CHAPTER 4

CONCEPTUALIZATION AND MEASUREMENT

LEARNING OBJECTIVES

1. Define and distinguish *conceptualization* and *operationalization.*
2. List four different means of operationalizing concepts.
3. Give two examples of constant and two of variable phenomena.
4. Identify the different forms of single questions and response choices.
5. Give examples of the four levels of measurement.
6. Compare the advantages and disadvantages of the three approaches to testing the validity of measures.
7. Define the five methods of evaluating measurement reliability.

©iStockphoto.com/Terminator3D

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.* You could then measure terrorism by counting, say over a 5-year period, the number of violent attacks that have explicit political aims. You will also need to define and measure *religious fundamentalism*—an even more difficult 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? 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: 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 (e.g., social health) 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 biologically female (gender—man or woman, say—is a bit trickier). 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, at least by contemporary American definitions. Do cousins living next door count? Cousins living in California? Or maybe the parents are divorced, the children are adopted, or 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 does biology? Or is it a connecting 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* affects your results.

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

Often social concepts can be used sloppily or even misleadingly. In some years, you may hear that "the economy" is doing well, but even then, many people may be faring badly. Typically in news reports, *the economy* refers to the gross domestic product (GDP)—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 this person would have 30 years ago—in fact, the average American man makes a little less than 30 years ago, and for women it's close. We could use the concept of *the economy* to refer to the economic well-being of actual people, but that's not typically how it's used.

Defining concepts clearly can be 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 (1947/1997: 152), 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, very well known, 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 always 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 that person's acts over the past 10 years and see when the person exerted his or her will over others? Should we try to uncover the true extent of the individual's 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.

Likewise, describing what causes *crime*, or even what causes *theft*, is inherently problematic because 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 who is 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 is the debates about, for instance, what is a theft or what is crime.

Conceptualization: The process of specifying what we mean by a term. In deductive research, conceptualization helps 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.

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 (gender, 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 terms (family, sex, happiness, power) to clarified definitions for them—really, figuring out what are the social "things" you'll be talking about.

It is especially important to define clearly concepts that are abstract or unfamiliar. When we refer to such concepts as *social capital, whiteness*, or *dissonance*, we cannot count on others knowing exactly what we mean. Even experts may disagree

Research That Matters

Excessive use of alcohol, illicit drugs, and cigarettes predict long-term differences in the life course. Bohyun Joy Jang and Megan Patrick at the University of Michigan and Megan Schuler at the Harvard Medical School studied whether substance use by young adults predicts delays in family formation.

The concept of substance use was measured with three questions about frequency of smoking cigarettes, binge drinking, and using marijuana. Their measures of the concept of family formation were questions about their marital, cohabitation, and parental status.

By the end of the chapter, you will understand why defining concepts and developing measures are critical steps in research.

Source: Adapted from Jang, Bohyun Joy, Megan E. Patrick, and Megan S. Schuler. 2018. Substance use behaviors and the timing of family formation during young adulthood. *Journal of Family Issues* 39(5).

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; rather, 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? Smartphones, for instance, may be seen as communication devices, like telephones, radios, telegraphs, or two tin cans connected by a string. Or they can be seen primarily as entertainment devices, like television sets or basketballs. Or they can be conceptualized as being essentially devices for the government to track our activities with—a kind of electronic ankle bracelet that everyone voluntarily carries around. Or they can also be conceived in yet another way: A college administrator we know, seeing students leaving class outside her building, said, "Phones have replaced cigarettes." She reconceptualized smartphones, seeing them not as communication tools but as something to nervously fiddle with, like cigarettes, chewing gum wrappers, keys on a lanyard, or the split ends of long hair—just "something to do." 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, when they enjoy their courses, and so on. We are interested, in other words, in changes in engagement—how and

In the News

Research in the News

Are Teenagers Replacing Drugs With Smartphones?

As high school age teens' use of smartphones and tablets has accelerated in recent years, their use of illicit drugs other than marijuana has been dropping. Could the first trend be responsible to some extent for the second? Substance abuse expert Dr. Silvia Martins, at Columbia University, thinks this "is quite plausible." According to Dr. Nora Volkow, the director of the National Institute on Drug Abuse, "teens can get literally high when playing these [computer] games." Teens quoted in the article agreed, but other experts proposed other explanations. Professor James Anthony at Michigan State University admitted that "there is very little hard, definitive evidence on the subject."

For Further Thought

1. Should the concept of "addiction" be applied to behavior on modern technology devices? How would you define the concept of addiction?
2. Can we depend on self-report measures of drug (and technology) use?

when it varies. Engagement, then, is a *variable*; it can be high, or it can be low. It's not just a fixed thing. Next, when we try to explain those different levels of student engagement (what causes them), we have to talk about changes in still other things—for instance, in who the teacher is, or what subject teachers offer, or what pedagogical techniques the teachers use. The whole effort to explain something relies on saying, basically, that a change in A causes a change in B. So both A and B have to be changeable things: They must be what scientists call *variables*.

We could use any number of variables to measure engagement: the student's reported interest in classes, teacher evaluations of student engagement, the number of hours spent on homework, or an index summarizing 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.

Constant: A number that has a fixed value in a given situation; a characteristic or value that does not change.

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 a **constant** (it's always the same), not a variable. You can't explain, for instance, low student engagement in classes by just saying "students are just like that, that's all." If engagement varies, it can only be explained by another variable, not by something that's a constant, or always the case. Or to take a different example, if you studied binge drinking in all-male fraternities, you might believe that the male atmosphere matters. But unless you compared them with female groups (sororities, say), gender wouldn't be a variable in your research—because it wouldn't "vary"—and you couldn't determine if gender made a difference.

As mentioned, many variables could be used to measure student engagement. Which ones 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 phenomenon 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 analysis?

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—after "conceptualizing"—and we have identified the variables that we want to measure, we must develop our exact measurement procedures; we need to specify the **operations** for measuring the variables we've chosen.

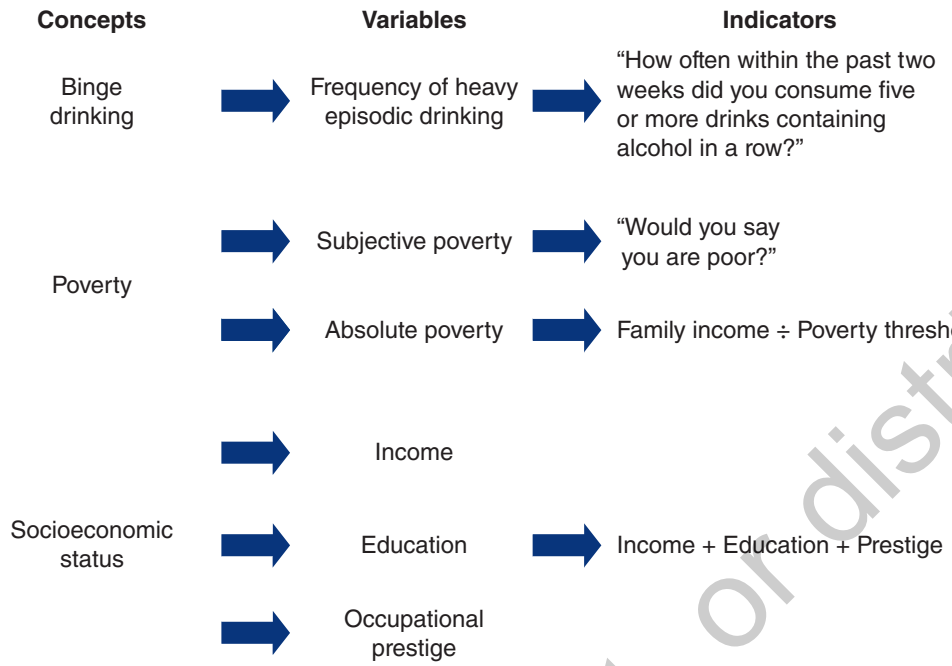
Operation: A procedure for identifying or indicating the value of cases on a variable.

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

Exhibit 4.1 represents the **operationalization** process for three different concepts. The first researcher defines her concept, binge drinking, and chooses one variable—frequency of heavy episodic drinking—to represent it. This variable is then measured by a specific *indicator*, which in this case will be responses to a single question: "How often within the last 2 weeks did you consume five or more drinks containing alcohol in a row?" (Because "heavy" drinking is defined differently for men and women, the question is phrased as "four or more drinks" for women.) The researcher—moving from left to right on the chart—developed a concept, chose a variable to measure it, then specified the exact *operation* for measuring that variable. Operationalization is the process of turning an abstract concept into a clearly measured variable.

The second researcher defines her concept—poverty—in a more complicated way. She decides that being poor has both subjective and objective components, and both should be measured. (In the research literature, these components are referred to as "subjective" and "absolute" poverty—*absolute* meaning that it's not compared to other people but to some objective standard.) The variable subjective poverty is then measured (operationalized) with responses to a survey question: "Would you say that you are poor?" Absolute poverty, however, is measured by comparing family income to the poverty threshold. The researcher has operationalized her concept in two different ways.

Exhibit 4.1 /// Concepts, Variables, and Indicators: Operationalizing Concepts



Finally, the third researcher decides that his concept—socioeconomic status—is multidimensional and should be operationalized by three different variables put together: (1) income, (2) education, and (3) occupational prestige. Only all three of these combined, he feels, really capture what we mean by socioeconomic status. So he picks indicators for each, and then puts those all together to provide ratings of a person's socioeconomic status. Three different operations are used to define socioeconomic status.

Indicators 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, but others do not. In this chapter, we will briefly introduce the operations of using published data, doing content analysis, asking questions, and observing behavior. All of these are covered in more detail later.

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 the junior year have a more valuable experience than do those who remain at the 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—for example, asking students in each program, "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 considered 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.

A historical question about the types of social bonds among combat troops in wars since 1940 probably requires retrospective interviews with surviving veterans. The validity of the data is lessened by the unavailability of many veterans from World War II and by problems of recall, but direct observation of their behavior during the war is certainly not an option.

Using Available Data

Data can be collected in a wide variety of ways; indeed, much of this book describes different technologies for data collection. But some data are already gathered and ready for analysis (such data will be described in more detail in Chapters 8 and 11). Government reports, for instance, 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. Data from 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, other jobs are hard to find, and people live on low wages, then 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 (i.e., 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 midwestern county; an intensive search in that county in 1992 located 172 churches (Hadaway, Marler,

and Chaves 1993: 744). Still, when legal standards, enforcement practices, and measurement procedures have been considered, comparisons among communities become more credible.

However, such adjustments may be less necessary when the operationalization of a concept is seemingly unambiguous, as with the homicide rate: After all, dead is dead, right? 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 even here, careful review of measurement operations is still important because (for instance) procedures for classifying a death as a homicide can vary between jurisdictions and over time.

Another rich source of already-collected data is survey data sets archived and made available to university researchers by the Inter-university Consortium for Political and Social Research (1996). One of its most popular survey data sets 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. College students can easily use this data set 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?

Content Analysis

Content analysis: A research method for systematically and quantitatively analyzing characteristics of messages.

One particular method for using available data is **content analysis**, a method for systematically and quantitatively analyzing characteristics of messages (Neuendorf 2002: 1). You can think of a content analysis as a "survey" of messages, 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. As a simple example of content analysis, you might look at a variety of women's magazines from 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 during 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 and Miles 1994). The computer is programmed with certain rules for coding text so that these rules will be applied consistently. We describe content analysis in detail in Chapter 11.

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 can be measured with such simple questions. In this section, we introduce some options for writing 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 7; 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 Fund for the Improvement of Postsecondary Education (FIPSE) Core Analysis Grantee Group (Presley, Meilman, and Lyerla 1994).

Closed-ended (fixed-choice) question: A survey question that provides preformatted response choices for the respondent to circle or check.

> *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.

Mutually exclusive: A variable's attributes (or values) are mutually exclusive when every case can be classified as having only one attribute (or value).

Exhaustive: Every case can be classified as having at least one attribute (or value) for the variable.

Most important, response choices should be **mutually exclusive** and **exhaustive**, so that respondents can each find *one and only one* choice that applies to them (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 might skip the question entirely or choose a response option that does not indicate what they are really thinking.

Open-ended question: A survey question to which respondents reply in their own words, either by writing or by talking.

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 1987." "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

Index: A composite measure based on summing, averaging, or otherwise combining the responses to multiple questions that are intended to measure the same concept.

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 forms 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 proven 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 the comparison of findings from other studies.

The questions in Exhibit 4.2 represent a short form of an index used to measure depression; it is called the Center for Epidemiologic Studies Depression Index (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.)

Exhibit 4.2 /// Examples of Indexes: Short Form of the Center for Epidemiologic Studies Depression Index (CES-D) and "Negative Outlook" Index

CES-D Index			
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
Negative Outlook Index			
How often was each of these things true during the past week? (Circle one response on each line)	**A Lot, Most, or All of the Time**	**Sometimes**	**Never or Rarely**
a. You felt that you were just as good as other people.	0	1	2
b. You felt hopeful about the future.	0	1	2
c. You were happy.	0	1	2
d. You enjoyed life.	0	1	2

Source: Hawkins, Daniel N., Paul R. Amato, and Valarie King. 2007. Nonresident father involvement and adolescent well-being: Father effects or child effects? *American Sociological Review* 72: 990.

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 results from rape or incest" and a 4 to agreement with the statement that abortion should be allowed "whenever a woman decides she wants one." A 4 rating is much stronger, in that anyone who gets a 4 would probably agree to all lower-number questions as well.

Scale: A composite measure based on combining the responses to multiple questions pertaining to a common concept after these questions are differentially weighted, such that questions judged on some basis to be more important for the underlying concept contribute more to the composite score.

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 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 Daniel 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 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 modified his original theory 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 their own senses and perspectives filter what they do see. 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 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 9, but it is important to begin to consider them whenever you read about observational measures.

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 considers 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. Conversely, 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 multiple methods to study one research question.

Triangulation—the use of two or more different measures of the same variable—can strengthen measurement considerably (Brewer and 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 produce different results. In a contrasting example, postcombat interviews of U.S. 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 1947/1978). 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 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 4 to 10 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 the car is stolen, the victim will definitely report it to the police to claim the insurance. Plus, some other people might report cars stolen when they haven'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 the person wasn't murdered. There he or she is, 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.

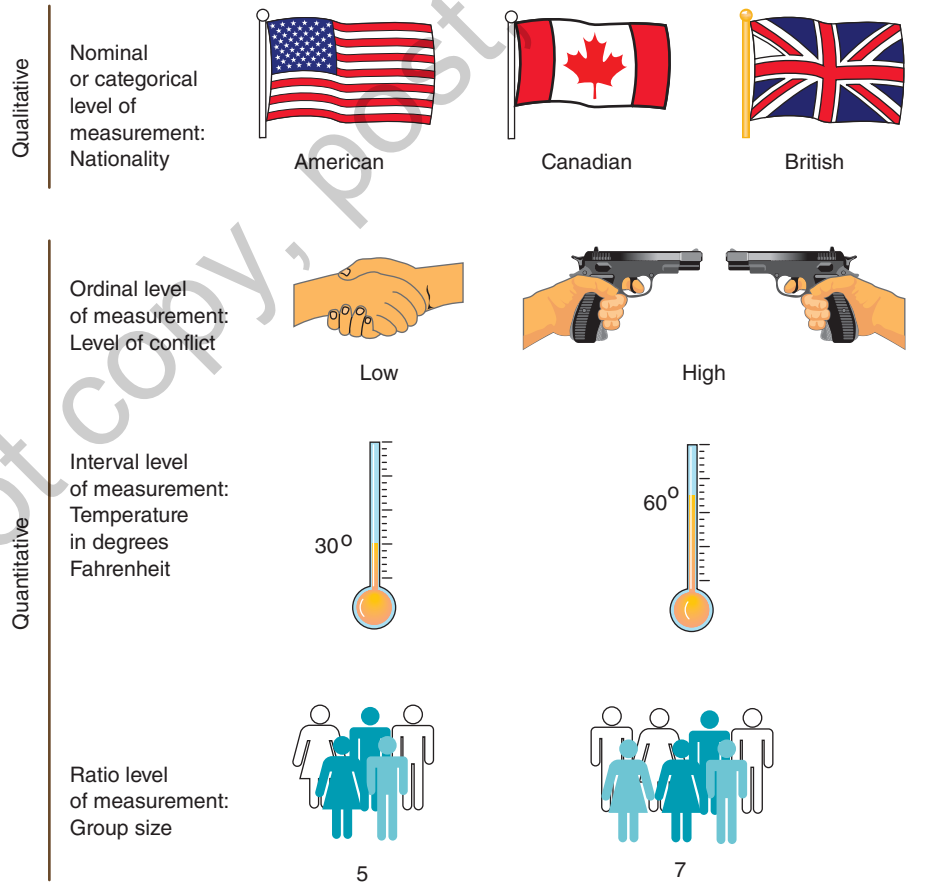
So if you can, it's best to use multiple measures of the same variable; that way, each measure helps to check the validity of the others.

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.

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 these data contain mathematically detailed information; they represent a higher level of measurement. There are four **levels of measurement**: (1) nominal, (2) ordinal, (3) interval, and (4) ratio. Exhibit 4.3 depicts the differences among these four levels.

Exhibit 4.3 /// Levels of Measurement



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, Muslim, 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 4.3). 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.

Nominal level of measurement: Variables whose values have no mathematical interpretation; they vary in kind or quality but not in amount.

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.

Ordinal level of measurement: A measurement of a variable in which the numbers indicating a variable's values specify only the order of the cases, permitting *greater than* and *less than* distinctions.

The properties of variables measured at the ordinal level are illustrated in Exhibit 4.3 by the contrast between the levels 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 1 to the low-conflict group and 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: A measurement of a variable in which the numbers indicating a variable's values represent fixed measurement units but have no absolute, or fixed, zero point.

Interval Level of Measurement

At the **interval level of measurement**, numbers represent fixed measurement units but have no absolute zero point. For example, in America temperatures are

measured on the Fahrenheit scale (see Exhibit 4.3), in which "zero" degrees isn't really "no heat"; it just is defined as the temperature around which concentrated salt water freezes. (Most of the world uses the Celsius scale, in which pure water freezes at 0 degrees and boils at 100 degrees). So 60 degrees Fahrenheit isn't really "twice as hot" as 30 degrees. Still, saying there was a "30-degree temperature increase" since yesterday definitely provides more information than just saying, "It's hotter today," which would be an ordinal description. Interval measures provide more information.

Sometimes social scientists create internal-level measures by combining responses to a series of ordinal measurements into an index. An index, for instance, could be created with responses to the Core Institute's questions about friends' disapproval of substance use (Exhibit 4.4). The survey has 13 questions, each of which has 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 4 more units than a score of 16. But it would still be a little misleading to say a 39 is "three times as disapproving" as a 13.

Ratio level of measurement: A measurement of a variable in which the numbers indicating the variable's values represent fixed measuring units *and* an absolute zero point.

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 (e.g., money, or the number of books in a house). Ratio numbers can be added and subtracted; and because the numbers begin at a truly absolute

Exhibit 4.4 /// 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 bodybuilding or improved athletic performance	○	○	○

Source: Core Institute 1994. *Core alcohol and drug survey.* Carbondale, IL: Core Institute.

zero point, they can also be multiplied and divided (so ratios can be formed between the numbers). Because they carry more information, they can be used in more complex data analyses.

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 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 have to be any "group" with a size of zero; 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 4.5 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*; as 0 to 10, 11 to 20, 21 to 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.

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

Exhibit 4.5 /// 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	× (/)				✓

Careers and Research

Dana Hunt, PhD, Principal Scientist

Dana Hunt

In the study site video for this chapter, Dana Hunt discusses two of the many lessons she has learned about measurement in a decades-long career in social research. Hunt received her BA in sociology from Hood College in Pennsylvania and then earned her PhD in sociology at the University of Pennsylvania. After teaching at Hood for several years, she took an applied research position at National Development and Research Institutes (NDRI) in New York City. NDRI's description on its website gives you an idea of what drew the attention of a talented young social scientist.

Founded in 1967, NDRI is a nonprofit research and educational organization dedicated to advancing scientific knowledge in the areas of drug and alcohol abuse, treatment, and recovery; HIV, AIDS, and HCV (hepatitis C virus); therapeutic communities; youth at risk; and related areas of public health, mental health, criminal justice, urban problems, prevention, and epidemiology.

Hunt moved from New York to the Boston area in 1990, where she is now a principal scientist at Abt Associates, Inc., in Cambridge, a large for-profit government and business research and consulting firm. Abt Associates applies scientific research, consulting, and technical assistance expertise on a wide range of issues in social, economic, and health policy; international development; clinical trials; and registries.

Two of Hunt's major research projects in recent years are the nationwide Arrestee Drug Abuse Monitoring Program for the Office of National Drug Control Policy and a study of prostitution and sex trafficking demand reduction for the National Institute of Justice.

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 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 will asking respondents for their income in dollars.

DID WE MEASURE WHAT WE WANTED TO MEASURE?

A good measurement needs to be both *valid* and *reliable*. "Valid," as we've discussed in Chapter 1, means that an operation should actually measure what it's supposed to. "Reliable" means that a measurement produces essentially the same result and time you use it; it's stable.

Measurement Validity

Let's start with validity. To determine a person's age, you could try to measure by (a) guessing, or (b) asking them. Guessing can be wildly inaccurate; it's not a very

"valid" measure. Asking is probably better. But they may still lie, or even forget, so validity is still a bit shaky. Finally, you could obtain the person's birth certificate, read the year given, and subtract that from the current year. The result is likely to be a valid measure of the person's age. That would be ideal, although usually just asking is probably sufficient.

Measurement validity can be assessed in several ways: (1) face validation, (2) criterion validation, and (3) construct validation.

Face Validity

Face validity (the simplest kind) is 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 and Hunter 1989: 131). For example, a count of the number of drinks people have consumed in the past week would be a measure of their alcohol consumption that has face validity. It just seems obviously appropriate.

Face validity: The type of validity that exists when an inspection of items used to measure a concept suggests that they are appropriate "on their face."

Although every measure should initially be inspected in this way, face validity is not scientifically convincing. Face validity helps, 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 (her measurement!), "Jeremy, do you really love me?" He replies, "Sure, baby, you know I do." That's face validity; she doesn't think he's a shameless liar. And yet Jeremy routinely goes out with other women, only calls Sara once every 3 weeks, and isn't particularly nice to her when they do go out. His answer that he loves her has face validity, but Sara should probably look for other validating measures—for instance, how he actually treats her and their relationship.

Criterion Validity

Much stronger (and more scientifically sophisticated) than face validity is **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.

Criterion validity: The type of validity that is established by comparing the scores obtained on the measure being validated to those obtained with a more direct or already validated measure of the same phenomenon (the criterion).

Criterion validity is established, then, when a more direct measure of a phenomenon regularly points to the same answer as the measure we seek to validate. A store might validate a written test of sales ability comparing test scores to peoples' actual sales performance. Or, a measure of walking speed based on mental counting might be validated with a stopwatch. Sometimes a criterion measured in the future can validate a measure—for instance, if SAT scores accurately predict college grades, that would validate the SAT.

Behaviors may be easy to measure. If you and your roommate are together every evening, you can actually count the beers he seems to be drinking every night. You definitely know about his drinking. But for many concepts social scientists are interested in—for instance, human emotions—it's difficult to find a well-established

criterion. Suppose you want to measure your roommate's feelings of social awkwardness or exclusion; what direct indicator could serve as a criterion? How do you really know if he's feeling bad? A tax return can validate self-reported income, but what would you use to measure misery?

Construct Validity

Finally, when no clear criterion exists, measurement validity can be established by relating a measure to other measures, used in a theory. Different parts of a theory should "hang together"; if they do, this helps to validate the measures. This approach is known as **construct validity**.

Construct validity: The type of validity that is established by showing that a measure is related to other measures as specified in a theory.

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 named chancellor (something like president) of Germany, following a bitterly divided election. Millions of people supported him enthusiastically, although more did not.

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 "construct," 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 measure this idea, then, the researchers created a number of different scales made up of interview questions. Each scale was to measure one element of Nazi authoritarianism. One scale was called the "anti-Semitism" scale, in which hatred of Jews was measured. Another was a "fascism" scale, measuring a tendency toward favoring a militaristic, nationalist government. Still 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 was validated through construct validity.

In a more contemporary example, A. Thomas McLellan and his associates (1985) developed a list of questions called the Addiction Severity Index (ASI), which they believed would measure levels of substance abuse. They did not have more direct measures, such as observation reports, so they couldn't use criterion validation—there were no solid criteria available.

However, prior research had suggested that substance abuse is often related to problems with physical and mental health, employment, and family relationships. And in fact, they found that individuals with higher ASI scores did indeed suffer more in all of these areas—providing construct validation of their index.

Both criterion and construct validity work by comparing results of one measure with some other measure that you think is probably related, and seeing if they match up. The vital step, though, is to make sure that the two measures are really independently produced. For example, if you ask a person two different questions about their own drinking ("Are you a heavy drinker?" and "How many drinks do you have in a week?"), of course they will be related; the same person gave both answers to questions on the topic. You aren't really establishing the validity of either. But if you compare one such self-report answer with, say, the report of an outside observer, then if these two match up you've established some validity.

Reliability

Reliability means that a measurement yields consistent scores (so scores change only when the phenomenon changes). 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 being valid but *reliable* nonetheless.

Establishing reliability of a measure is much more straightforward than establishing validity. Essentially, you will be comparing the measure with itself, in various ways. 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. This is **test–retest reliability**. The test would have **interitem reliability (internal consistency)** if doing well on some questions (items) matched up with doing well on others. When the wording of questions is altered slightly, your overall grade should still stay roughly the same (**alternate-forms reliability**). If you make an A on the first half of the test, you shouldn't get an F on the second half (**split-halves reliability**). Finally, whether your professor, or your TA, or another expert in the field evaluates your test shouldn't affect your grade (**interobserver reliability**).

Reliability: A measurement procedure yields consistent scores when the phenomenon being measured is not changing.

Test–retest reliability: A measurement showing that measures of a phenomenon at two points in time are highly correlated, if the phenomenon has not changed or has changed only as much as the phenomenon itself.

Interitem reliability (internal consistency): An approach that calculates reliability based on the correlation between multiple items used to measure a single concept.

Alternate-forms reliability: A procedure for testing the reliability of responses to survey questions in which subjects' answers are compared after the subjects have been asked slightly different versions of the questions or when randomly selected halves of the sample have been administered slightly different versions of the questions.

Split-halves reliability: Reliability achieved when responses to the same questions by two randomly selected halves of a sample are about the same.

Interobserver reliability: When similar measurements are obtained by different observers rating the same persons, events, or places.

Can We Achieve Both Reliability and Validity?

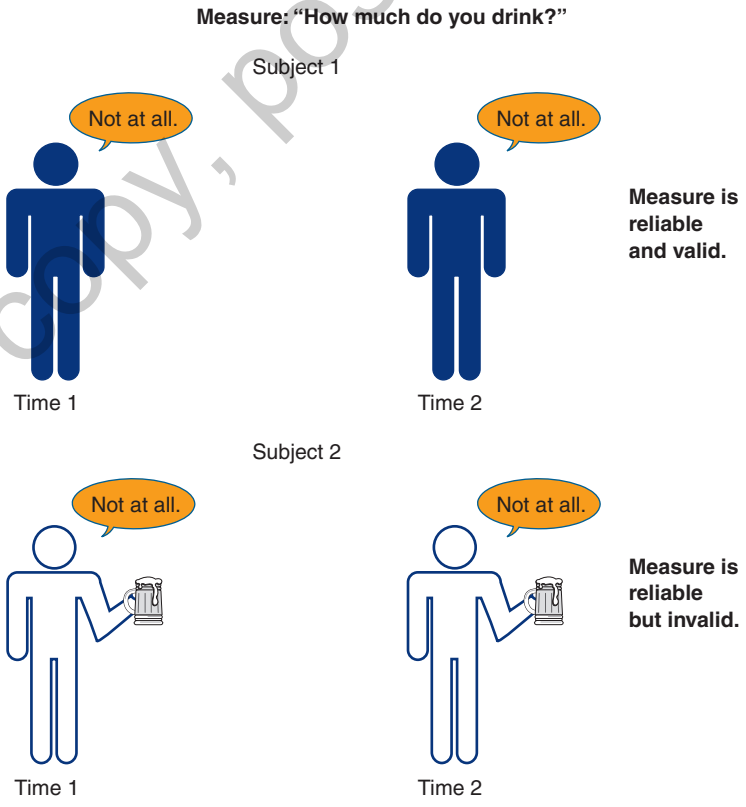
The reliability and validity of measures in any study must be tested after the fact to assess the quality of the information obtained. But then, if it turns out that a measure cannot be considered reliable and valid, little can be done to save the study. Hence, it is supremely important to select in 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 U.S. 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 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 4.6 illustrates. The discrepancy shown is a common flaw of self-report measures of substance abuse. People's answers to the questions are consistent (reliable), but they are consistently misleading (not valid): A number of respondents will not admit to drinking, even though they drink a lot. Most respondents answer the

Exhibit 4.6 /// The Difference Between Reliability and Validity: Drinking Behavior



multiple questions in self-report indexes of substance abuse 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 you 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

/// 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 three basic approaches: face validation, criterion validation, 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 component measures within an index, a comparison of responses to alternate forms of the test or by randomly selected ("split") halves of a sample to the same test, or the consistency of findings among observers.

/// STUDENT STUDY SITE

SAGE edge™

The Student Study Site, available at **edge.sagepub.com/chamblissmssw6e**, includes useful study materials including practice quizzes, eFlashcards, videos, audio resources, journal articles, and more.

/// EXERCISES

Discussing Research

1. What does *trust* mean to you? Identify two examples of "trust in action," and explain how they represent your concept of trust. Now develop a short definition of *trust* (without checking a dictionary). Compare your definition to those of your classmates and what you find in a dictionary. Can you improve your definition based on some feedback?
2. What questions would you ask to measure the level of trust among students? How about feelings of being "in" or "out" with regard to a group? Write five questions for an index, and suggest response choices for each. How would you validate this measure using a construct validation approach? Can you think of a criterion validation procedure for your measure?
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 the researcher could keep track of who responded? What criterion validation procedure would you suggest for assessing measurement validity?

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. Census

Bureau website (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 American Council for Drug Education, an affiliate of Phoenix House, and read some their facts about alcohol (http://www.phoenixhouse.org/prevention/). Is this information consistent with your definition?

Critiquing Research

1. Shortly before the year 2000 national census of the United States, a heated debate arose in Congress about whether instead of a census—a total headcount—a sample should 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 4.7 are selected from a survey of homeless shelter staff (Schutt and 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.
4. 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?

Exhibit 4.7 /// 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

(Continued)

Exhibit 4.7 /// (Continued)

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: Based on Schutt, Russell K. 1988. *Working with the homeless: The backgrounds, activities and beliefs of shelter staff*. Boston: University of Massachusetts. Unpublished report: 7–10, 15, 16. Results reported in Schutt, Russell K., and M. L. Fennell. 1992. Shelter staff satisfaction with services, the service network, and their jobs. *Current Research on Occupations and Professions* 7: 177–200.

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." 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 searched. 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 Sociological Abstracts or Psychological 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.

Ethics Questions

1. The ethical guidelines for social research require that subjects give their "informed consent" before participating in an interview. How "informed" do you think subjects have to be?

   If you are interviewing people to learn about substance abuse and its impact on other aspects of health, is it OK just to tell respondents in advance that you are conducting a study of health issues? What if you plan to inquire about victimization experiences? Explain your reasoning.
2. Both some Homeland Security practices and inadvertent releases of web searching records have raised new concerns about the use of unobtrusive measures of behavior and attitudes. If all identifying information is removed, do you think social scientists should be able to study the extent of prostitution in different cities by analyzing police records? How about how much alcohol different types of people use by linking credit card records to store purchases?

Video Interview Questions

Listen to the researcher interview for Chapter 4 at edge.sagepub.com/chamblissmssw6e, found in the Video and Multimedia Section.

1. What problems does Dana Hunt identify with questions designed to measure frequency of substance abuse and aggressive feelings?
2. What could be done to overcome these problems?