavas-Walt, Proctor, & Smith, 2012). What does poverty mean? The Official Poverty Line definition used in this report is conceptualized as an absolute standard, based on the amount of money required to purchase an emergency diet adequate for about two months multiplied by three. But other social scientists reject the notion that a poverty measure should be based on an emergency diet and suggest that poverty means having sufficient income to purchase adequate amounts of goods such as housing, food, shelter, transportation and the like in a particular geographical region (Lin & Bernstein, 2008). Still other researchers disagree with absolute standards and have urged adoption of a relative poverty standard, defining those persons who live in poverty based on their incomes relative to the general population. In fact, the term *poverty* means different things to different people and its measurement has always been somewhat controversial. These discussions are important because different notions about poverty shape estimates of how prevalent it is and what can be done about it.

We refer to terms like poverty as a **concept**, that is, a name for an image that summarizes a set of similar observations, feelings, or ideas. Many topics studied by social work researchers involve abstract concepts or ideas, not just simple objects. Some concepts are relatively straightforward and there is little confusion about their meaning; the concept *age* can readily be defined as "years since birth." When we refer to concepts like *homelessness*, *poverty*, or *community empowerment*, we cannot count on others knowing exactly what we mean. Even the experts may disagree about the meaning of frequently used concepts, just as we saw with the different definitions of poverty. That's okay. The point is not that there can be only one definition of a concept, but that we have to specify clearly what we mean when we use a concept.

Concept A mental image that summarizes a set of similar observations, feelings, or ideas.

So **conceptualization**—working out what your key terms will mean in your research—is a crucial part of the research process. Conceptualization is the process of matching up terms to definitions of the terms. Since many concepts of interest are abstract, we often examine social theory and prior research to review appropriate definitions. We may need to identify the different dimensions or aspects of the concept. We should understand how the definition we choose fits within the theoretical framework guiding the research and what assumptions underlie this framework.

Researchers start with a **nominal definition** by which the concept is defined in terms of other concepts. Nominal definitions are like the definitions found in dictionaries: You get an understanding of the word and its dimensions but you still do not have a set of rules to use to measure the concept. For example, child abuse might be defined as evident when either severe physical or emotional harm is inflicted on a child or there is contact of a sexual nature. The nominal definition of child abuse includes concepts such as *severe harm*, *physical abuse*, and *emotional abuse*, but the definition does not provide the set of rules to identify the forms of abuse or distinguish between severe and not severe harm. The actual measures of child abuse should be consistent with the nominal definition.

Conceptualization The process of specifying what we mean by a term. In deductive research, **conceptualization** helps to translate portions of an abstract theory into testable hypotheses involving specific variables. In inductive research, conceptualization is an important part of the process used to make sense of related observations.

Concepts and Variables

After we define the concepts in a study, we must identify corresponding variables and develop procedures to measure them. For example, we might be interested in the concept *substance abuse,* which is defined in the *DSM-IV-TR* as the "repeated use of a substance to the extent that it interferes with adequate social, vocational, or self-care functioning" (APA, 2004). We could convert this concept to any number of variables. One variable might be the count of alcoholic drinks; another variable might involve asking about the presence of blackouts; a third variable may ask about binge drinking; and a fourth variable might reflect a score on a rating scale of 10 questions. Any of these variables could show low or high degrees of substance abuse. If we are to study variation in substance abuse, we must identify the variables to measure that are most pertinent to the research question.

Where do variables fit in the continuum from concepts to operational indicators? Think of it this way: Usually, the term *variable* is used to refer to some specific aspect of a concept that varies and for which we then have to select even more concrete indicators. Concepts vary in their level of abstraction, and this in turn affects how readily we can specify the variables pertaining to the concept. We may not think twice before we move from a conceptual definition of age as time elapsed since birth to the concrete indicator, *years since birth.* Binge drinking is also a relatively concrete concept, but it requires a bit more thought. We may define binge drinking conceptually as episodic drinking and select for our research on binge drinking the variable, *frequency of five or more drinks in a row.* A single question is sufficient.

A very abstract concept like social status may have a clear role in social theory but a variety of meanings in different social settings. Variables that pertain to social status may include level of esteem in a group, extent of influence over others, level of income and education, or number of friends. It is very important to specify what we mean by an abstract concept like social status in a particular study and to choose appropriate variables to represent this meaning.

Not every concept in a particular study is represented by a variable. If we were to study clients' alcohol abuse at an inpatient treatment unit, there is no variation, rather all the clients are clients. In this case, client is called a **constant**; it is always the same and therefore, is not a variable. Or course, this does not mean we cannot study differences, such as gender, among the clients. In this case, gender is the variable and client is still a constant.

It's very tempting to try to measure everything by including in a study every variable we can think of that might have something to do with our research question. This haphazard approach will inevitably result in the collection of data that are useless and the failure to collect some data that are important. In choosing variables for a study, examine relevant theories to identify key concepts, review prior research to learn how useful different indicators have been, and assess the resources available for adequately measuring variables in the specific setting to be studied.

Operationalization

Once we have defined our concepts in the abstract—that is, we have provided a nominal definition—and we have identified the specific variables we want to measure, we must develop measurement procedures. The goal is to devise operations, that is, procedures to indicate the values of cases on a variable. **Operationalization** is the process of specifying the operations that will indicate the value of cases on a variable.

Exhibit 4.1 represents part of the operationalization process in three studies. The first researcher defines the concept, income, and chooses one variable, annual earnings, to represent it. This variable is then measured with responses to a single question, or item: "What was your total income from all sources in 2012?" The second researcher defines the concept, poverty, as having two aspects or dimensions, subjective poverty and absolute poverty. Subjective poverty is measured with responses to a survey question: "Do you consider yourself poor?" Absolute poverty is measured by comparing family income to the poverty threshold. The third researcher decides that the concept, social class, is defined by the sum of measurements of three variables: income, education, and occupational prestige.

Exhibit 4.1 Concepts, Variables, and Indicators

Concepts	Variables	Indicators
Binge drinking	Frequency of heavy episodic drinking	"How often within the past two weeks did you consume five or more drinks containing alcohol in a row?"
Poverty	Subjective poverty	"Would you say you are poor?"
	Absolute poverty	Family income ÷ Poverty threshold
Socioeconomic status	Income	
	Education	Income + Education + Prestige
	Occupational prestige	

One consideration is the precision of the information that is necessary. The first researcher in Exhibit 4.1 is seeking information that is quite precise. She assumes that respondents will be able to accurately report the information. As an alternative, she might have asked respondents: "Please identify the income category that includes your total income from all sources in 2012." For this question, she will get less exact information. Generally, the decision about precision is based on the information that is needed for the research. It may also be based on what the researcher believes people can recall and the content people may be willing to report.

The variables and particular measurement operations chosen for a study should be consistent with the purpose of the research question. Take the evaluative research question: Are self-help groups more effective in increasing the likelihood of abstinence among alcohol abusers than hospital-based treatments? We may operationalize the variable, *form of treatment* in terms of participation in these two types of treatment, self-help or hospital based. However, if we are answering the explanatory question, "What influences the success of alcohol abuse treatment?" we should probably consider what it is about these treatment alternatives that is associated with successful abstinence. Prior theory and research suggest that some of the important variables that differ between these treatment approaches are level of peer support, beliefs about the causes of alcoholism, and financial investment in the treatment.

Researchers provide an **operational definition**, which includes what is measured, how the indicators are measured, the rules used to assign a value to what is observed, and how to interpret the value. An operational definition for alcoholism might include the following content:

> The Michigan Alcoholism Screening Test (MAST) is a 24-item instrument that includes a variety of indicators of symptoms such as seeing drinking as a problem, seeking treatment for problem drinking, delirium tremens, severe shaking, hearing voices, complaints from others about drinking, memory loss from drinking, job loss due to drinking, social problems from drinking, arrests for drunk driving or for drunken behavior, guilt feelings about drinking, and ability to stop drinking. The scale may be administered orally or may be self-administered. Respondents respond yes or no to each item and each item is given a weighted score ranging from 0 to 5. There are four items for which the alcoholic response is "no." The weighted item responses are summed, with a score of 0 to 3 indicating no problem with alcoholism, 4 considered to be suggestive of a problem, and 5 or above an indication of alcoholism.

As you can see from this definition, we are provided with the specific indicators included in the measure, the method(s) for data collection, specific scoring of the responses and the interpretation of scale scores.

Using Scales to Measure Variables

When several questions are used to measure one concept, the responses may be combined by taking the sum or average of responses. A composite measure based on this type of sum or average is termed a **scale** (or index). The idea is that idiosyncratic variation in response to particular questions will average out so that the main influence on the combined measure will be the concept on which all the questions focus. Each item is an indicator of the concept, but the item alone is often not a sufficient measure of the concept. A scale can be considered as a more complete measure of the concept than any single component question.

Creating a scale is not just a matter of writing a few questions that seem to focus on a concept. Questions that seem to you to measure a common concept might seem to respondents to concern several different issues. The only way to know that a given set of questions does form a scale is to administer the questions to people like those you plan to study. If a common concept is being measured, people's responses to the different questions should display some consistency.

Scales have already been developed to measure many concepts, and some of these scales have been demonstrated to be accurate in a range of studies. It usually is much better to use such a scale to measure a

concept than it is to try to devise questions to form a new scale. Use of a preexisting scale both simplifies the work involved in designing a study and facilitates comparison of findings to those obtained in other studies. Scales can be found in research articles; on the Internet, for example the ERIC/AE Test Locator (www.ericae.net/testcol.htm); or in compilations such as *Measures for Clinical Practice* (Fischer & Corcoran, 2007).

The Center for Epidemiologic Studies Depression Scale (CES-D) (see Exhibit 4.2) is used to measure the concept of depression. The aspect of depression measured by the scale is the level (the frequency and number combined) of depressive symptoms. Given that depression consists of negative affect, lack of positive affect, and somatic behaviors, the developers of the scale identified questions to assess these dimensions. Many researchers in different studies have found that these questions form an accurate scale. Note that each question concerns a symptom of depression. People may have idiosyncratic reasons for having a particular symptom without being depressed; for example, persons who have been suffering a physical ailment may say that they have a poor appetite. By combining the answers to questions about several symptoms, the scale score reduces the impact of this idiosyncratic variation.

Exhibit 4.2 **Example of a Scale: The Center for Epidemiologic Studies Depression Scale (CES–D)**

INSTRUCTIONS FOR QUESTIONS. Below is a list of the ways you might have felt or behaved in the past week.

Please tell me how often you have felt this way during the past week:

0 = Rarely or none of the time (less than 1 day)
1 = Some or a little of the time (1 to 2 days)
2 = Fairly often (3 to 4 days)
3 = Most or all of the time (5 to 7 days)

During the past week:

1. I was bothered by things that usually don't bother me.
2. I did not feel like eating; my appetite was poor.
3. I could not shake off the blues even with help from my family or friends.
4. I felt I was just as good as other people.
5. I had trouble keeping my mind on what I was doing.
6. I felt depressed.
7. I felt everything I did was an effort.
8. I felt hopeful about the future.
9. I thought my life had been a failure.
10. I felt fearful.
11. My sleep was restless.
12. I was happy.
13. I talked less than usual.
14. I felt lonely.
15. People were unfriendly.
16. I enjoyed life.
17. I had crying spells.
18. I felt sad.
19. I felt people disliked me.
20. I could not "get going."

Source: Radloff (1977).

Some questions in a scale may cluster together in subscales or subsets. All the questions may measure the intended concept, but we may conclude that the concept has several different aspects; this results in a **multidimensional scale**. For example, the CES-D has some items that measure only negative affect, other questions that measure only lack of positive affect, and other questions measuring somatic symptoms. Each of these concepts is an indicator of depression. Researchers may choose to use a variable that summarizes the total scale score or they may choose to use variables that summarize the subscale scores.

The individual items in the CES-D have equal weight, that is, each item makes the same contribution to the depressive symptom score. Some scales have questions that are more central to the concept being measured than other questions and so may be given greater weight when computing the scale score. For example, the MAST asks questions that are assigned different weights. A positive response to the question, "Have you ever been in a hospital because of your drinking?" is given 5 points (weighted higher) while a positive response to the question, "Do you feel you are a normal drinker?" is assigned 2 points.

Treatment as a Variable

Frequently, social work researchers will examine the effectiveness of an intervention or compare two different intervention approaches. When an intervention is compared to no intervention or when two or more interventions are compared, the intervention becomes the independent variable. It is important that a researcher provide a clear nominal definition of the intervention. It is not enough for the researcher to say that the study is comparing one method to another, such as traditional case management to intensive case management. Although the general meaning of such an approach may be familiar to you, the researcher must define what each approach involves. For example, case management may include full support, so that the social worker provides a variety of services and supports including rehabilitation, social skill building, counseling, linking to resources, identifying work and social opportunities, and money management whereas another social worker providing case management may only assess the client, link the client to other services, and periodically reassess the client.

Nominal definitions of an intervention only provide the characteristics or components of the intervention, but fail to fully describe how the intervention was implemented. Researchers provide varying amounts of specificity regarding the actual operationalization of the intervention as is illustrated in the following example. Robert Newcomer, Taewoon Kang, and Carrie Graham (2006) evaluated a specialized case management (Providing Assistance to Caregivers in Transition; PACT) for nursing home individuals returning to the community. They specified the five components of the program and provided details about what each component included. In describing caregiver assessment and care management, they identified who carried out the task, where the assessment was completed, the topics covered in the assessment, the process for care planning, and the activities covered by case management. Yet, some important information is not included such as the case manager's frequency of contact or frequency of periodic reassessment. Therefore, even with a great deal of information, it would still be hard for you to replicate the intervention.

From Observations to Concepts

Qualitative research projects usually take an inductive approach to the process of conceptualization. In an inductive approach, concepts emerge from the process of thinking about what has been observed, as compared to the deductive approach that we just described, in which we develop concepts on the basis of theory and then decide what should be observed to indicate the concept. Instead of deciding in advance which concepts are important for a study, what these concepts mean, and how they should be measured, if

you take an inductive approach, you will begin by recording verbatim what you hear in intensive interviews or see during observational sessions. You will then review this material to identify important concepts and their meaning for participants. At this point, you may identify relevant variables and develop procedures for indicating variation between participants and settings or variation over time.

Qualitative researchers often develop key concepts inductively, in the course of the research, and continue to refine and evaluate the concepts throughout the research. Conceptualization, operationalization, and validation are ongoing and interrelated processes. You will learn more about qualitative research in Chapter 9.

Gathering Data

Social work researchers and practitioners have many options for operationalizing their concepts. We briefly mention these options here but go into much greater depth in subsequent chapters.

Researchers may use a direct measure, such as visual or recorded observation or a physical measure such as a pulse rate. Alternatively, data may be gathered by interviews or self-administered scales and questionnaires. These methods appear to be direct in that we gather the information directly from the respondent or client. Yet what we are trying to do is infer behavior, attitudes, emotions, or feelings because we cannot observe these directly.

There are other sources of information from which measures can be operationalized that are based on information collected by other researchers. Many large data sets have been collected by the federal government, state governments, and nongovernmental sources. Many of these data sets have social indicators that are relevant to social services, such as employment, program participation, income, health, crime, and mental health. Information may also be available from an agency's client records and variables can be operationalized using this information. Regardless of the source, when you rely on data collected by other sources, you are constrained by how variables were operationalized by those who collected the data.

When we have reason to be skeptical of potential respondents' answers to questions, when we cannot observe the phenomena of interest directly, and when there are no sources of available data, we can use indirect or unobtrusive measures, which allow us to collect data about individuals or groups without their direct knowledge or participation (Webb, Campbell, Schwartz, & Sechrest, 2000).

Two types of unobtrusive measures are physical traces and content analysis. The physical traces of past behavior are most useful when the behavior of interest cannot be directly observed. To measure the prevalence of drinking in college dorms or fraternity houses, we might count the number of empty bottles of alcoholic beverages in the surrounding Dumpsters. Content analysis studies are representations of the research topic in such media forms as news articles, chat rooms, and Twitter messages. An investigation of what motivates child abuse reporting might include a count of the amount of space devoted to newspaper articles in a sample of issues of the local newspaper or the number of television newscasters reporting on the maltreatment of children.

Combining Measurement Operations

The choice of a particular measurement method is often determined by available resources and opportunities, but measurement is improved if this choice also takes into account the particular concept or concepts to be measured. Responses to such questions as "How socially engaged were you at the party?" or "How many days did you use sick leave last year?" are unlikely to provide information as valid, respectively, as direct observation or agency records. However, observations at social gatherings may not answer questions about why some people do not participate; we may just have to ask people.

Triangulation—the use of two or more different measures of the same variable—can make for even more accurate measurement (Brewer & Hunter, 2005). When we achieve similar results with different measures of the same variable, particularly when the measures are based on such different methods as survey questions and field-based observations, we can be more confident in the validity of each measure. If results diverge with different measures, it may indicate that one or more of these measures is influenced by more measurement error than we can tolerate. Divergence between measures could also indicate that they actually operationalize different concepts.

2 Levels of Measurement

The final part of operationalization is to assign a value or symbol to represent the observation. Each variable has categories of some sort, and we need to know how to assign a value—typically a number—to represent what has been observed or learned. We may have a discrete variable, whereby the symbol represents a separate category or a different status. The variable may be a **continuous variable**, for which the symbol represents a quantity that can be described in terms of order, spread between the numbers, or relative amounts.

Level of measurement The mathematical precision with which the values of a variable can be expressed. The nominal level of measurement, which is qualitative, has no mathematical interpretation; the quantitative levels of measurement—ordinal, interval, and ratio—are progressively more precise mathematically.

When we know a variable's **level of measurement**, we can better understand how cases vary on that variable and so understand more fully what we have measured. Level of measurement also has important implications for the type of mathematical operations and statistics that can be used with the variable. There are four levels of measurement: nominal, ordinal, interval, and ratio. Exhibit 4.3 depicts the differences among these four levels.

Nominal Level of Measurement

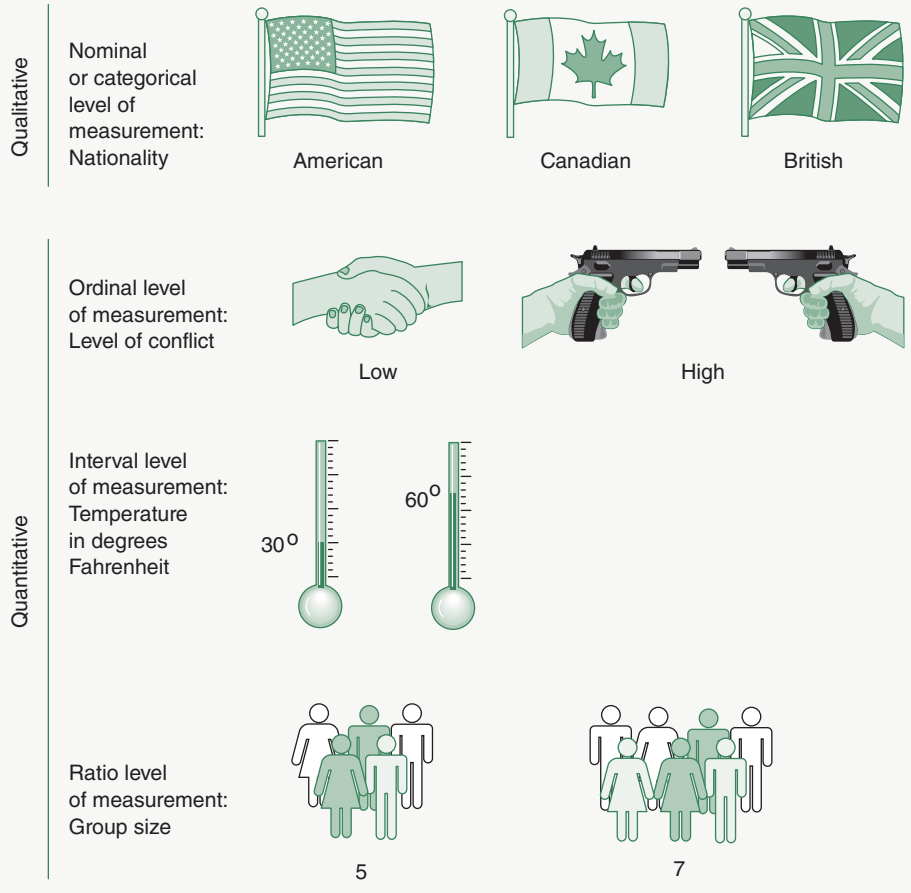
The **nominal level of measurement** identifies variables whose values have no mathematical interpretation; they vary in kind or quality but not in amount. The variable *gender* has two categories or attributes: male and female. We might represent male with the value 1 and female with the value 2, but these numbers do not tell us anything about the difference between male and female except that they are different. Female is not one unit more of gender than male, nor is it twice as much gender.

Nominal level variables are commonplace in social work research. Client characteristics such as ethnicity (e.g., African American, Hispanic, Asian American, White, Native American), marital status (e.g., Married Spouse Present, Married Spouse Absent, Widowed, Divorced, Separated, Never Married), or mental health diagnosis (e.g., Mood Disorder, Personality Disorder) are nominal level variables. Program-related variables such as referral source or type of service used are nominal variables. In each case, the variable has a set of categories whose order has no meaning.

Although the attributes of nominal variables do not have a mathematical meaning, they must be assigned to cases with great care. The attributes we use to categorize cases must be mutually exclusive and exhaustive:

- A variable's attributes or values are **mutually exclusive** if every case can have only one attribute.
- A variable's attributes or values are **exhaustive** when every case can be classified into one of the categories.

Exhibit 4.3 Levels of Measurement



When a variable's attributes are mutually exclusive and exhaustive, every case corresponds to one and only one attribute.

The only mathematical operation we can perform with nominal level variables is a count. We can count how many current clients are females and how many are males. From this count, we can calculate the percentage or proportion of females to males among our clients. If the agency serves 150 women and 100 men, then we can say that 60% of the clients are female. But we cannot identify an average gender nor can we add or subtract or compute any other kind of number.

Ordinal Level of Measurement

The first of the three quantitative levels is the **ordinal level of measurement**. At this level, the numbers assigned to cases specify only the order of the cases, permitting *greater than* and *less than* distinctions. For example, at the coffee shop you might choose between a small, medium, or large cup of coffee—that is ordinal measurement. The categories represent relative cup sizes but the gaps between the various responses do not have any particular meaning. As with nominal variables, the different values of a variable measured at the ordinal level must be mutually exclusive and exhaustive.

The properties of variables measured at the ordinal level are illustrated in Exhibit 4.3 by the contrast between the levels of agreement in two groups. The first group, symbolized by two people shaking hands, has a high level of agreement. The second group, symbolized by two persons pointing guns at each other, has a low level of agreement. To measure agreement, we would put the groups in order by assigning the number 1 to the high agreement group and the number 2 to the low agreement group. The numbers indicate only the relative position or order of the cases. Although high level of agreement is represented by the number 1, it is not one less unit of agreement than low level of agreement, which is represented by the number 2.

A common ordinal measure used in social service agencies is client satisfaction. Often, agencies will ask a client a global question about satisfaction with the services provided by the agency using a rating system such as *4=very satisfied, 3=satisfied*, *2=dissatisfied*, and *1=very dissatisfied*. Someone who responds very satisfied, coded as 4, is clearly more satisfied than someone who responds dissatisfied, coded as 2, but the respondent with a 4 is not twice as satisfied as the respondent with a 2. Nor is the respondent with a 4 two units more satisfied than the respondent with a 2. We only know that the first person is more satisfied than the second person, and therefore, the order has meaning. We can count the number of clients who fall into each category. We can also compute an average satisfaction, but the average is not a quantity of satisfaction; rather, the number summarizes the relative position of the group on the scale.

Agencies sometimes use goal attainment scales to evaluate client outcomes. These scales are usually developed by describing the worst indicators, the best indicators, and several steps in between. The gap between the steps has no meaning, but the scoring represents the progress of the client. Exhibit 4.4 provides an example of a goal attainment scale to measure self-esteem and mother's attitude toward children. The social worker evaluates the extent to which there is improvement in self-esteem based on the nature of the

Exhibit 4.4 Example of a Goal Attainment Scale

Problem Area	Client Outcome Goal	No Achievement	Some Achievement	Major Achievement
Self-esteem	To develop increased feeling of self-esteem	Makes only negative statements Does not identify strengths No verbal expression of confidence No sense of self-worth	Some positive statements Some negative statements Can identify some strengths but overly critical about self Emerging confidence Emerging self-worth	Makes many positive statements Few to no negative statements Can identify strengths without qualifying statements Is confident Has self-worth
Mother's attitude toward child	Less of a negative attitude toward child	Resists child's affection Constantly shows anger verbally and nonverbally Constantly shows frustration Constantly shows hostility Constantly impatient	Occasional affection Occasional anger Occasional frustration Occasional hostility Occasional impatience	Accepts child's affection No verbal or nonverbal signs of anger, hostility, or frustration Patient

verbal and nonverbal responses of the client. There is an order to the levels of achievement, and we can describe how many clients fall into each category.

Interval Level of Measurement

At the **interval level of measurement** numbers represent fixed measurement units but have no absolute or fixed zero point. An interval level of measurement also has mutually exclusive categories, the categories are exhaustive, and there is an order to the responses. This level of measurement is represented in Exhibit 4.3 by the difference between two Fahrenheit temperatures. Although 60 degrees is 30 degrees hotter than 30 degrees, 60 in this case is not twice as hot as 30. Why not? Because heat does not begin at 0 degrees on the Fahrenheit scale. Therefore, the numbers can be added and subtracted but ratios between them (2 to 1 or twice as much) are not meaningful.

There are few true interval level measures in social work, but many social work researchers treat scales created by combining responses to a series of ordinal level variables as interval level measures. This is frequently done because there are more mathematical operations associated with interval level variables. A scale of this sort could be created with responses to the Core Institute's (1994) questions about friends' disapproval of substance use (Exhibit 4.5). The survey has 13 questions on the topic, each of which has the same three response choices: "Don't disapprove" is valued at 1, "Disapprove" is valued at 2, and "Strongly disapprove" is valued at 3. Each question can be used independently of the other questions to provide useful information: an ordinal level of measurement. Alternatively, the responses to the 13 questions can be summed to reflect overall disapproval. The scale would then range from 13 to 39, with higher scores representing greater disapproval. A score of 24 could be treated as if it were 12 more units than a score of 12, but that does not mean that there is twice as much disapproval. Or the responses could be averaged to retain the original 1 to 3 range.

Exhibit 4.5 Example of Interval-Level Measures: Core Alcohol and Drug Survey

26. How do you think your close friends feel (or would feel) about you...
(mark one for each line)

		Don't disapprove	Disapprove	Strongly disapprove
a.	Trying marijuana once or twice	○	○	○
b.	Smoking marijuana occasionally	○	○	○
c.	Smoking marijuana regularly	○	○	○
d.	Trying cocaine once or twice	○	○	○
e.	Taking cocaine regularly	○	○	○
f.	Trying LSD once or twice	○	○	○
g.	Taking LSD regularly	○	○	○
h.	Trying amphetamines once or twice	○	○	○
i.	Taking amphetamines regularly	○	○	○
j.	Taking one or two drinks of an alcoholic beverage (beer, wine, liquor) nearly every day	○	○	○
k.	Taking four or five drinks nearly every day	○	○	○
l.	Having five or more drinks in one sitting	○	○	○
m.	Taking steroids for body building or improved athletic performance	○	○	○

Ratio Level of Measurement

The **ratio level of measurement** represents fixed measuring units with an absolute zero point; in this situation, zero means absolutely no amount of whatever the variable indicates. On a ratio scale, 10 is two points higher than 8 and is also two times greater than 5. Ratio numbers can be added and subtracted, and because the numbers begin at an absolute zero point, they can be multiplied and divided (so ratios can be formed between the numbers). For example, people's ages can be represented by values ranging from zero years (or some fraction of a year) to 120 or more. A person who is 30 years old is 15 years older than someone who is 15 years old (30 – 15 = 15) and is twice as old as that person (30/15 = 2). Of course, the numbers also are mutually exclusive, are exhaustive, have an order, and there are equal gaps.

Exhibit 4.3 displays an example of a variable measured at the ratio level. The number of people in the first group is 5, and the number in the second group is 7. The ratio of the two groups' sizes is then 1.4, a number that mirrors the relationship between the sizes of the groups. Note that there does not actually have to be any group with a size of 0; what is important is that the numbering scheme begins at an absolute zero—in this case, the absence of any people.

Ratio level variables are common in social work research. We can count the number of clients in a program, the time spent in a particular activity, or the number of hot meals delivered to homebound elderly. We can describe a community by the number of community development organizations, number of after school programs, or the number of low-income households. In each case, the answer *zero* is meaningful, representing the complete absence of the variable.

The Case of Dichotomies

Dichotomies, variables having only two values, are a special case from the standpoint of levels of measurement. The values or attributes of a variable such as depression clearly vary in kind or quality, not in amount. Thus, the variable, depression, is categorical—measured at the nominal level. Yet in practical terms, we can think of the variable in a slightly different way, as indicating the presence of the attribute *depressed* or *not depressed*. Viewed in this way, there is an inherent order; a depressed person has more of the attribute (it is present) than a person who is not depressed (the attribute is not present). Nonetheless, although in practical terms there is an order, we treat dichotomous variables as a nominal variable.

Mathematical Comparisons

Exhibit 4.6 summarizes the types of comparisons that can be made with different levels of measurement, as well as the mathematical operations that are legitimate with each. All four levels of measurement allow researchers to assign different values to different cases. All three quantitative measures allow researchers to rank cases in order.

Researchers choose levels of measurement in the process of operationalizing the variables; the level of measurement is not inherent in the variable. Many variables can be measured at different levels with different procedures. A variable to describe alcoholic drinking can be measured by asking respondents to identify how many alcoholic drinks they had in the last week, a ratio variable, or answer the same question by checking *None*, *1 to 4*, *5 to 9*, or *10 or more*, an ordinal variable. A nominal variable about drinking

Exhibit 4.6 Properties of Measurement Levels

Examples of Comparison Statements	Appropriate Math Operations	Relevant Level of Measurement			
		Nominal	Ordinal	Interval	Ratio
A is equal to (not equal to) B	= (≠)	✓	✓	✓	✓
A is greater than (less than) B	> (<)		✓	✓	✓
A is three more than (less than) B	+ (−)			✓	✓
A is twice (half) as large as B	× (÷)				✓

could be created by asking, "Did you consume any alcoholic drink in the last week" with response categories *yes* or *no*.

It is a good idea to try to measure variables at the highest level of measurement possible. The more information available, the more ways we have to compare cases. There are more possibilities for statistical analysis with quantitative than with qualitative variables. You can create ordinal or nominal variables from ratio level variables, but you cannot go in the reverse direction. If you know the actual number of alcoholic drinks, you can combine the reports into categories at a later time, but if you ask respondents to check the category, you cannot later modify that variable to reflect the actual number of drinks consumed.

Be aware that other considerations may preclude measurement at a high level. For example, many people are reluctant to report their exact incomes even in anonymous questionnaires. So asking respondents to report their income in categories (such as less than $10,000, $10,000–19,999, $20,000–29,999, or $30,000 and higher) will elicit more responses, and, thus, more valid data, than asking respondents for their income in dollars.

Oftentimes, researchers treat variables measured at the interval and ratio levels as comparable. They then refer to this as the interval-ratio level of measurement. You will learn in Chapter 12 that different statistical procedures are used for variables with fixed measurement units, but it usually does not matter whether there is an absolute zero point.

2 Measurement Error

No matter how carefully we operationalize and design our measures, no measure is perfect, and there will be some error. What respondents report (the reported score) is not necessarily the true response (the true score) because of the imperfections of measurement. The true response differs from the reported response because of measurement error, of which there are two types: systematic error and random error.

Systematic error is generally considered to be a predictable error, in that we can predict the direction of the error. Think about weighing yourself on a scale each day. If you put a scale on a particular part of the floor in your house, you will always weigh less (reported score) than you actually do (true score). The direction of the error is predictable: In this case, your scale will always underreport your true weight.

There are different forms of systematic error and each of these forms of systematic error reflects some bias. The various forms include:

- *Social desirability.* Social desirability bias occurs when respondents wish to appear most favorable in the eyes of the interviewer or researcher.
- *Acquiescence bias.* There is a tendency for some respondents to agree or disagree with every statement, regardless of whether they actually agree.
- *Leading questions.* Leading questions have language that is designed to influence the direction of a respondent's answer. There are many different ways in which this might be done, such as using words that have a negative connotation in society (e.g. government regulation or liberal), using the names of controversial people, or including some but not all responses to a question in the actual question.
- *Differences in subgroup responses according to gender, ethnicity, or age.* Differences in cultural beliefs or patterns, socialization processes, or cohort effects may bias findings from what otherwise might be a set of neutral questions.

To avoid systematic error requires careful construction of scales and questions and the testing of these questions with different population groups. We explore these methods in depth in Chapter 8.

Unlike systematic error, **random error** is unpredictable in terms of its effects. Random error may be due to the way respondents are feeling that particular day. Respondents may be having a great day or, in contrast, they may be fatigued, bored, or not in a cooperative mood. Perhaps the weather is making them less willing to cooperate. Respondents may also be affected by the conditions of the testing. The lighting may be bad, the room may be noisy, the seating may be cramped, the lack of walls in the cubicle may mean other people can hear, there may be other people in the room, or they may not like the looks of the person gathering the information.

Another form of random error is *regression to the mean*. This is the tendency of persons who score very high on some measure to score lower the next time, or the reverse, for persons who score very low to score higher. What might have influenced the high or low score on the first test may not operate in the second test.

Random error might occur when researchers rating behavior are not adequately trained to do the rating. For example, two people grading an essay test might come up with different grades if they have not discussed the grading criteria beforehand. A field supervisor and a beginning student might assess a client differently given the variation in their years of experience.

As we have already said, the effects of random error cannot be predicted: Some responses overestimate the true score, whereas other responses underestimate the true score. Many researchers believe that if the sample size is sufficiently large, the effects of random error cancel each other out. Nonetheless, we want to use measurement scales and questions that are stable to minimize the effects of random error as much as possible.

2 How to Assess Measurement Accuracy

Do the operations to measure our concepts provide stable or consistent responses—are they reliable? Do the operations developed to measure our concepts actually do so—are they valid? Why are these questions important? When we test the effectiveness of two different interventions or when we monitor a client's progress, we want the changes we observe to be due to the intervention and not due to the instability or inaccuracy of the measurement instrument. We also want to know that the measure we use is really a measure of the outcome and not a measure of some other outcome. If we have weighed our measurement options, carefully constructed our questions and observational procedures, and selected from the available data indicators, we should be on the right track. But we cannot have much confidence in a measure until we have evaluated with data its reliability and validity.

Measurement Reliability

Reliability means that a measurement procedure yields consistent or equivalent scores when the phenomenon being measured is not changing. If a measure is reliable, it is affected less by random error or chance variation than if it is unreliable. Reliability is a prerequisite for measurement validity: We cannot really measure a phenomenon if the measure we are using gives inconsistent results. The methods to evaluate measurement reliability include test-retest reliability, internal consistency, alternate forms, and interrater and intrarater reliability.

Test-Retest Reliability

When researchers measure a phenomenon that does not change at two different time points, the degree to which the two measurements are related is the **test-retest reliability** of the measure. If you take a test of your research methodology knowledge and retake the test 2 months later, the test is reliable if you receive a similar score both times, presuming that nothing happened during the 2 months to change your research methodology knowledge. We hope to find a correlation between the two tests of about .7 and prefer even a higher correlation, such as .8.

Of course, if events between the test and the retest have changed the variable being measured, then the difference between the test and retest scores should reflect that change. As the gap in time between the two tests increases, there is a greater likelihood that real change did occur. This also presumes that you were not affected by the conditions of the testing: a **testing effect**. The circumstances of the testing, such as how you were given the test, or environmental conditions, such as lighting or room temperature, may impact on test scores. A testing effect may extend to how you felt the first time you took the test; because you did not know what to expect the first time, you may have been very nervous, as opposed to the second time, when you knew what to expect.

Internal Consistency

When researchers use multiple items to measure a single concept, they are concerned with **internal consistency**. For example, if the items comprising the CES-D (like those in Exhibit 4.2) reliably measure depressive symptoms, the answers to the different questions should be highly associated with one another. The stronger the association among the individual items and the more items that are included the higher the reliability of the scale.

One method to assess internal consistency is to divide the scale into two parts, or **split-half reliability**. We might take a 20-item scale, such as the CES-D, sum the scores of the first 10 items, sum the scores of the second 10 items (items 11 through 20), and then correlate the scores for each of the participants. If there is internal consistency, the correlation should be fairly high, such as .8 or .9. The correlation typically gets higher the more items there are in the scale.

There are countless ways in which you might split the scale, and in practical terms, it is nearly impossible to split the scale by hand into every possible combination. The speed of computers enables us to calculate a score that splits the scale in every combination. A summary score, such as **Cronbach's alpha coefficient**, is the average score of all the possible split-half combinations.

Alternate-Forms Reliability

Researchers are testing **alternate forms reliability** (or parallel forms reliability) when they compare subjects' answers to slightly different versions of survey questions (Litwin, 1995). A researcher may reverse the order of the response choices in a scale, modify the question wording in minor ways, or create a set of different questions. The two forms are then administered to the subjects. If the two set of responses are not too different, alternate forms reliability is established. For example, you might remember taking the SATs or ACTs when you were in high school. When you compared notes with your friends, you found that each of you had taken different tests. The developers had evaluated the tests using alternate forms reliability to ensure that the different forms were equivalent and comparable.

Interrater Reliability

When researchers use more than one observer to rate the same people, events, or places, **interrater reliability** is their goal. If observers are using the same instrument to rate the same phenomenon, their

ratings should be similar. If they are similar, we can have much more confidence that the ratings reflect the phenomenon being assessed rather than the orientations of the raters.

Assessments of interrater reliability may be based on the correlation of the rating between two raters. Two raters could evaluate the quality of play between five teenage mothers and their children on a 10-point scale. The correlation would show whether the direction of the raters' scores was similar as well as how close the agreement was for the relative position for each of the five scores. One rater may judge the five interactions as 1, 2, 3, 4, and 5, whereas the second rater scores the interactions as 6, 7, 8, 9, and 10. The correlation would be quite high—in fact, the correlation would be perfect. But as demonstrated by this example, the agreement about the quality of the interactions was quite different. So an alternative method is to estimate the percentage of exact agreement between the two raters. In this example, the rater agreement is zero.

Assessing interrater reliability is most important when the rating task is complex. Consider a commonly used measure of mental health, the Global Assessment of Functioning Scale (GAF). The rating task seems straightforward, with clear descriptions of the characteristics that are supposed to determine GAF scores. But in fact, the judgments that the rater must make while using this scale are very complex. They are affected by a wide range of respondent characteristics, attitudes, and behaviors as well as by the rater's reactions. As a result, interrater agreement is often low on the GAF, unless the raters are trained carefully.

Intrarater Reliability

Intrarater reliability occurs when a single observer is assessing an individual at two or more points in time. It differs from test-retest reliability in that the ratings are done by the observer as opposed to the subjects. Intrarater reliability is particularly important when you are evaluating a client's behavior or making judgments about the client's progress.

Measurement Validity

Validity refers to the extent to which your indicators measure what they are intended to measure. Technically, a valid measure of a concept is one that is (a) closely related to other apparently valid measures, (b) closely related to the known or supposed correlates of that concept, and (c) not related to measures of unrelated concepts (Brewer & Hunter, 2005). A good measure of your current age should correspond to your age calculated from your birth certificate. Measurement validity is assessed with four different approaches: face validation, content validation, criterion validation, and construct validation.

Face Validity

Researchers apply the term **face validity** to the confidence gained from careful inspection of a concept to see whether it is appropriate "on its face." A measure has face validity if it obviously pertains to the meaning of the concept being measured more than to other concepts (Brewer & Hunter, 2005). For example, a count of how many drinks people consumed in the past week would be a face valid measure of their alcohol consumption.

Although every measure should be inspected in this way, face validation does not provide any evidence of measurement validity. The question "How much beer or wine did you have to drink last week?" looks valid on its face as a measure of frequency of drinking, but people who drink heavily tend to underreport the amount they drink. So the question would be an invalid measure in a study that includes heavy drinkers.

Content Validity

Content validity establishes that the measure covers the full range of the concept's meaning. To determine that range of meaning, the researcher may solicit the opinions of experts and review literature that

identifies the different aspects or dimensions of the concept. Like face validity, content validity lacks empirical support, and experts may disagree with the range of content provided in a scale.

Criterion Validity

Criterion validity is established when the results obtained from one measure are similar to the results obtained with a more direct or already validated measure of the same phenomenon (the criterion). The criterion that researchers select can itself be measured either at the same time as the variable to be validated or after that time. **Concurrent validity** exists when a measure yields scores that are closely related to scores on a criterion measured at the same time. A measure of blood-alcohol concentration or a urine test could serve as the criterion for validating a self-report measure of drinking as long as the questions we ask about drinking refer to the same period. **Predictive validity** is the ability of a measure to predict scores on a criterion measured in the future. SAT or ACT scores as a measure of academic ability can be validated when compared with college grades.

Criterion validation greatly increases confidence that the measure is measuring what was intended. It is a stronger form of validity than face or content validity as it is based on empirical evidence rather than subjective assessment.

Construct Validity

Construct validity is demonstrated by showing that a measure is related to a variety of other measures of other concepts as specified in a theory. This validation approach is commonly used in social work research when no clear criterion exists for validation purposes. The construct validation process relies on using a deductive theory with hypothesized relationships among the concepts (Koeske, 1994). The measure has construct validity (or theoretical construct validity) if it behaves as it should relative to the other concepts in the theory. For example, Hann, Winter, and Jacobsen (1999) compared subject scores on the CES-D to a number of indicators that they felt from previous research and theory should be related to depression: fatigue, anxiety, and global mental health. They found that individuals with higher CES-D scores tended to have more problems in each of these areas, giving us more confidence in the CES-D's validity as a measure.

There are other approaches to establish construct validity that you are likely to encounter when reading research literature. **Convergent validity** is when you can show a relationship between two measures of the same construct that are assessed using different methods (Koeske, 1994). **Discriminant validity** is achieved if the measure to be validated has a weak or no relationship to other unrelated concepts. The CED-D would demonstrate convergent validity if the scale scores correlated strongest with the Beck Depression Inventory (a validated scale to measure depression) and discriminant validity if the scores correlated lower with the Beck Anxiety Inventory (a validated scale to measure anxiety).

Another form of construct validity is **known groups validity**, which is demonstrated by comparing the scale scores to groups with and without the characteristic measured by the scale. We would expect the CES-D scores to be higher among people who have a clinical diagnosis of major depression than people who have a clinical diagnosis of anxiety.

The distinction between criterion and construct validation is not always clear. Opinions can differ about whether a particular indicator is indeed a criterion for the concept that is to be measured. A key difference is simply that with criterion validity, "the researcher's primary concern is with the criterion in a practical context, rather than with the theoretical properties of the construct measure" (Koeske, 1994, p. 50). What if you want to validate a question-based measure of the amount of social support that people receive from their friends? Should you just ask people about the social support they have received? Could friends' reports of the amount of support they provided serve as a criterion? Are verbal accounts of the amount of support provided adequate? What about observations of social support that people receive?

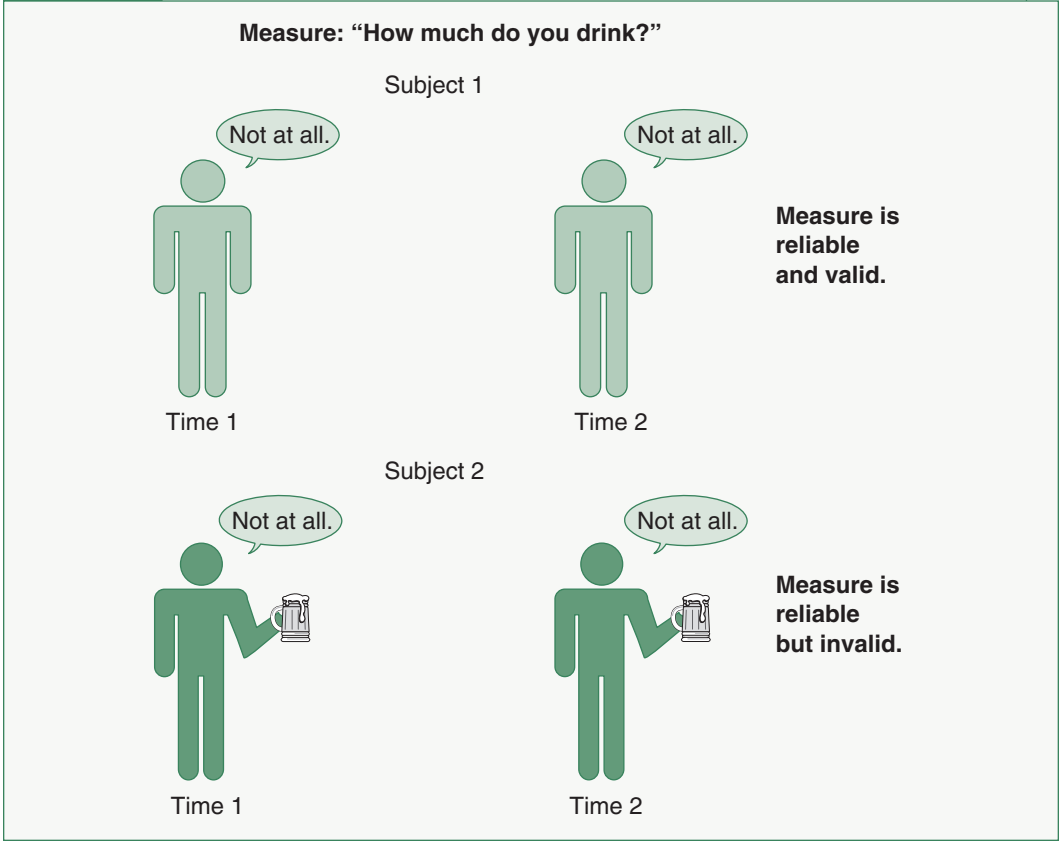
Even if you could observe people in the act of counseling or otherwise supporting their friends, can an observer be sure that the interaction is indeed supportive? There is not really a criterion here, just related concepts that could be used in a construct validation strategy.

What construct and criterion validation have in common is the comparison of scores on one measure to scores on other measures that are predicted to be related. It is not so important that researchers agree that a particular comparison measure is a criterion rather than a related construct. But it is very important to think critically about the quality of the comparison measure and whether it actually represents a different view of the same phenomenon.

Reliability and Validity of Existing Measures

A reliable measure is not necessarily a valid measure, as Exhibit 4.7 illustrates. This discrepancy is a common flaw of self-report measures of substance abuse. Most respondents answer questions in a consistent manner, so the scales are reliable. However, a number of respondents will not admit that they drink even though they drink a lot. Their answers to the questions are consistent, but they are consistently misleading. So the scales based on self-report are reliable but invalid. Unfortunately, many measures are judged to be worthwhile on the basis only of a reliability test.

Exhibit 4.7 The Difference Between Reliability and Validity: Drinking Behavior



The reliability and validity of measures in any study must be tested after the fact to assess the quality of the information obtained. If it turns out that a measure cannot be considered reliable and valid, little can be done to save the study. Hence, it is important to select in the first place measures that are likely to be reliable and valid. Consider the different strengths of different measures and their appropriateness to your study. Conduct a pretest in which you use the measure with a small sample and check its reliability. Provide careful training to ensure a consistent approach if interviewers or observers will administer the measure. In most cases, however, the best strategy is to use measures that have been used before and whose reliability and validity have been established in other contexts. But the selection of tried and true measures still does not absolve researchers from the responsibility of testing the reliability and validity of the measure in their own studies.

2 Using Scales to Identify a Clinical Status

Many scales do not just measure the range or intensity of some phenomenon but are also used by researchers and practitioners as screening tools to make educated guesses about the presence or absence of clinical conditions. For example, the CES-D has been used to determine the extent of depression in the community. CES-D scale scores may range from 0 to 60; people with scores 16 or higher may be classified as depressed whereas people scoring below 16 may be classified as not depressed. This score is called a **cut-off score**.

Cut-off score A scale score used to define the presence or absence of a particular condition.

Cut-off scores should be as accurate as possible. If not, we risk expending limited resources on what may turn out to be an inaccurate assessment, we risk missing individuals with the condition, and we risk labeling clients with a condition they might not actually have. Typically, the validity of a cut-off score is assessed by comparing the scale's classifications to an established clinical evaluation method. The CES-D cut-off score might be compared with a clinical diagnosis using the *DSM-IV-TR*.

A summary of the analysis of the validity of a cut-off is presented in Exhibit 4.8. If the cut-off scale provides an accurate assessment, there should be a high proportion of cases classified as either a **true negative** (cell a) or a **true positive** (cell d). A true negative occurs when based on the scale the client is assessed as not having a problem and really does not have the problem. A true positive occurs when it is determined from the obtained scale score that the client has a problem and the client really does have the problem based on the clinical evaluation. There should be few **false negative** (cell b) cases when based on the scale score you conclude that the client does not have the problem, but the client really does have the problem and few **false positive** (cell c) cases when you conclude from the scale score that the client does have a significant problem, but in reality does not have the problem.

Researchers use different measures to establish the validity of the cut-off scores. **Sensitivity** describes the true positive cell; it reflects a proportion based on the number of people who are assessed as having the condition (d) relative to the number of people who actually have the condition (b+d), or d/b+d. **Specificity** describes the true negative cell. It is a proportion based on the number of people assessed as not having a condition (a) relative to the number who really do not have the condition (a + c); its mathematical formula is a/(a+c). False negative rates and false positive rates are similarly calculated.

Exhibit 4.8 Outcomes of Screening Scale Versus Clinical Assessment

	Actual Diagnosis for the Clinical Condition		
Screening Scale Result	Client does not have clinical condition	Client has clinical condition	Total
Assessed as not having condition	True negative (a)	False negative (b)	a + b
Assessed as having the condition	False positive (c)	True positive (d)	c + d
Total	a + c	b + d	

Ideally, we would like the sensitivity and specificity of the scale's cut-off scores to be very high so that we make few mistakes. Yet there are tradeoffs. To identify all the true positives, the cut-off score would need to be eased; in the case of the CES-D, it would need to be lowered. This will increase sensitivity but will also likely result in more false positives, which means a lower specificity. Making it more difficult to test positive requires setting a higher score; this will increase the specificity but will also produce more false negatives and the sensitivity score will decline.

Two other types of estimates you will see are the positive predictive value and the negative predictive value. The positive predictive value is the proportion of people who actually have the condition (d) to the number who were assessed by the screening tool as having the condition (c + d), that is, d/(c + d). The negative predictive value is the proportion of all those who actually do not have the condition (a) compared to all those who were assessed as having the condition (a + b); this is calculated by a/(a + b). The ability to predict accurately is useful when we decide to use a screening scale to get some sense of how prevalent a particular condition is in the community. So if we wanted to assess how common depression is in the community, we would want high predictive values.

2 Measurement in a Diverse Society

Although it is crucial to have evidence of reliability and validity, it is also important that such evidence generalize to the different populations social workers serve. Often people of color, women, the poor, and other groups have not been adequately represented in the development or testing of various measurement instruments (Witkin, 2001). Just because a measure appears valid does not mean that you can assume cross-population generalizability.

It is reasonable to consider whether the concepts we use have universal meaning or differ across cultures or other groups. C. Harry Hui and Harry C. Triandis (1985) suggest that there are four components that must be evaluated to determine whether a concept differs across cultures including:

1. *Conceptual equivalence*. The concept must have the same meaning, have similar precursors and consequences, and relate to other concepts in the same way.
2. *Operational equivalence*. The concept must be evident in the same way so that the operationalization is equivalent.
3. *Item equivalence*. Items used must have the same meaning to each culture.
4. *Scaler equivalence*. The values used on a scale mean the same in intensity or magnitude.

Take the concept, *self-esteem*. Bae and Brekke (2003) note that cross-cultural research has found that Asian Americans typically have lower self-esteem scores than other ethnic groups. They hypothesized that Korean Americans would have lower scores on positively worded items than other ethnic groups but would have similar scores on negatively worded items. They suggested that this response pattern was due to culture: "Giving high scores on the positive items is intrinsically against their collective culture in which presenting the self in a self-effacing and modest manner is regarded as socially desirable behavior to maintain social harmony" (Bae & Brekke, 2003, p. 28). Bae and Brekke did find that overall self-esteem scores were lower among Korean Americans and that it was due to Korean Americans scoring lower on the positively worded items while scoring the same or higher than other ethnic groups on the negatively worded items.

Similar concerns have been noted for scales measuring depression. For example, Joy Newmann (1987) has argued that gender differences in levels of depressive symptoms may reflect differences in the socialization process of males and females. She suggests that some scales ask questions such as crying, being lonely, and feeling sad, which are more likely to be responded to in the affirmative by women and not by men because men are socialized to not express such feelings. Stephen Cole, Ichiro Kawachi, Susan Maller, and Lisa Berkman (2000) did find that women were much more likely than men to endorse the item "feeling like crying" and suggested that the item be dropped from the scale. Debra Ortega and Cheryl Richey (1998) note that people of color may respond differently to questions used in depression scales. Some ethnic groups report feelings of sadness or hopelessness as physical complaints and therefore have high scores on these questions but low scores on emotion-related items. Ortega and Richey also note that some items in depression scales, such as suicidal ideation, are not meaningful to some ethnic groups. The elderly are more likely to endorse some items that also measure physical changes as opposed to changes brought about by depression (Sharp & Lipsky, 2002).

Biased scores can result in practical problems. For example, many scales include cut-off scores to demonstrate the presence or absence of a condition. If there is a response bias, the result could be the treatment of a condition that does not exist or not identifying a condition that does exist (Bae & Brekke, 2003; Ortega & Richey, 1998). The failure to measure correctly may affect the ability to identify effective interventions. The relationship of different phenomena may be distorted because of measurement bias. Therefore, it is important to assess the samples used for validation and to use measures that have been validated with the population group to whom it will be administered.

2 Measurement Implications for Evidence-Based Practice

Measurement is an essential ingredient in social work practice whether it is your assessment of a client or your monitoring and evaluation of your practice. Further, the studies you review depend, in part, on

the quality of the measurement; systematic errors can negate the validity of a particular study (Johnston, Sherer, & Whyte, 2006). You need to be confident that the evidence presented is due to the intervention and not the instability of the measurement instrument.

What should you consider when you examine the efficacy of a measure for your agency? In the previous sections, we stressed the importance of measurement reliability and validity. That alone is insufficient because there should be evidence of the appropriateness of the measure for the population with whom it will be used. Therefore, when you review research about the reliability and validity of a measure, you need to look at the samples that were used in the studies. Too often these studies are done without consideration of gender, race, ethnicity, or age. It may be that the samples used in the studies look nothing like the population you are serving. If that is the case, the instrument may not be appropriate for your agency or setting.

The same holds true for scales that can be used for diagnostic purposes; there should be statistical evidence that the scale is accurate in its determination of correct (true positives and true negatives) diagnoses with few wrong (false positives and false negatives) diagnoses (Warnick, Weersing, Scahill, & Woolston, 2009). Earlier, we described the CES–D as a commonly used scale with a more or less acceptable cut-off score of 16. On further inspection, researchers found that this score was too low to be useful with the elderly. Some item reports in the CES–D can be due to physical conditions that are common among the elderly. As a result, an appropriate cut-off score for elderly people with physical ailments has been determined to be 20 (Schein & Koenig, 1997). The bottom line is to take nothing for granted about cut-off scores described in the literature.

Of course, you should also keep in mind practical considerations in selecting a measurement scale. These considerations include:

- *Administration of the scale*. Different methods of administration require different amounts of time to complete, as well as skill to gather the data. For example, self-report takes less time than interviewing the client.
- *Cost*. The instrument should be affordable. Many useful measures and scales can be found in the public domain, but many other scales have to be purchased, and sometimes you must also pay for their scoring.
- *Sensitivity to change*. The measure you use should be sufficiently sensitive to pick up changes in the desired outcome, and there should be a sufficient number of items that you are able to identify changes.
- *Reactivity*. To the extent possible, you want nonreactive measures, that is, measures that do not influence the responses that people provide.
- *Acceptability*. The measures have to be accepted by staff as measures that will provide valid information

All of these were considerations we had to take into account when we were asked by a family service agency's senior adult unit to recommend a short and simple screen for pathological gambling. The agency uses a 25- to 30-minute psychosocial assessment at intake, screening for a variety of social, economic, health, and mental health concerns, so they did not want something that would add terribly to the length of the assessment. At the same time, they wanted something that would be accurate, easy to use, and not offend their older clients. Ultimately, we found a reliable and valid two-item screen that could be added to the intake assessment.

Just as there are systematic reviews of intervention research, you may find systematic reviews of different measurement and screening instruments. For example, Henry O'Connell and his colleagues

(O'Connell et al., 2004) recently reviewed self-report alcohol screening instruments for older adults and Warnick, Weersing, Scahill, and Woolston (2009) reviewed measures to predict youth mental health.

As you read intervention research or other types of research studies or you develop a research proposal, there are important questions for you to consider. You should identify the major concepts in the study and assess whether the measure is clearly defined. Next, you should examine how the concepts are operationalized. Is the operational definition sufficient to capture the various dimensions of the concept? When scales are used, is there evidence of reliability and validity as well as the scale's appropriateness for the specific study population? Our confidence in the measure is enhanced when the author reports methods used to enhance the reliability of the measure, such as the specific training in collecting the information or using multiple measures.

2 Conclusion

Remember always that measurement validity is a necessary foundation for social work research and professional practice. Gathering data without careful conceptualization or conscientious efforts to operationalize key concepts often is a wasted effort. The difficulties of achieving valid measurement vary with the concept being operationalized and the circumstances of the particular study.

Planning ahead is the key to achieving valid measurement in your own research; careful evaluation is the key to sound decisions about the validity of measures in others' research. Statistical tests can help to determine whether a given measure is valid after data have been collected, but if it appears after the fact that a measure is invalid, little can be done to correct the situation. If you cannot tell how key concepts were operationalized when you read a research report, do not trust the findings. If a researcher does not indicate the results of tests used to establish the reliability and validity of key measures, remain skeptical.

Key Terms

Alternate forms reliability
Concept
Conceptualization
Concurrent validity
Constant
Construct validity
Content validity
Continuous variable
Convergent validity
Criterion validity
Cronbach's alpha coefficient
Cut-off score
Discriminant validity
Exhaustive
Face validity
False negative
False positive
Internal consistency
Interrater reliability
Interval level of measurement
Intrarater reliability
Known groups validity
Level of measurement
Multidimensional scale
Mutually exclusive
Nominal definition
Nominal level of measurement
Operational definition
Operationalization
Ordinal level of measurement
Predictive validity
Random error
Ratio level of measurement
Reliability
Scale
Sensitivity
Specificity
Split-half reliability
Systematic error
Testing effect
Test-retest reliability
Triangulation
True negative
True positive

Highlights

- Conceptualization plays a critical role in research. In deductive research, conceptualization guides the operationalization of specific variables.
- Concepts may refer to either constant or variable phenomena. Concepts that refer to variable phenomena may be very similar to the actual variables used in a study, or they may be much more abstract.
- Concepts should have a nominal definition and an operational definition. A nominal definition defines the concept in terms of other concepts while the operational definition provides the specific rules by which you measure the concept.
- In social work research, a treatment or intervention is often a variable. The intervention should have an operational definition, that is, a description of the intervention process.
- Scales measure a concept by combining answers to several questions and so reducing idiosyncratic variation. Several issues should be explored with every scale: Does each question actually measure the same concept? Does combining items in a scale obscure important relationships between individual questions and other variables? Is the scale multidimensional?
- Measures are not perfect, and there may be two types of measurement error. Systematic error refers to predictable error and should be minimized. Random error is unpredictable in terms of effect on measurement.
- Level of measurement indicates the type of information obtained about a variable and the type of statistics that can be used to describe its variation. The four levels of measurement can be ordered by complexity of the mathematical operations they permit: nominal (least complex), ordinal, interval, ratio (most complex). The measurement level of a variable is determined by how the variable is operationalized.
- The validity of measures should always be tested. There are four basic approaches: face validation, content validation, criterion validation, and construct validation.
- Measurement reliability is a prerequisite for measurement validity, although reliable measures are not necessarily valid. The forms of reliability include: test-retest, internal consistency, parallel forms, interrater, and intrarater.
- Some scales are used to screen for the presence or absence of a clinical condition and, therefore, use cut-off scores. The accuracy of cut-off scores is assessed using measures of sensitivity and specificity.
- In examining studies of measurement reliability and validity, it is important to look at the samples to ensure that there is evidence of reliability and validity for different population subgroups.

Discussion Questions

1. Describe the relationship between a nominal definition and an operational definition of a concept. How are these two types of definitions related?
2. What does "global assessment of functioning" mean to you? What behaviors would you look for to assess global assessment of functioning? Identify two such behaviors. What questions would you ask to measure global assessment of functioning? Create a scale by writing five questions with response choices. How would you assess the reliability and validity of your scale?
3. If you were given a questionnaire right now that asked you about your use of alcohol and illicit drugs in the past year, would you disclose the details fully? How do you think others would respond? What if the questionnaire was anonymous? What if there was a confidential ID number on the questionnaire so that only the researcher could keep track of who responded? What criterion validation procedure would you suggest for assessing measurement validity?

Critiquing Research

Using one of the articles provided on the website, *Learning From Journal Articles*, www.sagepub.com/fswr2e, answer the following questions:

1. What are the major concepts used in the study? What are the nominal definitions? Does the author provide clear and complete nominal definitions for each concept? Are some concepts treated as unidimensional that you think might best be thought of as multidimensional?
2. What are the variable operational definitions? Are the operational definitions adequate? Do the measures of the variables seem valid and reliable? How does the author establish measurement reliability and measurement validity? Is there evidence that the reliability and validity of the measurements have been assessed with populations or samples similar to the study sample?

Making Research Ethical

1. Why is it important that the reliability and validity of any scale be evaluated with different populations?

Developing a Research Proposal

At this point you can begin the process of conceptualization and operationalization.

1. Identify the concepts you will use in the study. Provide a nominal definition for each concept. When possible, this definition should come from the existing literature—either a book you have read for a course or a research article.
2. How will the concepts be operationalized? Identify the variables you will use to study the research question. Which of these variables are independent or dependent variables? What is the level of measurement for each variable? How will these variables be coded?
3. Develop measurement procedures or identify existing instruments that might be used. If you are using a new measure, what procedures will you use to determine the reliability and validity of the measure? If you are using an existing instrument, report the evidence for the instrument's reliability and validity.

Web Exercises

1. How would you define alcoholism? Write a brief definition. Based on this conceptualization, describe a method of measurement that would be valid for a study of alcoholism.
2. Now go to the National Institute on Alcohol Abuse and Alcoholism and read some facts about alcohol (http://www.niaaa.nih.gov/alcohol-health/overview-alcohol-consumption/alcohol-use-disorders). Is this information consistent with the definition you developed for Question 1?

STUDENT STUDY SITE

Visit **www.sagepub.com/fswr2e** to access additional study tools including eFlashcards, web quizzes, web resources, interactive exercises, data sets, links to SAGE journal articles and more.