Chapter 3

Conceptualization and Measurement

Every time you begin to review or design a research study, you will have to answer two questions: (1) What do the main concepts mean in this research? (2) How are the main concepts measured? Both questions must be answered to evaluate the validity of any research. For instance, to study a hypothesized link between religious fundamentalism and terrorism, you may conceptualize terrorism as "nongovernmental political violence," and measure incidents of terrorism by counting for a 5-year period the number of violent attacks that have explicit political aims. You will also need to define and measure "religious fundamentalism," no easy task. What counts? And how should you decide what counts? We cannot make sense of a researcher's study until we know how the concepts were *defined* and *measured.* Nor can we begin our own research until we have defined our concepts clearly and constructed valid measures of them.

In this chapter, we briefly address the issue of conceptualization, or defining your main terms. We then describe measurement sources such as available archive data, questions, observations, and less direct, or unobtrusive, measures. We then discuss the level of measurement reflected in different measures. The final topic is to assess the validity and reliability of these measures. By the chapter's end, you should have a good understanding of measurement, the first of the three legs (measurement, generalizability, and causality) on which a research project's validity rests.

WHAT DO WE HAVE IN MIND?

A May 2000 *New York Times* article (Stille, 2000) announced that the "social health" of the United States had risen a bit, after a precipitous decline in the 1970s and 1980s. Should we be relieved? Concerned? What, after all, does "social health" mean? To social scientist Marc Miringoff, it has to do with social and economic inequalities. To political pundit William J. Bennett, it is more a matter of moral values. In fact, the **concept** of social health means different things to different people. Most agree that it has to do with "things that are not measured in the gross national product" and is supposed to be "a more subtle and more meaningful way of measuring what's important to [people]" (Stille, 2000:A19). But until we agree on a definition of social health, we can't decide whether it has to do with child poverty, trust in government, out-of-wedlock births, alcohol-related traffic deaths, or some combination of these or other phenomena.

Conceptualization

A continuing challenge for social scientists, then, rests on the fact that many of our important topics of study (social health, for instance) are not clearly defined things or objects (like trees or rocks), but are abstract concepts or ideas. A concept is an image or idea, not a simple object. Some concepts are relatively simple, such as a person's age or sex: almost everyone would agree what it means to be 14 years old or female. But other concepts are more ambiguous. For instance, if you want to count the number of families in Chicago, what counts as a family? A husband and wife with two biological children living in one house—yes, that's a family.

Do cousins living next door count? Cousins living in California? Or maybe the parents are divorced; or the children adopted; or maybe the children are grown. Maybe two women live together with one adopted child and one biological child fathered by a now-absent man. So perhaps "living together" is what defines a family—or is it biology? Or is it a crossing of generations—that is, the presence of adults and children? The particular definition you develop will affect your research findings, and some people probably won't like it whatever you do, but how you define "family" obviously affects your results.

Often social concepts can be used sloppily or even misleadingly. Nowadays you may hear that "the economy" is doing well, but it may seem to you that many people are worse off. Typically, in news reports "the economy" refers to the Gross Domestic Product—the total amount of economic activity (value of goods and services, precisely) in the country in a given year. When the GDP goes up, reporters say "the economy is improving." But that's very different from saying that the average working person makes more money than he or she would have 20 years ago—in fact, he or she makes less. We could use the concept of "the economy" to refer to the economic well-being of actual people, but that's not typically in fact how it's used.

Defining concepts clearly can be quite difficult because many concepts have several meanings and can be measured in many ways. What is meant, for instance, by the idea of "power?" The classic definition, provided by German sociologist Max Weber, is that power is the ability to meet your goals over the objections of other people. That definition implies that unknown people can be quite powerful, whereas certain presidents of the United States have been relatively powerless. A different definition might equate power to one's official position; in that case, the president of the United States would be powerful. Or perhaps power is equated with prestige, so famous intellectuals like Albert Einstein would be considered powerful. Or maybe power is defined as having wealth, so that rich people are seen as powerful.

And even if we can settle on a definition, how then do we actually measure power? Should we ask a variety of people if a certain person is powerful? Should we review their acts over the last 10 years and see when they exerted their will over others? Should we try to uncover the true extent of their wealth and use that? How about power at a lower level, say, as a member of student government? The most visible and vocal people in your student assembly may be, in fact, quite unpopular and perhaps not very powerful at all—just loud. At the same time, there may be students who are members of no official body whatsoever, but somehow they always get what they want. Isn't that power? From these varied cases, you can see that power can be quite difficult to conceptualize.

Or describing what causes "crime," or even what causes "theft," is inherently problematic, since the very definition of these terms is spectacularly flexible and indeed forms part of their interest for us. What counts as theft varies dramatically, depending on the thief—a next-door neighbor, a sister, or a total stranger wandering through town—and what item is taken: a bottle of water, your watch, a lawn mower, a skirt, your reputation, or $5. Indeed, part of what makes social science interesting are the debates over, for instance, what is a theft, or what is crime.

So conceptualization—working out what your key terms will mean, in your research—is a crucial part of the research process. Definitions need to be explicit. Sometimes conceptualization is easy: "Older men are more likely to suffer myocardial infarction than younger men," or "Career military officers mostly vote for Republican candidates in national elections." Most of the concepts used in those statements are easily understood and easy to measure (sex, age, military status, voting). In other cases, conceptualization is quite difficult: "As people's moral standards deteriorate, the family unit starts to die," or "intelligence makes you more likely to succeed."

Conceptualization, then, is the process of matching up terms (family, sex, happiness, power) to clarified definitions for them—really, figuring out what are the social "things" you'll be talking about.

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

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.

It is especially important to define clearly concepts that are abstract or unfamiliar. When we refer to concepts like "social control," "anomie," or "social status," we cannot count on others knowing exactly what we mean. Even experts may disagree about the meaning of frequently used concepts if they base their conceptualizations on different theories. That's OK. 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, and we should expect others to do the same.

Conceptualization also involves creating concepts, or thinking about how to conceive of the world: What things go together? How do we slice up reality? Cell phones, for instance, may be seen as communication devices, like telephones, radios, telegraphs, or two tin cans connected by a string. But they can also be conceived in another way: a college administrator we know, seeing students leaving class outside her building, said, "Cell phones have replaced cigarettes." She reconceptualized cell phones, seeing them not as communication tools but as something to fiddle with, like cigarettes, chewing gum wrappers, keys on a lanyard, or the split ends of long hair. In conceptualizing the world, we create the lenses through which we see it.

Our point is not that conceptualization problems are insurmountable, but that (1) you need to develop and clearly state what you *mean* by your key concepts, and (2) your measurements will need to be clear and consistent with the definitions you've settled on (more on that topic shortly).

Variables and Constants

After we define the concepts for a study, we must identify variables that correspond to those concepts. For example, we might be interested in what affects students' engagement in their academic work—when they are excited about their studies, when they become eager to learn more, and so on. Our main concept, then, would be "engagement." We could use any number of different variables to measure engagement: the student's reported interest in classes, a teacher's evaluation of student engagement, the number of hours spent on homework, or an index including a number of different questions. Any of these variables could show a high or low level of student engagement. If we are to study variation in engagement, we must identify variables to measure that are most pertinent to our theoretical concerns.

You should be aware that not every concept in a particular study is represented by a variable. In our student engagement study, all of the students *are* students—there is no variation in that. So "student," in this study, is called a **constant** (it's always the same), not a variable.

There are many variables that could measure student engagement. Which variables should we select? It's very tempting, and all too common, to simply try to "measure everything," by including in a study every variable we can think of. We could collect self-reports of engagement, teacher ratings, hours studied per week, pages of essays written for class, number of visits to the library per week, frequency of participation in discussion, times met with professors, and on and on. This haphazard approach will inevitably result in the collection of some useless data and the failure to collect some important data. Instead, we should take four steps:

1. Examine the theories that are relevant to our research question to identify those concepts that would be expected to have some bearing on the phenomena we are investigating.

2. Review the relevant research literature and assess the utility of variables used in prior research.

3. Consider the constraints and opportunities for measurement that are associated with the specific setting(s) we will study. Distinguish constants from variables in this setting.

4. Look ahead to our analysis of the data. What role will each variable play in our analyses?

Remember: A few well-chosen variables are better than a barrel full of useless ones.

HOW WILL WE KNOW WHEN WE'VE FOUND IT?

Once we have defined our concepts in the abstract—that is, after conceptualizing—and we have identified the variables that we want to measure, we must develop our measurement procedures. The goal is to devise **operations** that actually measure the concepts we intend to measure—in other words, to achieve measurement validity.

Exhibit 3.1 represents the **operationalization** process in three studies. The first researcher defines his or her concept, binge drinking, and chooses one variable—frequency of heavy episodic drinking—to represent it. This variable is then measured with responses to a single question, or *indicator:* "How often within the last 2 weeks did you drink five or more drinks containing alcohol in a row?" Because "heavy" drinking is defined differently for men and women (relative to their different metabolisms), the question is phrased in terms of "four or more drinks" for women. The second researcher defines her concept—poverty—as having two aspects or dimensions, subjective poverty and absolute poverty. Subjective poverty is measured with responses to a survey question: "Would you say that you are poor?" Absolute poverty is measured by comparing family income to the poverty threshold. The third researcher decides that her concept—social class—is defined by a position on three measured variables: income, education, and occupational prestige.

Operationalization: The process of specifying the operations that will indicate the value of cases on a variable.

Measures can be based on activities as diverse as asking people questions, reading judicial opinions, observing social interactions, coding words in books, checking census data tapes, enumerating the contents of trash receptacles, or drawing urine and blood samples. Experimental researchers may operationalize a concept by manipulating its value; for example, to operationalize the concept of exposure to anti-drinking messages, some subjects may listen to a talk about binge drinking while others do not. We will focus here on the operations of using

Exhibit 3.1 Concepts, Variables, and Indicators: Operationalizing Concepts

Concept	Variable	Indicator
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
Social class	Income	Annual total self-reported income
	Education	Years of schooling
	Occupational prestige	Rated prestige of current occupation, based on national polls
		Total = Social class

published data, asking questions, observing behavior, and using unobtrusive means of measuring people's behavior and attitudes.

The variables and measurement operations chosen for a study should be consistent with the purpose of the research question. Suppose we hypothesize that college students who go abroad for their junior year have a more valuable experience than those who remain at their college. If our purpose is *evaluation* of different junior-year options, we can operationalize "junior-year programs" by comparing (1) traditional coursework at home, (2) study in a foreign country, and (3) internships at home that are not traditional college courses. A simple question, asking students in each program, for example, "How valuable do you feel your experience was?" would help to provide the basis for determining the relative value of these programs. But if our purpose is *explanation,* we would probably want to interview students to learn what features of the different programs made them valuable, to find out the underlying dynamics of educational growth.

Time and resource limitations also must be taken into account when we select variables and devise measurement operations. For many sociohistorical questions (such as "How has the poverty rate varied since 1950?"), census data or other published counts must be used. On the other hand, a historical question about the types of social bonds among combat troops in 20th-century wars probably requires retrospective interviews with surviving veterans. The validity of the data is lessened by the unavailability of many veterans from World War I and by problems of recall, but direct observation of their behavior during the war is certainly not an option.

Using Available Data

Government reports are rich, accessible sources of social science data. Organizations ranging from nonprofit service groups to private businesses also compile a wealth of figures that may be available to some social scientists for some purposes. In addition, the data collected in many social science surveys are archived and made available for researchers who were not involved in the original survey project.

Before we assume that available data will be useful, we must consider how appropriate they are for our concepts of interest, whether other measures would work better, or whether our concepts can be measured at all with these data. For example, many organizations informally (and sometimes formally) use turnover—that is, how many employees quit each year—as a measure of employee morale (or satisfaction). If turnover is high (or retention rates are low), morale must be bad and needs to be raised. Or so the thinking goes.

But obviously, factors other than morale affect whether people quit their jobs. When a single chicken-processing plant is the only employer in a small town, and other jobs are hard to find, and people live on low wages, turnover may be very low even among miserable workers. In the "dot-com" companies of the late 1990s, turnover was high—despite amazingly good conditions, salary, and morale—because the industry was so hungry for good workers that companies competed ferociously to attract them. Maybe the concepts "morale" and "satisfaction," then, can't be measured adequately by the most easily available data, that is, turnover rates.

We also cannot assume that available data are accurate, even when they appear to measure the concept. "Official" counts of homeless persons have been notoriously unreliable because of the difficulty of locating homeless persons on the streets, and government agencies have at times resorted to "guesstimates" by service providers. Even available data for such seemingly straightforward measures as counts of organizations can contain a surprising amount of error. For example, a 1990 national church directory reported 128 churches in a Midwest county; an intensive search in that county in 1992 located 172 churches (Hadaway, Marler, & Chaves, 1993:744).

When legal standards, enforcement practices, and measurement procedures have been taken into account, comparisons among communities become more credible. However, such adjustments may be less necessary when the operationalization of a concept is relatively unambiguous, as with the homicide rate: Dead is dead. And when a central authority imposes a common data-collection standard, as with the FBI's Uniform Crime Reports, data become more comparable across communities. But careful review of measurement operations is still important, because procedures for classifying a death as a homicide can vary between jurisdictions and over time.

Another rich source of already-collected data is survey datasets archived and made available to university researchers by the Inter-University Consortium for Political and Social Research. One of its most popular survey datasets is the General Social Survey (GSS). The GSS is administered regularly by the National Opinion Research Center (NORC) at the University of Chicago to a sample of more than 1,500 Americans (annually until 1994; biennially since then). GSS questions vary from year to year, but an unchanging core of questions includes measures of political attitudes, occupation and income, social activities, substance abuse, and many other variables of interest to social scientists. This dataset can easily be used by college students to explore a wide range of interesting topics. However, when surveys are used in this way, after the fact, researchers must carefully evaluate the survey questions. Are the available measures sufficiently close to the measures needed that they can be used to answer the new research question?

Constructing Questions

Asking people questions is the most common, and probably most versatile, operation for measuring social variables. Do you play on a varsity team? What is your major? How often, in a week, do you go out with friends? How much time do you spend on schoolwork? Most concepts about individuals are measured with such questions. In this section, we'll introduce some options for writing single questions, explain why single questions can sometimes be inadequate measures, and then examine the use of multiple questions to measure a concept.

In principle, questions, asked perhaps as part of a survey, can be a straightforward and efficient means by which to measure individual characteristics, facts about events, level of knowledge, and opinions of any sort. In practice, though, survey questions can easily result in misleading or inappropriate answers. All questions proposed for a survey must be screened carefully for their adherence to basic guidelines and then tested and revised until the researcher feels some confidence that they will be clear to the intended respondents (Fowler, 1995). Some variables may prove to be inappropriate for measurement with any type of question. We have to recognize that memories and perceptions of the events about which we might like to ask can be limited.

Specific guidelines for reviewing questions are presented in Chapter 6; here, our focus is on the different types of survey questions.

Single Questions

Measuring variables with single questions is very popular. Public opinion polls based on answers to single questions are reported frequently in newspaper articles and TV newscasts: "Do you favor or oppose U.S. policy in Iraq?" "If you had to vote today, for which candidate would you vote?" Social science surveys also rely on single questions to measure many variables: "Overall, how satisfied are you with your job?" "How would you rate your current health?"

Single questions can be designed with or without explicit response choices. The question that follows is a **closed-ended**, or **fixed-choice, question**, because respondents are offered explicit responses from which to choose. It has been selected from the Core Alcohol and Drug Survey distributed by the Core Institute, Southern Illinois University, for the FIPSE Core Analysis Grantee Group (Presley, Meilman, & Lyerla, 1994).

Compared with other campuses with which you are familiar, this campus's use of alcohol is . . . (Mark one)

___ *Greater than other campuses*

___ *Less than other campuses*

___ *About the same as other campuses*

Most surveys of a large number of people contain primarily fixed-choice questions, which are easy to process with computers and analyze with statistics. However, fixed-response choices can obscure what people really think unless the choices are designed carefully to match the range of possible responses to the question.

Most important, response choices should be mutually exclusive and exhaustive, so that every respondent can find *one and only one* choice that applies to him or her (unless the question is of the "Check all that apply" variety). To make response choices exhaustive, researchers may need to offer at least one option with room for ambiguity. For example, a questionnaire asking college students to indicate their school status should not use freshman, sophomore, junior, senior, and graduate student as the only response choices. Most campuses also have students in a "special" category, so you might add "Other (please specify)" to the five fixed responses to this question. If respondents do not find a response option that corresponds to their answer to the question, they may skip the question entirely or choose a response option that does not indicate what they are really thinking.

Researchers who study small numbers of people often use **open-ended questions**, which don't have explicit response choices and allow respondents to write in their answers. The next question is an open-ended version of the earlier fixed-choice question:

How would you say alcohol use on this campus compares to that on other campuses?

An open-ended format is preferable when the full range of responses cannot be anticipated, especially when questions have not been used previously in surveys or when questions are asked of new groups. Open-ended questions also can allow clear answers when questions involve complex concepts. In the previous question, for instance, "alcohol use" may cover how many students drink, how heavily they drink, if the drinking is public or not, if it affects levels of violence on campus, and so on.

Just like fixed-choice questions, open-ended questions should be reviewed carefully for clarity before they are used. For example, if respondents are asked, "When did you move to Boston?" they might respond with a wide range of answers: "In 1944." "After I had my first child." "When I was 10." "20 years ago." Such answers would be very hard to compile. To avoid such ambiguity, rephrase the question to clarify the form of the answer, for instance, "In what year did you move to Boston?" Or provide explicit response choices (Center for Survey Research, 1987).

Indexes and Scales

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 an **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 that all the questions focus on. In addition, the index can be considered a more complete measure of the concept than can any one of the component questions.

Creating an index 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, in fact, form an index 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.

Because of the popularity of survey research, indexes already have been developed to measure many concepts, and some of these indexes have proved to be reliable in a range of studies. Usually it is much better to use such an index than it is to try to form a new one. Use of a preexisting index both simplifies the work of designing a study and facilitates comparison of findings from other studies.

The questions in Exhibit 3.2 represent a short form of an index used to measure depression; it is called the Center for Epidemiologic Studies Depression Index

Exhibit 3.2 Example of an Index: Excerpt From the Center for Epidemiologic Studies Depression Index (CES-D)

At any time during the past week . . . (Circle one response on each line)	*Never*	*Some of the time*	*Most of the time*
a. Was your appetite so poor that you did not feel like eating?	1	2	3
b. Did you feel so tired and worn out that you could not enjoy anything?	1	2	3
c. Did you feel depressed?	1	2	3
d. Did you feel unhappy about the way your life is going?	1	2	3
e. Did you feel discouraged and worried about your future?	1	2	3
f. Did you feel lonely?	1	2	3

Source: Lenore Radloff, 1977. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population." *Applied Psychological Measurement, 1:* 385–401. Reprinted by permission of Sage Publications, Inc. Copyright 1977 West Publishing Company/Applied Psychological Measurement Inc. Reproduced by permission.

(CES-D). Many researchers in different studies have found that these questions form a reliable index. Note that each question concerns a symptom of depression. People may well have one particular symptom without being depressed; for example, persons who have been suffering from a physical ailment may say that they have a poor appetite. By combining the answers to questions about several symptoms, the index reduces the impact of this idiosyncratic variation. (This set of questions uses what is termed a "matrix" format, in which a series of questions that concern a common theme are presented together, with the same response choices.)

Usually an index is calculated by simply averaging responses to the questions, so that every question counts equally. But sometimes, either intentionally by the researcher or by happenstance, questions on an index arrange themselves in a kind of hierarchy in which an answer to one question effectively provides answers to others. For instance, a person who supports abortion on demand almost certainly supports it in cases of rape and incest as well. Such questions form a **scale**. In a scale, we give different weights to the responses to different questions before summing or averaging the responses. Responses to one question might be counted two or three times as much as responses to another. For example, based on Christopher Mooney and Mei Hsien Lee's (1995) research on abortion law reform, a scale to indicate support for abortion might give a "1" to agreement that abortion should be allowed "when the pregnancy resulted from rape or incest" and a

"4" to agreement with the statement that abortion should be allowed "whenever a woman decided she wanted one." A "4" rating is much stronger, in that anyone who gets a "4" would also probably agree to all "lower-number" questions as well.

Making Observations

Asking questions, then, is one way to operationalize, or measure, a variable. *Observations* can also be used to measure characteristics of individuals, events, and places. The observations may be the primary form of measurement in a study, or they may supplement measures obtained through questioning.

Direct observations can be used as indicators of some concepts. For example, Albert J. Reiss Jr. (1971) studied police interaction with the public by riding in police squad cars, observing police-citizen interactions, and recording the characteristics of the interactions on a form. Notations on the form indicated such variables as how many police-citizen contacts occurred, who initiated the contacts, how compliant citizens were with police directives, and whether or not police expressed hostility toward the citizens.

Often, observations can supplement what is initially learned from interviews or survey questions, putting flesh on the bones of what is otherwise just a verbal self-report. In Chambliss's (1996) book, *Beyond Caring,* a theory of the nature of moral problems in hospital nursing that was originally developed through interviews was expanded with lessons learned from observations. Chambliss found, for instance, that in interviews nurses described their daily work as exciting, challenging, dramatic, and often even heroic. But when Chambliss himself sat for many hours and watched nurses work, he found that their daily lives were rather humdrum and ordinary, even to them. Occasionally there were bursts of energetic activity and even heroism—but the reality of day-to-day nursing was far less exciting than interviews would lead one to believe. Indeed, Chambliss's original theory was modified to include a much broader role for routine in hospital life.

Direct observation is often the method of choice for measuring behavior in natural settings, as long as it is possible to make the requisite observations. Direct observation avoids the problems of poor recall and self-serving distortions that can occur with answers to survey questions. It also allows measurement in a context that is more natural than an interview. But observations can be distorted, too. Observers do not see or hear everything, and what they do see is filtered by their own senses and perspectives. Moreover, in some situations the presence of an observer may cause people to act differently from the way they would otherwise (Emerson, 1983). If you set up a video camera in an obvious spot on campus, in order to monitor traffic flows, you may well change the flow—just because people will see the camera and avoid it (or come over to make faces). We will discuss these issues in more depth in Chapter 8, but it is important to begin to consider them whenever you read about observational measures.

Content Analysis

Content analysis is a research method for systematically analyzing and making inferences from text (Weber, 1985:9). You can think of a content analysis as a "survey" of documents ranging from newspapers, books, or TV shows to persons referred to in other communications, themes expressed in government documents, or propositions made in tape-recorded debates. Words or other features of these units are then "coded" to measure the variables involved in the research question (Weber, 1985). As a simple example of content analysis, you might look at a variety of women's magazines over the past 25 years, and count the number of articles in each year devoted to various topics, such as makeup, weight loss, relationships, sex, and so on. You might count the number of articles on different subjects as a measure of the media's emphasis on women's anxiety about these issues and see how that emphasis (i.e., the number of articles) has increased or decreased over the past quarter-century. At the simplest level, you could code articles by whether key words ("fat," "weight," "pounds," etc.) appeared in the titles.

After coding procedures are developed, their reliability should be assessed by comparing different coders' results for the same variables. Computer programs for content analysis can be used to enhance reliability (Weitzman & Miles, 1994). The computer is programmed with certain rules for coding text so that these rules will be applied consistently.

Collecting Unobtrusive Measures

Unobtrusive measures allow us to collect data about individuals or groups without their direct knowledge or participation. In their recently revised classic book, Eugene Webb and his colleagues (2000) identified four types of unobtrusive measures: physical trace evidence, archives (available data), simple observation, and contrived observation (using hidden recording hardware or manipulation to elicit a response). These measures provide valuable supplements or alternatives to more standard survey-based measures, because they are not affected by an interviewer's appearance or how he or she asks questions. We have already considered some types of archival and observational data (Webb, Campbell, Schwartz, & Sechrest, 2000).

The physical traces of past behavior are one type of unobtrusive measure that provides creative opportunities. Patterns of grass wear on a campus quad show where most people walk. 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. Student interest in their college courses might be measured by counting the number of times that books on reserve as optional reading are checked out, or the number of class handouts left in recycling bins outside a lecture hall. The most popular stalls in restrooms can be determined by their high rate of toilet paper replacement.

Potential unobtrusive measures are everywhere. Webb and his colleagues (2000:37) suggested measuring the interest in museum exhibits by the frequency with which tiles in front of the exhibits needed to be replaced. If auto mechanics note the radio dial settings in cars brought in for repairs, they can target their advertising to those stations to which their customers listen most. A quick glance at the hands of men in a bar could help you see if the patrons do heavy manual work (calluses).

Unobtrusive measures can also be created from such diverse forms of media as newspaper archives, magazine articles, TV or radio talk shows, legal opinions, historical documents, personal letters, or e-mail messages. Researchers may read and evaluate the text of Internet listservs, as Fox and Roberts (1999) did in a study of British physicians. We could even learn about cities by comparing their telephone directory yellow pages! For example, we find that Sarasota, Florida, has many pages devoted to nursing homes and hospital appliances; Chattanooga, Tennessee, which has approximately the same number of people, rather than having pages devoted to medical care, has many pages listing churches. Admittedly, that's a rough way to compare cities, but it may alert us to key differences between them.

Combining Measurement Operations

The choice of a particular measurement method—questions, observations, archives, and the like—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 questions such as "How socially adept were you at the party?" or "How many days did you use sick leave last year?" are unlikely to provide valid information on shyness or illness. Direct observation or company records may work better. On the other hand, observations at cocktail parties may not fully answer our questions about why some people are shy; we may just have to ask people. Or if a company keeps no record of sick leave, we may have to ask direct questions and hope for accurate memories. Every choice of a measurement method entails some compromise between the perfect and the possible.

Triangulation—the use of two or more different measures of the same variable—can strengthen measurement considerably (Brewer & Hunter, 1989:17). When we achieve similar results with different measures of the same variable, particularly when they are based on such different methods as survey questions and field-based observations, we can be more confident of the validity of each measure. In surveys, for instance, people may say that they would return a lost wallet they found on the street. But field observation may prove that in practice many succumb to the temptation to keep the wallet. The two methods produced different results. In a contrasting example, post-combat interviews of American soldiers in World War II found that most GIs never fired their weapons in battle; and the written, archival records of ammunition resupply patterns confirmed this interview finding (Marshall, 1978). If results diverge when using different measures, it may indicate that we are sustaining more measurement error than we can tolerate.

Divergence between measures could also indicate that each measure actually operationalizes a different concept. An interesting example of this interpretation of divergent results comes from research on crime. Crime statistics are often inaccurate measures of actual crime; what gets reported to the police and shows up in official statistics is not at all the same thing as what happens according to victimization surveys (in which random people are asked if they have been a crime victim). Social scientists generally regard victim surveys as a more valid measure of crime than police-reported crime. We know, for instance, that rape is a dramatically underreported crime, with something like four to ten times the number of rapes occurring as are reported to police. But auto theft is an *overreported* crime: More auto thefts are reported to police than actually occur. This may strike you as odd, but remember that almost everyone who owns a car also owns car insurance; if their car is stolen, they will definitely report it to the police in order to claim the insurance. Some other people might report a car stolen when it hasn't been, because of the financial incentive. (By the way, insurance companies are quite good at discovering this scam, so it's a bad way to make money.)

Murder, however, is generally reported to police at roughly the same rate at which it actually occurs (i.e., official police reports generally match victim surveys). When someone is killed, it's very difficult to hide the fact. A body is missing, a human being doesn't show up for work, people find out. At the same time, it's very hard to pretend that someone was murdered when they weren't. There they are, still alive, in the flesh. Unlike rape or auto theft, there are no obvious incentives for either underreporting or overreporting murders. The official rate is generally valid.

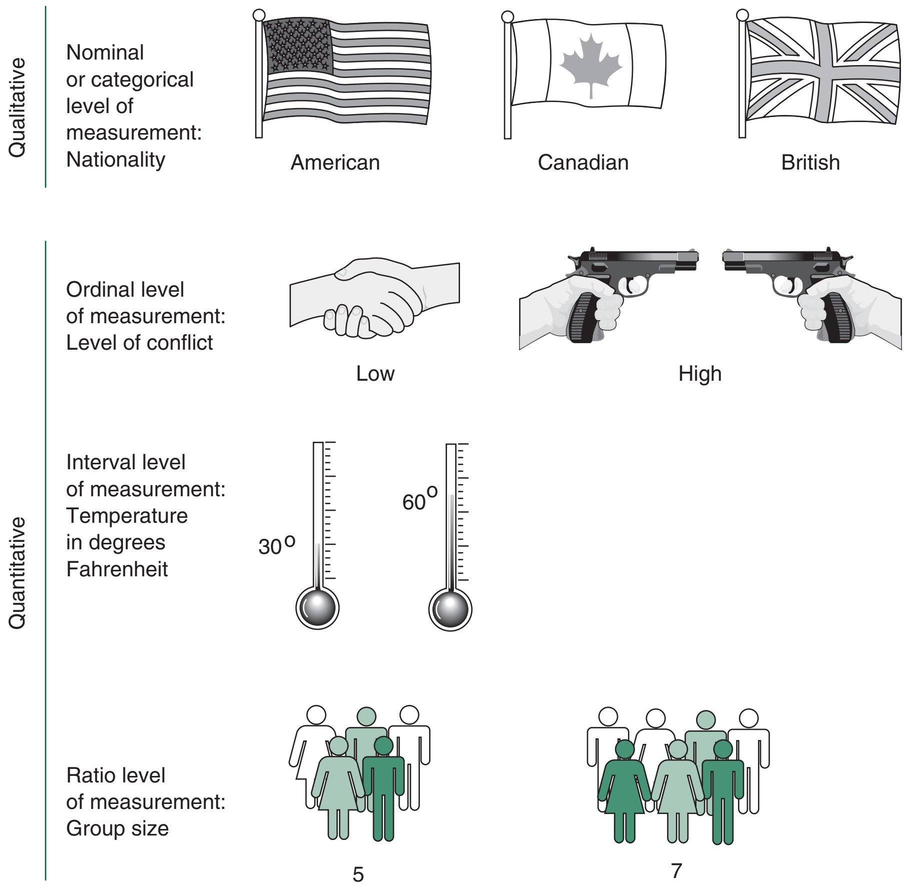
So if you can, it's best to use multiple measures of the same variable; that way each measure helps to check the validity of the others.

HOW MUCH INFORMATION DO WE REALLY HAVE?

There are many ways of collecting information, or different *operations* for gathering data: asking questions, using previously gathered data, analyzing texts, and so on. Some of this data contains mathematically detailed information; it represents a higher level of measurement. There are four **levels of measurement:** nominal, ordinal, interval, and ratio. Exhibit 3.3 depicts the differences among these four levels.

> ***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.*

Exhibit 3.3 Levels of Measurement


Qualitative
Nominal or categorical level of measurement: Nationality
American
Canadian
British
Ordinal level of measurement: Level of conflict
Low
High
Quantitative
Interval level of measurement: Temperature in degrees Fahrenheit
30°
60°
Ratio level of measurement: Group size
5
7

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. “State” (referring to the United States) is one example. The variable has 50 attributes (or categories or qualities), but none of them is more “state” than another. They’re just different. Religious affiliation is another nominal variable, measured in categories: Christian, Moslem, Hindu, Jewish, and so on. Nationality, occupation, and region of the country are also measured at the nominal level. A person may be Spanish or Portuguese, but one nationality does not represent more nationality than another—just a different nationality (see Exhibit 3.3). A person may be a doctor or a truck driver, but one does not represent three units more occupation than the other. Of course, more people may identify themselves as being of one nationality than of another, or one occupation may have a higher average income than another occupation, but these are comparisons involving variables other than “nationality” or “occupation” themselves.

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 measure, or 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.

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

Ordinal Level of Measurement

The first of the three quantitative levels is the **ordinal level of measurement**. At this level, you specify only the order of the cases, in “greater than” and “less than” distinctions. At the coffee shop, for example, you might choose between a small, medium, or large cup of decaf—that’s ordinal measurement.

The properties of variables measured at the ordinal level are illustrated in Exhibit 3.3 by the contrast between the level of conflict in two groups. The first group, symbolized by two people shaking hands, has a low level of conflict. The second group, symbolized by two people pointing guns at each other, has a high level of conflict. To measure conflict, we could put the groups “in order” by assigning the number 1 to the low-conflict group and the number 2 to the high-conflict group, but the numbers would indicate only the relative position or order of the cases.

As with nominal variables, the different values of a variable measured at the ordinal level must be mutually exclusive and exhaustive. They must cover the range of observed values and allow each case to be assigned no more than one value.

Interval Level of Measurement

At the **interval level of measurement**, numbers represent fixed measurement units but have no absolute zero point. This level of measurement is represented in Exhibit 3.3 by the difference between two Fahrenheit temperatures. Note, for example, that 60 degrees is 30 degrees higher than 30 degrees; but 60 is not "twice as hot" as 30. Why not? Because heat does not "begin" at 0 degrees on the Fahrenheit scale. The numbers can therefore be added and subtracted, but ratios of them (2 to 1 or "twice as much") are not meaningful. There are thus few true interval-level measures in the social sciences; most are ratio-level, because they have zero points.

Sometimes, though, social scientists will create indexes by combining responses to a series of variables measured at the ordinal level and then treat these indexes as interval-level measures. An index of this sort could be created with responses to the Core Institute's questions about friends' disapproval of substance use (see Exhibit 3.4). The survey has 13 questions on the topic, each of

Exhibit 3.4 Ordinal Measures: Core Alcohol and Drug Survey. Responses could be combined to create an interval scale (see text).

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	○	○	○

Source: Core Institute, Core Alcohol and Drug Survey, 1994. Carbondale, IL: Core Institute.

which has the same three response choices. If "Don't disapprove" is valued at 1, "Disapprove" is valued at 2, and "Strongly disapprove" is valued at 3, the summed index of disapproval would range from 13 to 39. A score of 20 could be treated as if it were four more units than a score of 16. Or the responses could be averaged to retain the original 1–3 range.

Ratio Level of Measurement

A **ratio level of measurement** represents fixed measuring units with an absolute zero point. Zero, in this situation, 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 as great as 5. Ratio numbers can be added and subtracted, and because the numbers begin at an absolute zero point, they can also be multiplied and divided (so ratios can be formed between the numbers).

For example, people's ages can be represented by values ranging from 0 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 also twice as old as that person (30/15 = 2). Of course, the numbers also are mutually exclusive and exhaustive, so that every case can be assigned one and only one value. Age (in years) is clearly a ratio-level measure.

Exhibit 3.3 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.

Comparison of Levels of Measurement

Exhibit 3.5 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 variables; the level of measurement is not inherent in the variable itself. Many variables can be measured at different levels, with different procedures. Age can be measured as young or old; 0–10, 11–20, 21–30, and so on; or as 1, 2, or 3 years old. We could gather the data by asking people their age, by having an observer guess ("Now *there's* an old guy!"), or by searching through hospital records for exact dates and times of birth. Any of these approaches could work, depending on our research goals.

Exhibit 3.5 Properties of Measurement Levels

		Relevant level of measurement			
Examples of comparison statements	*Appropriate math operations*	*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	× (/)				✓

Usually, though, it is a good idea to measure variables at the highest level of measurement possible. The more information available, the more ways we have to compare cases. We also have more possibilities for statistical analysis with quantitative than with qualitative variables. Even if your primary concern is only to compare teenagers to young adults, you should measure age in years rather than in categories; you can always combine the ages later into categories corresponding to "teenager" and "young adult."

Be aware, however, that other considerations may preclude measurement at a high level. For example, many people are very 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, and so on) will elicit more responses, and thus more valid data, than asking respondents for their income in dollars.

DID WE MEASURE WHAT WE WANTED TO MEASURE?

Do the operations developed to measure our variables actually do so—are they valid? If we have weighed our measurement options, carefully constructed our questions and observational procedures, and selected sensibly from the available data indicators, we should be on the right track. But we cannot have much confidence in a measure until we have empirically evaluated its *validity.* We must also evaluate the *reliability* of our measures. The reliability of a measure is the degree to which it produces a consistent answer; such reliability (consistency) is a prerequisite for measurement validity.

Measurement Validity

In Chapter 1, you learned that measurement validity refers to how well your indicators measure what they are intended to measure. For instance, a good measure of a person's age is the current year minus the year given on their birth certificate. Very probably, the resulting number accurately represents the person's age. A less valid measure would be for the researcher to ask the person (who may lie, or forget), or for the researcher to simply guess. Measurement validity can be 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 if it is appropriate "on its face." More precisely, we can say that a measure has face validity if it obviously pertains to the meaning of the concept being measured more than to other concepts (Brewer & Hunter, 1989:131). For example, a count of the number of drinks people have 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 in itself does not provide convincing evidence of measurement validity. Face validity has some plausibility but often not much. For instance, let's say that Sara is having some worries about her boyfriend, Jeremy. She wants to know if he loves her. So she asks him, "Jeremy, do you really love me?" He replies, "Sure, baby, you know I do." And yet he routinely goes out with other women, only calls Sara once every three weeks, and isn't particularly nice to her when they do go out. His answer that he loves her has a certain "face validity," but Sara should probably look for other validating measures. (At the least, she might ask her friends if they think his claim has validity.)

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. A measure of student engagement based on how much one talks in class won't count the quiet, but attentive and hardworking, person in the front row. Or if you measure power by listing only elected government officials, you may miss the most important people altogether.

Criterion Validity

Criterion validity is established when the results from one measure match those obtained with a more direct or an already validated measure of the same phenomenon (the "criterion"). A measure of blood-alcohol concentration, for instance, could be the criterion for validating a self-report measure of drinking. In other words, if Jason says he hasn't been drinking, we establish criterion validity by giving him a "breathalyzer" test. Observations of drinking by friends or relatives could also, in some limited circumstances, serve as a criterion for validating a self-report.

The criterion that researchers select can be measured either at the same time as the variable to be validated or after that time. **Concurrent validity** exists when a criterion conducted at the same time yields scores that are closely related to scores on a measure. A store might validate a test of sales ability by administering the test to its current salespeople and then comparing their test scores to their actual sales performance. Or a measure of walking speed based on mental counting might be validated concurrently with a stopwatch. With **predictive validity**, a measure is validated by predicting scores on a criterion measured in the future—for instance, SAT scores are validated when they predict a student's college grades.

Criterion validation greatly increases our confidence that a measure works, but for many concepts of interest to social scientists, it's difficult to find a criterion. Yes, if you and your roommate are together every evening, you can count the beers he drinks. You definitely know about his drinking. But if we are measuring feelings or beliefs or other subjective states, such as feelings of loneliness, what direct indicator could serve as a criterion? How do you know he's lonely? Even with variables for which a reasonable criterion exists, the researcher may not be able to gain access to the criterion—as would be the case with a tax return or employer document that we might wish we could use as a criterion for self-reported income.

Construct Validity

Measurement validity also can be established by relating a measure to other measures specified in a theory. This validation approach, known as **construct validity**, is commonly used in social research when no clear criterion exists for validation purposes.

A historically famous example of construct validity is provided by the work of Theodor W. Adorno, Nevitt Sanford, Else Frenkel-Brunswik, and Daniel Levinson (1950), in their book *The Authoritarian Personality.* Adorno and his colleagues, working in the United States and Germany immediately after World War II, were interested in a question that troubled much of the world during the 1930s and 1940s: Why were so many people attracted to Nazism and to its Italian and Japanese fascist allies? Hitler was not an unpopular leader in Germany. In fact, in January 1933 he came to power by being elected chancellor (something like president) of Germany, although some details of the election were a bit suspicious. Millions of people supported him enthusiastically. Why did so many Germans, during the 1930s, come to nearly worship Adolf Hitler and believe strongly in his program—which proved, of course, to be so disastrous for Europe and the rest of the world? The Adorno research group proposed the existence of what they called an "authoritarian personality," a type of person who would be drawn to a dictatorial leader of the Hitler type. Their key concept, then, was "authoritarianism."

But of course there's no such "thing" as authoritarianism; it's not like a tree, something you can look at. It's a *construct,* an idea that we use to help make sense of the world. To establish construct validity of this idea, the researchers created a number of different scales made up of interview questions. One scale was called the "anti-Semitism" scale, in which hatred of Jews was measured. Another measure was a "fascism" scale, measuring a tendency toward favoring a militaristic, nationalist government. Another was the "political and economic conservatism" scale, and so on. Adorno and his colleagues interviewed lots of Germans and found that high scores on these different scales tended to correlate; a person who scored high on one tended to score high on the others. Hence they determined that the "authoritarian personality" was a legitimate construct. The idea of authoritarianism, then, was validated through construct validity.

In short, a construct ("authoritarianism") was validated through the use of a number of other measures that all tended to be high or low at the same time. Simultaneous high scores on them validated the idea of authoritarianism.

Construct and criterion validation, then, compare scores on one measure to scores on other measures that are predicted to be related. Distinguishing the two forms (construct and criterion) matters less than thinking clearly about the comparison measures and whether they actually represent different views of the same phenomenon. For example, correspondence between scores on two different self-report measures of alcohol use is a weak indicator of measurement validity. A person just reports in two different ways how much she drinks; of course the two will be related. But the correspondence of a self-report measure with an observer-based measure of substance use is a much stronger demonstration of validity. The subject (a) reports how much she drinks, and then (b) an observer reports on the subject's drinking. If the results match up, it's strong evidence of validity.

Reliability

Reliability means that a measurement procedure yields consistent scores (or that the scores change only to reflect actual changes in the phenomenon). 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. Let's say, for example, that you would like to know your weight and have decided on two different measures: the scales in the bathroom and your mother's estimate. Clearly, the scales are more reliable, in the sense that they will show pretty much the same thing from one day to the next, unless your weight actually changes. But your mother, bless her, may say "You're so skinny!" on Sunday; but on Monday, when she's not happy, she may say "You look terrible! Have you gained weight?" Her estimates may bounce around quite a bit. The bathroom scales are not so fickle; they are *reliable.*

This doesn't mean that the scales are *valid*—in fact, if they are spring-operated and old, they may be off by quite a few pounds. But they will be off by the same amount every day—hence not valid, but *reliable* nonetheless.

There are four possible indications of unreliability. For example, a test of your knowledge of research methods would be unreliable if every time you took it you received a different score even though your knowledge of research methods had not changed in the interim, not even as a result of taking the test more than once. This is test-retest reliability. Similarly, an index composed of questions to measure knowledge of research methods would be unreliable if respondents' answers to each question were totally independent of their answers to the others. The index has interitem reliability if the component items are closely related. A measure also would be unreliable if slightly different versions of it resulted in markedly different responses (it would not achieve alternate-forms reliability). Finally, an assessment of the level of conflict in social groups would be unreliable if ratings of the level of conflict by two observers were not related to each other (it would then lack interobserver reliability).

Test-Retest Reliability

When researchers measure an unchanging phenomenon at two different times, the degree to which the two measurements are related is the **test-retest reliability** of the measure. If you take a test of your math ability and then retake the test two months later, the test is reliable if you receive a similar score both times, presuming that your math ability stayed constant. 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.

Interitem Reliability (Internal Consistency)

When researchers use multiple items to measure a single concept, they must be concerned with **interitem reliability** (or internal consistency). For example, if the questions in Exhibit 3.2 reliably measure depression, 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 index.

Alternate-Forms Reliability

When researchers compare subjects' answers to slightly different versions of survey questions, they are testing **alternate-forms reliability** (Litwin, 1995: 13–21). A researcher may reverse the order of the response choices in an index or may modify the question wording in minor ways, and then readminister the index to subjects. If the two sets of responses are not too different, alternate-forms reliability is established.

A related test of reliability is the **split-halves reliability** approach. A survey sample is divided in two by flipping a coin or using some other random assignment method. The two forms of the questions are then administered to the two halves of the sample. If the responses of the two halves of the sample are about the same, the reliability of the measure is established.

Interobserver Reliability

When researchers use more than one observer to rate the same people, events, or places, **interobserver reliability** is their goal. If observers are using the same instrument to rate the same thing, their ratings should be very 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 observers.

Assessing interobserver reliability is most important when the rating task is complex. Consider a commonly used measure of mental health, the Global Assessment of Functioning Scale (GAFS), a bit of which is shown in Exhibit 3.6. The rating task seems straightforward, with clear descriptions of the subject characteristics that are supposed to lead to high or low GAFS 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 subject characteristics, attitudes, and behaviors as well as by the rater's reactions. As a result, interobserver agreement is often low on the GAFS, unless the raters are trained carefully.

Can We Achieve Both Reliability and Validity?

The reliability and validity of measures in any study must be tested after the fact to assess the quality of the information obtained. But then, if it turns out that a measure cannot be considered reliable and valid, little can be done to save the study. Hence it is supremely important to select in the first place measures that are likely to be both reliable and valid. The Dow Jones Industrials Index is a perfectly *reliable* measure of the state of the American economy—any two observers of it will see the same numbers—but its validity is shaky: There's more to the economy than the rise and fall of stock prices. In contrast, a good therapist's interview of a married couple may produce a *valid* understanding of their relationship, but such

Exhibit 3.6 The Challenge of Interobserver Reliability: Excerpt From the Global Assessment of Functioning Scale (GAFS)

Consider psychological, social, and occupational functioning on a hypothetical continuum of mental health-illness. Do not include impairment in functioning due to physical (or environmental) limitations.

Code (Note: Use intermediate codes when appropriate, e.g., 45, 68, 72.)

100 **Superior functioning in a wide range of activities, life's problems never seem to get out**
of hand, is sought by others because of his or her many positive qualities. No
91 **symptoms.**

90 **Absent or minimal symptoms** (e.g., mild anxiety before an exam), **good functioning in all**
areas, interested and involved in a wide range of activities, socially effective, generally
satisfied with life, no more than everyday problems or concerns (e.g., an occasional
81 argument with family members).

80 **If symptoms are present, they are transient and expectable reactions to psychosocial**
stressors (e.g., difficulty concentrating after family argument); **no more than slight**
impairment in social, occupational, or school functioning (e.g., temporarily falling
71 behind in schoolwork).

70 **Some mild symptoms** (e.g., depressive mood and mild insomnia) **OR some difficulty in**
social, occupational, or school functioning (e.g., occasional truancy or theft within the
household), **but generally functioning pretty well, has some meaningful interpersonal**
61 **relationships.**

60 **Moderate symptoms** (e.g., flat affect and circumstantial speech, occasional panic attacks)
OR moderate difficulty in social, occupational, or school functioning (e.g., few friends,
51 conflicts with peers or co-workers).

50 **Serious symptoms** (e.g., suicidal ideation, severe obsessional rituals, frequent shoplifting)
OR any serious impairment in social, occupational, or school functioning (e.g., no
41 friends, unable to keep a job).

40 **Some impairment in reality testing or communication** (e.g., speech is at times illogical,
obscure, or irrelevant) **OR major impairment in several areas, such as work or school,**
family relations, judgment, thinking, or mood (e.g., depressed man avoids friends,
neglects family, and is unable to work, child frequently beats up younger children, is
31 defiant at home, and is failing at school).

30 **Behavior is considerably influenced by delusions or hallucinations OR serious**
impairment in communication or judgment (e.g., sometimes incoherent, acts grossly
inappropriately, suicidal preoccupation) **OR inability to function in almost all areas** (e.g.,
21 stays in bed all day, no job, home, or friends).

20 **Some danger of hurting self or others** (e.g., suicide attempts without clear expectation of
death, frequently violent, manic excitement) **OR occasionally fails to maintain minimal**
personal hygiene (e.g., smears feces) **OR gross impairment in communication** (e.g.,
11 largely incoherent or mute).

10 **Persistent danger of severely hurting self or others** (e.g., recurrent violence) **OR**
persistent inability to maintain minimal personal hygiene OR serious suicidal act with
1 **clear expectation of death.**

0 Inadequate information.

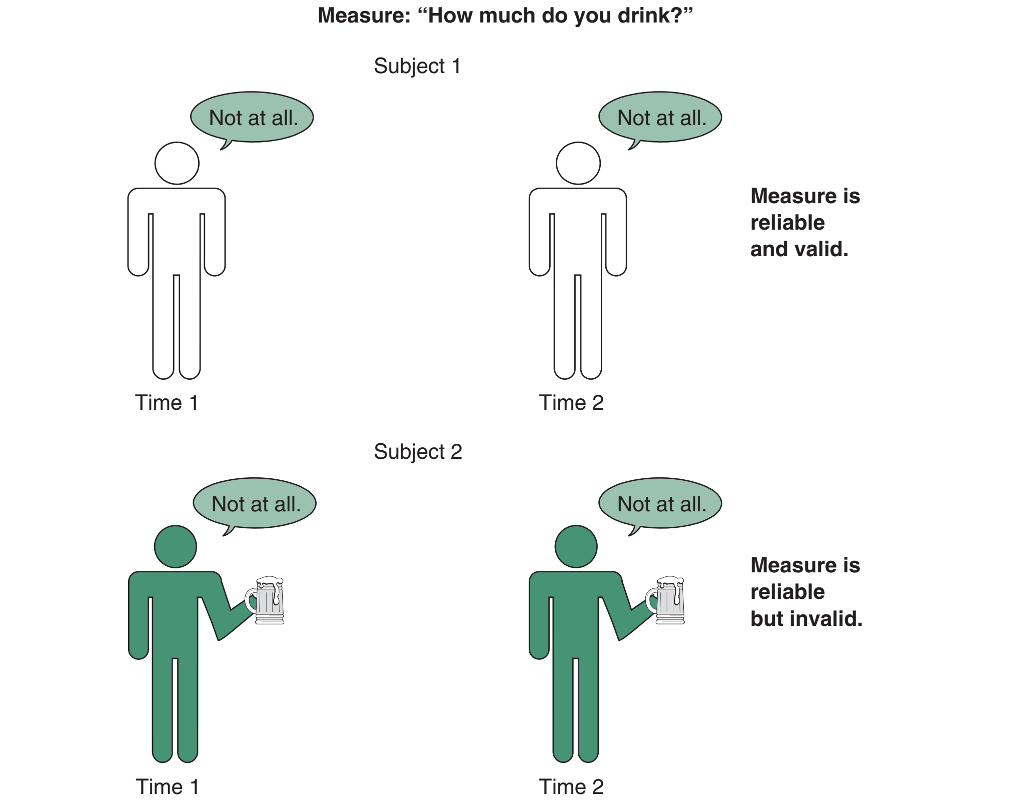
Source: Reprinted with permission from the *Diagnostic and Statistical Manual of Mental Disorders,* Fourth Edition. Copyright © 1994 American Psychiatric Association.

interviews are often not reliable, because another interviewer could easily reach different conclusions.

Finding measures that are both reliable and valid can be challenging. Don't just choose the first measure you find or can think of. 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 measures. 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 even the selection of "tried and true" measures does not absolve researchers from the responsibility of testing the reliability and validity of the measure in their own studies.

Remember that a reliable measure is not necessarily a valid measure, as Exhibit 3.7 illustrates. This discrepancy is a common flaw of self-report measures

Exhibit 3.7 The Difference Between Reliability and Validity: Drinking Behavior


Measure: "How much do you drink?"
Subject 1
Not at all.
Not at all.
Measure is reliable and valid.
Time 1
Time 2
Subject 2
Not at all.
Not at all.
Measure is reliable but invalid.
Time 1
Time 2

of substance abuse. People's answers to the questions are consistent, but they are consistently misleading: A number of respondents will not admit to drinking, even though they drink a lot. The multiple questions in self-report indexes of substance abuse are answered by most respondents in a consistent way, so the indexes are reliable. As a result, some indexes based on self-report are reliable but invalid. Such indexes are not useful and should be improved or discarded.

CONCLUSION

Remember always that measurement validity is a necessary foundation for social research. 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. The examples in this chapter of difficulties in achieving valid measures should sensitize you to the need for caution.

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, don't trust the findings. And 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
Closed-ended (fixed-choice) question
Concept
Conceptualization
Concurrent validity
Constant
Construct validity
Content analysis
Content validity
Criterion validity
Exhaustive
Face validity
Index
Interitem reliability
Interobserver reliability
Interval level of measurement
Level of measurement
Mutually exclusive
Nominal level of measurement
Open-ended question
Operationalization
Operations
Ordinal level of measurement
Predictive validity
Ratio level of measurement
Reliability
Scale
Split-halves reliability
Test-retest reliability
Triangulation
Unobtrusive measure

HIGHLIGHTS

- Conceptualization plays a critical role in research. In deductive research, conceptualization guides the operationalization of specific variables; in inductive research, it guides efforts to make sense of related observations.

- 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 are operationalized in research by one or more indicators, or measures, which may derive from observation, self-report, available records or statistics, books and other written documents, clinical indicators, discarded materials, or some combination.

- Indexes and scales measure a concept by combining answers to several questions and so reducing idiosyncratic variation. Several issues should be explored with every intended index: Does each question actually measure the same concept? Does combining items in an index obscure important relationships between individual questions and other variables? Is the index multidimensional?

- If differential weighting, based on differential information captured by questions, is used in the calculation of index scores, then we say that the questions constitute a scale.

- 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 (or qualitative), ordinal, interval, and 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 (either predictive or concurrent), and construct validation. Criterion validation provides the strongest evidence of measurement validity, but often there is no criterion to use in validating social science measures.

- Measurement reliability is a prerequisite for measurement validity, although reliable measures are not necessarily valid. Reliability can be assessed through a test-retest procedure, an interitem comparison of responses to alternate forms of the test, or the consistency of findings among observers.

To assist you in completing the Web Exercises, please access the Study Site at http://www.pineforge.com/mssw2 where you'll find the Web Exercises with accompanying links. You'll find other useful study materials like self-quizzes and e-flashcards for each chapter, along with a group of carefully selected articles from research journals that illustrate the major concepts and techniques presented in the book.

EXERCISES

Discussing Research

1. Pick one important, frequently used concept such as "power," "the economy," "fundamentalism," "poverty," "authoritarianism," "racism," or some other concept suggested by your instructor. Then find five uses of it in newspapers, magazines, or journals. Is the concept defined clearly in each article? How similar are the definitions? Write up what you have found in a short report.

2. Look at our definition of "terrorism" on the first page of this chapter. Do you agree with this definition? For each component of the definition (e.g., "political," "nongovernmental," etc.) give an example that might contradict the definition. Propose an alternative definition that might be better.

3. Do you and your classmates share the same beliefs about the meanings of important concepts? First, divide your class into several groups, each group having at least six students. Then, assign each group a concept from the following list: violence, stress, social support, mental illness, social norms. In each group, each student should independently write a brief definition of his or her concept and some different examples supporting it. Finally, all students who have worked on the same concept should meet together, compare their definitions and examples, and try to reach agreement on the meaning of the concept. Discuss what you learned from this effort.

4. Propose an open-ended question to measure one of the concepts you discussed in the preceding exercises. Compare your approach to those adopted by other students.

Finding Research

1. What are some of the research questions you could attempt to answer with available statistical data? Visit your library and ask for an introduction to the government documents collection. Inspect the U.S. Bureau of the Census Web site (www.census.gov) and find the population figures broken down by city and state. List five questions that you could explore with such data. Identify six variables implied by these research questions that you could operationalize with the available data. What are three factors that might influence variation in these measures, other than the phenomenon of interest? (Hint: Consider how the data are collected.)

2. 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 (as you define it). Now go to the National Council on Alcohol and Drug Dependence (NCADD) Web site (www.ncadd.org/facts/defalc.html) and read their official "Definition of Alcoholism." What is the definition of alcoholism used by NCADD? How is alcoholism conceptualized? How does this compare to your definition?

Critiquing Research

1. Shortly before the year 2000 national census of the United States, a heated debate arose in Congress over whether instead of a census—a total headcount—a sample should

Exhibit 3.8 Selected Shelter Staff Survey Questions

1. What is your current job title? ____

2. What is your current employment status?
   Paid, full-time 1
   Paid, part-time (less than 30 hours per week) 2

3. When did you start your current position? ________ / ________ / ________
   Month Day Year

4. In the past month, how often did you help guests deal with each of the following types of problems? (*Circle one response on each line*.)

	Very often		----------	----------	----------		Never
Job training/placement...................	1	2	3	4	5	6	7
Lack of food or bed.......................	1	2	3	4	5	6	7
Drinking problems.........................	1	2	3	4	5	6	7

5. How likely is it that you will leave this shelter within the next year?
   Very likely... 1
   Moderately... 2
   Not very likely... 3
   Not likely at all... 4

6. What is the highest grade in school you have completed at this time?
   First through eighth grade... 1
   Some high school... 2
   High school diploma... 3
   Some college... 4
   College degree... 5
   Some graduate work... 6
   Graduate degree... 7

7. Are you a veteran?
   Yes... 1
   No... 2

Source: Schutt & Fennell, 1992.

be used to estimate the number and composition of the U.S. population. As a practical matter, might a sample be more accurate in this case than a census? Why?

2. Develop a plan for evaluating the validity of a measure. Your instructor will give you a copy of a questionnaire actually used in a study. Pick out one question and define the concept that you believe it is intended to measure. Then develop a construct validation strategy involving other measures in the questionnaire that you think should be related to the question of interest—if it measures what you think it measures.

3. The questions in Exhibit 3.8 are selected from a survey of homeless shelter staff (Schutt & Fennell, 1992). First, identify the level of measurement for each question. Then

rewrite each question so that it measures the same variable but at a different level. For example, you might change a question that measures age at the ratio level, in years, to one that measures age at the ordinal level, in categories. Or you might change a variable measured at the ordinal level to one measured at the ratio level. For the categorical variables, those measured at the nominal level, try to identify at least two underlying quantitative dimensions of variation, and write questions to measure variation along these dimensions. For example, you might change a question asking which of several factors the respondent thinks is responsible for homelessness to a series of questions that ask how important each factor is in generating homelessness.

What are the advantages and disadvantages of phrasing each question at one level of measurement rather than another? Do you see any limitations on the types of questions for which levels of measurement can be changed?

4. A lengthy assessment of different measures of substance abuse is available at a site maintained by the National Institute on Alcoholism and Alcohol Abuse (http://www.niaaa.nih.gov/publications/Assesing%20Alcohol/selfreport.htm). Read through the summary of instruments (or review the instrument "fact sheets" at this NIAAA site). Pick two instruments. What concept of substance abuse is reflected in each measure? Is either measure multidimensional? What do you think the relative advantages of each measure might be? What evidence is provided about their reliability and validity? What other test of validity would you suggest?

Doing Research

1. Some people have said in discussions of international politics that "democratic governments don't start wars." How could you test this hypothesis? Clearly state how you would operationalize (1) "democratic," and (2) "start."

2. Now it's time to try your hand at operationalization with survey-based measures. Formulate a few fixed-choice questions to measure variables pertaining to the concepts you researched for Exercise 1 under "Discussing Research," such as poverty, power, or racism. Arrange to interview one or two other students with the questions you have developed. Ask one fixed-choice question at a time, record your interviewee's answer, and then probe for additional comments and clarifications. Your goal is to discover what respondents take to be the meaning of the concept you used in the question and what additional issues shape their response to it.

When you have finished the interviews, analyze your experience: Did the interviewees interpret the fixed-choice questions and response choices as you intended? Did you learn more about the concepts you were working on? Should your conceptual definition be refined? Should the questions be rewritten, or would more fixed-choice questions be necessary to capture adequately the variation among respondents?

3. Now try index construction. You might begin with some of the questions you wrote for Exercise 2. Write four or five fixed-choice questions that each measure the same concept. (For instance, you could ask questions to determine whether someone is alienated.) Write each question so it has the same response choices (a "matrix" design). Now conduct a literature search to identify an index that another researcher used to measure your concept

or a similar concept. Compare your index to the published index. Which seems preferable to you? Why?

4. List three attitudinal variables.

   a. Write a conceptual definition for each variable. Whenever possible, this definition should come from the existing literature—either a book you have read for a course or the research literature that you have been searching. Ask two class members for feedback on your definitions.
   b. Develop measurement procedures for each variable: Two measures should be single questions and one should be an index used in prior research (search the Internet and the journal literature in Soc Abstracts or Psych Abstracts). Ask classmates to answer these questions and give you feedback on their clarity.
   c. Propose tests of reliability and validity for the measures.

5. Exercise your cleverness on this question: For each of the following, suggest two unobtrusive measures that might help you discover (a) how much of the required reading for this course students actually complete; (b) where are the popular spots to sit in a local park, and (c) which major U.S. cities have the highest local taxes.