ROYCE A. SINGLETON, JR. • BRUCE C. STRAITS

Approaches to SOCIAL RESEARCH

FIFTH EDITION

4

Elements of Research Design

All social research focuses on a particular topic and addresses specific research questions. We begin this chapter by considering the social forces and personal motives that affect topic selection. We then consider three main concerns as the researcher narrows in on researchable questions: (1) what entities (e.g., individual people, groups, formal organizations, nations) are to be studied, (2) what aspects or characteristics of these entities are of interest, and (3) what kinds of relationships among the characteristics are anticipated. By considering these key elements of research design, we can grasp the language of social research and what it means to state the problem in researchable terms.

Origins of Research Topics

The starting point for research is the selection of a topic. Once a topic is chosen and the research question is set, we can discuss rules and guidelines for conducting research that will generate the most valid data and the most definitive answers. But there are no rules for selecting a topic. Given that anything that is "social" and "empirical" could be the subject of social research, there is a nearly endless variety of potential topics. So, *how* are specific topics likely to emerge in the social sciences? We have identified five factors that explain the origin of most topics.

1. *The structure and state of the scientific discipline.* With the scientific goal of advancing knowledge, most researchers select topics suggested by the ongoing development of theory and research in their particular fields of study. The organization of disciplines casts the framework for topic selection. Social psychologists, for example, divide their discipline with respect to various forms of social behavior, such as aggression, altruism, interpersonal attraction, and conformity, which act as organizing themes or areas of research interest. Similarly, sociologists frequently study aspects of various institutions, like religion, politics, education, and the family, around which the discipline is organized. As knowledge in an area develops, inconsistencies and gaps in information are revealed that researchers attempt to resolve.

2. *Social problems.* The focus and development of the social sciences are intimately related to interest in basic problems of the "human condition." Historically, this has been a major source of research topics, especially in sociology. The most eminent sociologists of the nineteenth and early twentieth centuries—people like Emile Durkheim, Karl Marx, Max Weber, and Robert Park—concerned themselves with problems emanating from great social upheavals of their day, such as the

French and Industrial revolutions and massive foreign immigration to the United States. The problems wrought by these events—alienation, deviance, urban crowding, racism, and many others—have remained a major focus of the discipline ever since. Indeed, many people are attracted to the social sciences because of their perceived relevance to social problems.

3. *Personal values of the researcher.* To carry out a research project, with its inevitable complications, obstacles, and demands for time and money, requires considerable interest and commitment on the part of the investigator. What sustains this interest more than anything else are highly personal motivations for doing research on a particular topic. Thus, an investigator may choose a topic not only because it is considered theoretically important, novel, or researchable but also because it stimulates his or her interest. According to social psychologist Zick Rubin (1976:508–9), one reason for his embarking on the study of romantic love was that he was "by temperament and avocation, a songwriter. Songwriters traditionally put love into measures." And so he set out to find a way of measuring love scientifically. In a similar fashion, it is not surprising that members of particular groups often have pioneered research on those groups; for example, women have led the way in research on women and African Americans in research on blacks (King, Keohane, and Verba, 1994:14).

4. *Social premiums.* There are also powerful social determinants of topic selection. Through the availability of supporting funds, the prestige and popularity of the research area, and pressures within the discipline and within society, social premiums are placed on different topics at different times. Typically, these premiums will reinforce one another, with the social climate affecting funding, which in turn affects prestige. This was certainly true of the space program in the 1960s. Today, in the social sciences, the aging of the population as a whole has raised interest and support for research on the elderly, just as it has caused a dramatic increase in federal expenditures on and services available to older people. Similarly, in the 1970s the women's movement spurred a dramatic increase in research on gender issues that has continued to this day.

5. *Practical considerations.* An overriding concern in any research project is cost. Research requires time, money, and personnel. Limitations on these resources, as well as other practical considerations such as the skill of the researcher and the availability of relevant data (see chapter 12), will shape both the nature and scope of the problem that the researcher can pursue.

The choice of any given research topic may be affected by any or all of the factors mentioned. Consider, a study by Beckett Broh (2002), which examined the effects of participation in extracurricular activities on high school academic achievement. Using data from a national survey of high school students, Broh's study was designed to find out who benefits from participating in sports and other school activities and why. The study continued a line of inquiry on the impact of the extracurriculum of theoretical interest to sociologists of education and sport. Social scientists, school officials, and the general public have long debated whether sport, in particular, builds character and has positive educational benefits. Given the costs of extracurricular programming, especially school sports, and the public concern about boosting academic achievement, Broh's study also had important practical implications. Finally, the topic was of special personal interest to Broh. She herself

was a high school athlete, and after her collegiate athletic career was cut short by an injury, she turned to coaching middle and high school basketball in Michigan and Ohio. These experiences naturally sparked her interest in sport and education as a PhD student in sociology. And when her mentor suggested that she look at the questions in the National Educational Longitudinal Study, she found the means of testing some of her ideas about the impact of athletes' network of relationships on their academic achievement.

Once a general topic has been chosen, it must be stated in researchable terms. This involves translating the topic into one or more clearly defined, specific questions or hypotheses. We'll discuss the process of moving from general topics to specific questions later. First, it is important to understand that the formulation of a researchable problem or question boils down to deciding what *relationships* among what *variables* of what *units* are to be studied. We will now turn our attention to these important terms.

Units of Analysis

The entities (objects or events) under study are referred to in social research as **units of analysis.** Social scientists study a variety of units (sometimes called "elements" or "cases"). These include individual people; social roles, positions, and relationships; a wide range of social groupings such as families, organizations, and cities; as well as various social artifacts such as books, periodicals, documents, and even buildings. Ordinarily, the unit of analysis is easily identified. The unit is simply what or who is to be described or analyzed. For example, a researcher wanting to determine if larger organizations (in terms of the number of employees) have more bureaucratic rules and regulations than smaller ones would treat the *organization* as the relevant unit and would gather information on the size and bureaucratic complexity of different organizations. In Broh's study of the impact of extracurricular activities on academic achievement, the unit of analysis was *individuals* or, more precisely, individual high school students.

In much social research the individual person is the unit of analysis. Social scientists are less interested in individual differences, however, than in the impact of the social context on people in general, the social relationships that individuals form, and large-scale social processes. To analyze "the social" often requires units beyond individual human beings. Consider Fred Markowitz's (2006) study of psychiatric deinstitutionalization. Before the 1960s substantial numbers of mentally ill persons were treated in large, publicly funded hospitals. But due to developments in psychiatric drugs, stricter standards for involuntary commitment, and federal funding cuts, the number of resident patients in state and county psychiatric hospitals sharply declined over the next few decades. Studies of *individuals* seem to indicate that this deinstitutionalization movement increased the number of mentally ill who are arrested or imprisoned. For Markowitz this raised a question about social control: If the behavior of mentally ill persons, who are more likely than others to be both perpetrators and victims of crime, cannot be managed by the mental health–care system, will the mentally ill tend to end up in the criminal justice system? Mental health–care and criminal

justice *systems* refer to *social* units; social systems are aspects of communities or nations, not individuals. Markowitz thus chose to analyze cities. Examining the number of psychiatric beds in hospitals and the crime and arrest rates in 81 U.S. cities, he found that the lower a city's capacity to place the mentally ill in public psychiatric hospitals, the higher the city's crime and arrest rates.

Data on individuals also may be inadequate for investigating social and cultural change, especially in the distant past. At mid-twentieth century, sociologist David Riesman (1950) published an influential book, *The Lonely Crowd*, in which he theorized a general trend toward "other-directedness." Because of changes wrought by the Industrial Revolution, such as the expansion of white-collar and service jobs, increasing material abundance, and the development of mass media of communication, people's actions were becoming less motivated by intrinsic values and more influenced by the actions of others. But how can one study long-range trends in individual motivation when it is impossible to analyze individuals from the past? One social scientific solution is to rely on various social artifacts and to assume that such artifacts reflect the individual values and behavior of direct interest. To test Riesman's theory, for example, Sanford Dornbusch and Lauren Hickman (1959) chose as their units of analysis advertisements in a mass-circulation women's magazine for the period 1890–1956, to see if the advertising had appealed increasingly over time to the standards of others (other-directedness). Indeed, it had.

Aggregate Data

One way of identifying the unit of analysis in reported research is to determine the properties or traits that are being described. Individual people have traits such as age, race, and gender; have attitudes and opinions; and act in certain ways, such as saying that they voted in the last civic election, committed a felony, or did volunteer work last week. Only social aggregates such as groups, organizations, cities, and nations have a size, average age, percentage who voted, or crime rate. It can be confusing, however, when individual-level data are combined to describe a social unit to which the individuals belong.

Information about one set of units that is statistically combined to describe a larger social unit is called **aggregate data**. Each year, *U.S. News and World Report* uses aggregate data to rank America's best colleges: They gather information about individual students at each college to characterize the college as a whole. Among their indicators of academic quality, for example, is each school's graduation rate, which is calculated by dividing the number of enrollees in a given class into the number who graduate within 6 years of their enrollment. Other aggregate measures include average SAT or ACT test score; the proportion of enrolled freshmen who graduated in the top 10 percent of their high school classes; and the acceptance rate, the ratio of the number of students admitted to the number who apply.

In *U.S. News'* rankings, the unit of analysis is colleges; that is what is being described and compared. But the use of aggregate data does not always imply a collective unit of analysis. In another study of adolescents' extracurricular participation, Andrew Guest and Barbara Schneider (2003) wondered whether the effect of such activities depended on the social context of the school and community.

For example, sports may have a greater effect on academic achievement in lower-class communities, which value sports as paths to financial gain, than in upper-class communities, which value sports for their health and aesthetic benefits. Their unit of analysis, like Broh's study, was individual high school students; however, they created aggregate measures to study the impact of the community and school context. Their measure of the community's socioeconomic class was based largely on the average income and education of the neighborhoods where the students lived, and a school context measure consisted of the percentage of students from each student's high school who went on to 4-year colleges after graduation.

Thus, when information about individuals is aggregated to describe groups or collectivities, the unit of analysis may be either the individual or the group. How can you tell? In their analyses, researchers ordinarily compare a number of instances of a particular unit—for example, a number of individuals, a number of cities, or a number of colleges. To identify the unit of analysis, therefore, ask yourself what is being described in each instance and what sorts of units are being compared—individuals or groups. If the aggregate data are used to compare different groups, the unit of analysis is the group. If the data are used to compare individuals who belong to different groups, the unit is the individual.

One reason it is important to identify accurately the unit of analysis is that confusion over units may result in false conclusions about research findings. Generally speaking, assertions should be made only about the particular unit under study. (Actually, assertions should be even more circumscribed than this, as we will see in chapter 6.) To draw conclusions about one unit on the basis of information about another is to risk committing a logical fallacy.

Ecological Fallacy

The most common fallacy involving the mismatching of units is the **ecological fallacy** (Robinson, 1950). This occurs when relationships between properties of groups or geographic areas are used to make inferences about the individual behaviors of the people within those groups or areas. This is quite similar to what logicians call the "fallacy of division": assuming that what holds true of a group also is true of individuals within the group. Knowing that Sally attended a college whose students had relatively low average SAT scores, you would commit this fallacy if you assumed that Sally herself had low SAT scores.

Political analysts who use aggregate data from elections to study individual voting behavior are particularly susceptible to the ecological fallacy. Suppose, for example, that a researcher wanted to know whether registered Democrats or Republicans were more likely to support an Independent Party candidate in a city election but that the only information available was the percentage of votes the candidate received and the percentages of Republican and Democratic voters *in each precinct*. In short, the researcher wants to draw conclusions about *individual voters* but has collective information only about *precincts*. Knowing that the candidate received a relatively larger number of votes in precincts with greater percentages of Republicans does not permit the conclusion that Republican voters were more likely to support the candidate than were Democratic voters. In fact, it is quite plausible that Democrats in predominantly

Republican precincts were more likely to support the Independent Party candidate than were Democrats in other precincts. Because the unit of analysis here is the precinct and not the individual voter, we simply do not know how individual voters within each precinct voted.

At one time social scientists frequently performed ecological analyses such as the one above. For example, criminologists analyzed crime and delinquency rates in relation to other characteristics of census tracts in order to draw conclusions about characteristics of individual criminals and delinquents. A typical erroneous conclusion might be that foreign-born persons commit more crimes than native-born persons because the crime rate is higher in areas with greater proportions of foreigners. But such a conclusion is clearly unwarranted because we do not know who actually committed the crimes—foreign or native-born persons. Similarly, Durkheim's classic study of suicide was subject to the ecological fallacy by inferring that Protestants commit more suicides than Catholics from the observation that suicide rates were higher in predominantly Protestant nations than in predominantly Catholic ones.[1]

It is not always wrong to draw conclusions about individual-level processes from aggregate or group-level data. Social scientists have identified conditions under which it is reasonable to make such inferences (e.g., Firebaugh, 1978), but it is often difficult to determine if these conditions are met. The implications of the ecological fallacy are clear: Carefully determine the units about which you wish to draw conclusions and then make sure that your data pertain to those units. If you are interested in individuals but only aggregate data are available, then draw conclusions very tentatively, recognizing the possibility of an ecological fallacy.

KEY POINT

If data describe social units (e.g., schools), one must be cautious in drawing conclusions about individuals (students within the schools).

Variables

While the researcher observes units of analysis, it is relationships among characteristics of units that are of primary interest. Characteristics of units that vary, taking on different values, categories, or attributes for different observations, are called **variables**. Variables may vary over cases, over time, or over both cases and time. For example, among individuals, any set of characteristics that may differ for different people, such as age (range of years), gender (male and female), and marital status (single, married, divorced, widowed, etc.), is a variable. And for an individual, any characteristic that may vary from one time period to the next, such as age, level of education (first grade, second grade, etc.), and income (dollars earned per year), is a variable.

It is not unusual to see some confusion between variables and the attributes or categories of which they consist. "Gender" is a variable consisting of the categories male and female; "male" and "female" by themselves are not variables but simply

categories that distinguish persons of different gender. Likewise, "divorced" and "Republican" are not variables but categories of the variables "marital status" and "political party affiliation," respectively. To keep this distinction clear, note that any term you would use to describe yourself or someone else (e.g., sophomore, sociology major) is an attribute or category of a variable (academic class, major).

You are now in a position to test your understanding of the concepts of unit of analysis and variable. Table 4.1 is designed to help you do this. The first column presents eight research questions or hypotheses; the second and third columns identify the relevant units of analysis and variables. Read the first column while covering up the second and third columns, and record the unit of analysis and variables on a separate sheet of paper. Then check your answers by examining the full table.

KEY POINT

Social scientists study *relationships* among *variables;* variables depict and differentiate *units of analysis.*

Types of Variables

Social scientists find it necessary to classify variables in several ways. One type of classification is necessitated by the complexity of social phenomena. For any given research problem, the researcher can observe and measure only a few of the many potentially relevant properties. Those variables that are the object of study—part of some specified relationship—are called **explanatory variables**, and all other variables are **extraneous** (Kish, 1959).

There are two types of explanatory variables: dependent and independent. The **dependent variable** is the one the researcher is interested in explaining and predicting. Variation in the dependent variable is thought to depend on or to be influenced by certain other variables. The explanatory variables that do the influencing and explaining are called **independent**. If we think in terms of cause and effect, the independent variable is the presumed cause and the dependent variable the presumed effect. Independent variables are also called "predictor variables" because their values or categories may be used to predict the values or categories of dependent variables. For example, when Broh studied the impact of extracurricular involvement on academic achievement, her independent variable consisted of whether students participated in specific school activities such as interscholastic sports and her dependent variable was level of academic achievement. One research question was whether sport participation explained (or predicted) differences in academic achievement.

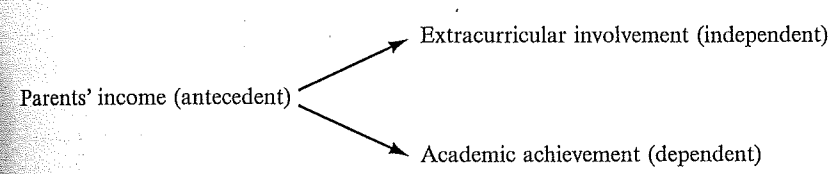
Research studies in the social sciences often involve several independent variables and sometimes more than one dependent variable. Also, a variable is not intrinsically independent or dependent. An independent variable in one study may well be a dependent variable in another, depending on what the researcher is trying to explain. Finally, it is conventional in mathematics and science for the letter X to symbolize the independent variable and for the letter Y to represent the dependent variable. This is a practice we shall follow in the remainder of the book.

TABLE 4.1. Sample Research Questions, Units of Analysis, and Variables

Research question/hypothesis [WHAT ONE WANTS TO KNOW]	*Unit of analysis* [WHAT ENTITIES ARE DESCRIBED AND COMPARED]	*Variables* [WITH RESPECT TO WHAT CHARACTERISTICS]
Are older people more afraid of crime than younger people?	Individuals	Age, fear of crime
The greater the growth of air passenger traffic at a city's airport, the greater the economic growth.	Cities	Growth of air traffic, economic growth
The higher the proportion of female employees, the lower the wages in nineteenth-century factories.	Factories	Proportion of employees who are female, average wage
Does economic development lower the birth rate?	Nations	Level of economic development, birth rate
The longer the engagement period, the longer the marriage.	Couples (dyads)	Length of engagement, marriage duration
Fan support in the NBA, as measured by attendance, is not related to the proportion of black players on the team.	NBA teams	Racial composition of the team, average attendance
Comic strips introduced in the 1930s were more likely to emphasize powerful heroes than strips introduced in the 1920s.	Comic strips	Whether main characters in strip were powerful, when comic strips were introduced (1920s or 1930s)
When television sports commentators make reference to individual competitors, they are more likely to use first names for female athletes than for male athletes.	References to competitors	Gender of athlete, whether athlete identified by first name

Extraneous variables, which are not part of the explanatory set, may be classified in two important ways. First, in relation to specific independent and dependent variables, they may be **antecedent** or **intervening**. An antecedent variable occurs prior in time to both the independent and dependent variable; a variable is intervening if it is an effect of the independent variable and a cause of the dependent variable. Antecedent variables in Broh's study were parents' income and a student's race and gender; each of these variables may affect both extracurricular involvement and academic achievement. An intervening variable was students' self-esteem. Extracurricular involvement could affect self-esteem, which in turn could affect a student's academic performance. Figure 4.1 depicts these examples of antecedent and intervening variables. Each arrow in the figure represents causal direction. Thus "Parents' income → Extracurricular involvement" means that parents' income influences or causes extracurricular involvement, and the absence of an arrow means that one variable does not cause another.

A. Antecedent Variable



B. Intervening Variable

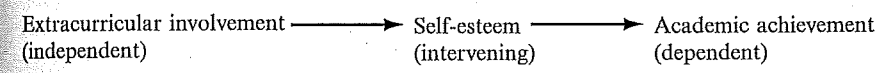


FIGURE 4.1. Antecedent and intervening variables.

Extraneous variables also may be categorized as controlled or uncontrolled. Controlled or, more commonly, **control variables** are held constant, or prevented from varying, during the course of observation or analysis. This may be done to limit the focus of the research or to test hypotheses pertaining to specific subgroups—for example, all males or all males under 18 years of age. Basically, the value or category of a control variable remains the same for a given set of observations. Several techniques for holding variables constant are discussed at length in the following chapters. Some examples would be selecting only individuals of the same age and gender, observing groups of the same size, creating uniform laboratory conditions or social settings in which to observe people or groups, and statistically controlling for specific attributes.

Whenever a variable is held constant in research, that variable cannot account for (or explain) any of the variation in the explanatory variables. Suppose, for example, that you wanted to explain differences (i.e., variation) between people in their level of aggression. If you controlled for gender by studying only males, then the variable "gender" could not account for any of the observed variation in aggression. Holding variables constant thus simplifies complex social situations. It is a means of ruling out variables that are not of immediate interest but that might otherwise explain part of the phenomenon that the investigator wishes to understand. One aim of efficient research design is to identify potentially relevant extraneous variables in order to measure and control as many as is feasible (Kish, 1959). Broh measured numerous extraneous variables, including parents' income and students' race, gender, and self-esteem. This enabled her to statistically control and test the effects of these variables on the relationship between extracurricular activities and academic achievement.

Another important distinction is made between quantitative and qualitative variables. Actually, this is a rather crude way of pointing to some significant differences in the numerical representation of variable categories. We make more precise distinctions along these lines in the next chapter but for now appeal to the common understanding of quantitative and qualitative. A variable is **quantitative** if its values or categories consist of numbers and if differences between its categories can be ex-

pressed numerically. For example, the variable "income," measured in dollars, signifies a quantitative difference—a certain number of dollars—between people with different incomes. **Qualitative variables** have discrete categories, usually designated by words or labels, and nonnumerical differences between categories. For example, the variable "gender," consisting of the discrete categories "male" and "female," makes a categorical, but not a numerical, distinction between people.[2]

Relationships

Social scientists' ultimate objective is to make sense of the social world by discovering enduring relationships among phenomena. Much research is therefore aimed directly at developing and testing relationships. There are other, more immediate purposes for which research is conducted, such as exploration and description, which we discuss later in the chapter. Yet, even when the researcher's goal is discovery or description, research findings will depend to a large extent on what particular *relationships* are anticipated. Research is not like the kind of fishing trip in which you drop your line anywhere, hoping you will catch something. Investigators do not make random observations. Whatever their goals, they must decide what to observe or ignore, how to go about making their observations, and how to interpret them. Such decisions inevitably are based on the researcher's expectations about how variables are related to one another.

Anticipated relationships are implicit in the researcher's disciplinary focus. If a sociologist looks at dormitory rooms and classrooms, for example, "it is not with the eyes of an architect" but with an interest in how these rooms and their contents are related to the people using them (Miles and Huberman, 1994:17). Similarly, you cannot do any kind of research without some guiding orientation. Consider the investigator who wants to study "everything" about families but "let the facts speak for themselves" (Batten, 1971:9–11). Without *any* expectations, how is the researcher to decide what constitutes a "family"? Should first cousins be included even if they live far away or seldom keep in touch? Should a family include all persons living in the same household, no matter how remote the blood tie (Batten, 1971)? When collecting data, should the researcher include the hat size and shoe size of each family member? Such considerations ultimately depend on the social researcher's *expectations* about salient properties and *relationships* regarding families.

Facts never speak for themselves. Not only which facts are sought but also how they are interpreted depends again on anticipated relationships—on what particular answers the researcher expects the "facts" to provide. A trenchant example is related by sociologist Herman Smith (1975:22). According to Smith, it is a fact that African Americans save less money than whites. However, if the fact that blacks earn less than whites is taken into account, it turns out that blacks save proportionately more of their incomes than whites. Thus, blacks would appear to be *more* frugal than whites. But this conclusion would not have been reached if the researcher had not entertained the notion that the relationship between race and savings might be affected by a more complex set of relationships involving how much of his or her income a person saved.

It is clear, then, that all research carries expectations about the nature of what is being investigated. As a result, an important tenet of social research, which unfortunately is not always followed, is that one's expectations—that is, anticipated relationships and guiding theoretical explanations—should always be identified as far as possible. Beckett Broh's (2002) selection of variables for her study was guided largely by three theoretical explanations of the link between sports participation and educational achievement: (1) developmental theory—sports teaches skills, such as a strong work ethic and perseverance that help students achieve in the classroom; (2) the leading crowd hypothesis—athletes gain social status and membership in a peer group of disproportionately college-oriented high achievers, which facilitates higher academic performance; and (3) social capital theory—sports create strong social ties among students, parents, and teachers that provide educational benefits. Broh derived the social capital model in part from her experience as a basketball coach; she noticed that her players tended to spend more time with their parents and had better relationships with them than other students.

KEY POINT

Facts or data do not speak for themselves; they must be interpreted based on anticipated and predicted relationships.

Thus far we have relied on the reader's intuitive grasp of the term "relationship." Everybody has experienced relationships at some time in their lives. One might think of relationships among kin; "serious" relationships, as between lovers; and relationships among the members of teams and work groups. We also have a sense of a relationship when one event regularly precedes or follows the occurrence of another, for example, when the appearance of dark clouds is regularly followed by rain. All such relationships have two features. First, they always involve two or more entities: persons, objects, or events—such as parent and child, leader and follower, or clouds and rain. Second, the pairs or combinations of things usually occur together and change together; thus, the appearance of one thing signals the appearance of the other and the absence of one implies the absence of the other. For example, by definition every parent has a child and every child has (or had) a parent; we cannot have one without the other. Also, by observation we know that certain kinds of clouds produce rain and that without clouds it cannot rain.

The kind of relationships with which social scientists are concerned, relationships among variables, have these same two features. Two or more variables are related, or form a relationship, to the extent that changes in one are accompanied by systematic changes in the other(s). Since the manner in which the variables change or vary together will depend on whether the variables are qualitative or quantitative, we will consider the nature of relationships separately for each of these types of variables.

Relationships among Qualitative Variables

Consider the two qualitative variables race and political party affiliation. If two individuals have the same party affiliation, say Democrat, then the category of this variable does not change as we look from one individual to the other. If they have

different affiliations, say one is a Democrat and the other is a Republican, then the category of the variable does change. Assuming similar statements about race, we would say that a relationship exists between race and political party affiliation if a comparison of a pair of individuals reveals that a change in race is accompanied by a change in party affiliation. More generally, the basic idea of a relationship, or association, between qualitative variables can be incorporated into two assertions (Leik, 1972:26): (1) if one variable changes, the other variable changes and (2) if one variable does not change, the other does not change. If these two assertions were true for all pairs of cases, the result would be a perfect one-to-one correspondence between the categories of one variable and the categories of the other. Table 4.2A depicts a perfect relationship between race and political party affiliation among twenty individuals.

In actual research, of course, we never see such perfect associations. To the researcher, therefore, the important question is not whether a given pair of variables is perfectly associated but how strongly they are related. That is, how closely do the data approximate a perfect association between variables? Statistical techniques to assess this are called "measures of association." These techniques, as well as the concept of "strength of relationship," are best understood if assertions about concomitant changes in variables are treated as predictions. Table 4.2A suggests two predictions: (1) if a person is black, then he or she will be a Democrat and (2) if a person is white, then he or she will be a Republican. The proportion of times such predictions are correct for all cases is an index of the strength of the relationship. A high proportion indicates that the variables are strongly related; a low proportion indicates a weak relationship. Furthermore, if the proportion is so low that knowledge of one variable is of no use in predicting the other, the variables are said to be unrelated. The rest of Table 4.2 gives a possible combination showing a moderate

TABLE 4.2. Varying Degrees of Association between Two Qualitative Variables: Race and Political Affiliation

A. Perfect association	*White*	*Black*	*Total*
Democrat	0	10	10
Republican	10	0	10
Total	10	10	20
B. Moderate association	*White*	*Black*	*Total*
Democrat	3	7	10
Republican	7	3	10
Total	10	10	20
C. No association	*White*	*Black*	*Total*
Democrat	5	5	10
Republican	5	5	10
Total	10	10	20

relationship between race and political party affiliation (B) and a distribution indicating no relationship (C). Statistical indices of association may be computed for each of these distributions. Ordinarily, these indices will range from 0, indicating no relationship, to 1.00, indicating a perfect relationship.

Relationships among Quantitative Variables

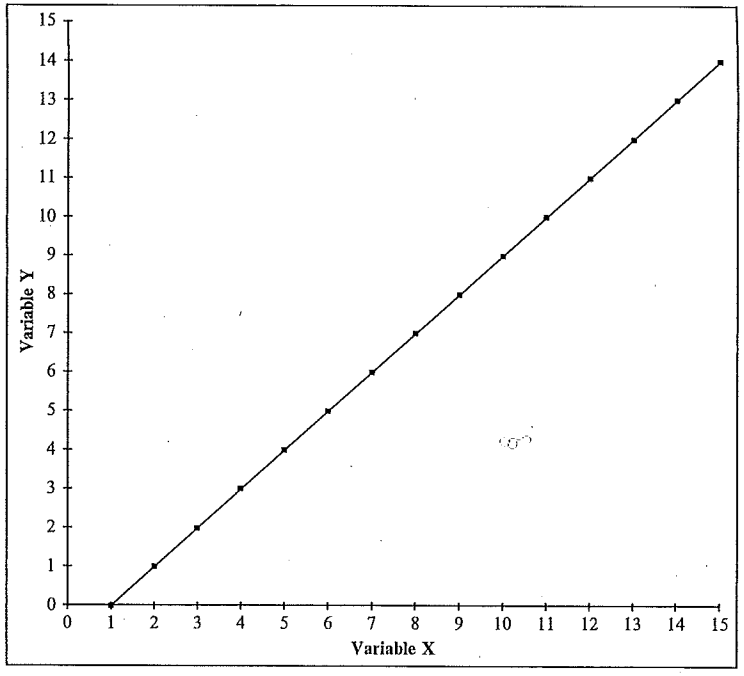
As we move from qualitative to quantitative variables, it becomes possible to say whether a change in a variable represents an increase or a decrease in value. With this additional property, we can measure not only the strength of the relationship but also two other aspects: direction and linearity.

A relationship may be either positive or negative in direction. A **positive** or **direct relationship** between variables exists if an increase in the value of one variable is accompanied by an increase in the value of the other or if a decrease in the value of one variable is accompanied by a decrease in the value of the other. In other words, the two variables consistently change in the same direction. We would expect a positive relationship between sons' heights and fathers' heights (the taller a father, the taller his son will tend to be) and between students' scores on the Scholastic Aptitude Test and their college grade-point averages (as scores increase, grades tend to increase). A **negative** or **inverse relationship** between variables exists if a decrease in the value of one variable is accompanied by an increase in the value of the other. Thus, changes in one variable are opposite in direction to changes in the other. We would find a negative relationship between a person's age and how long he or she is expected to live (as age increases, life expectancy decreases) and between the speed and accuracy with which people perform many tasks (the faster one does something, the less accurately one is likely to do it).

Relationships among quantitative variables are usually depicted with graphs, such as those in Figure 4.2. The lines in these two graphs illustrate the characteristic of "linearity." The straight line in Figure 4.2A depicts a *linear* relationship. (Actually, this line depicts a *positive linear* relationship. A straight line running from the upper left corner to the lower right corner would depict a *negative linear* relationship.) Notice that one variable changes at the same rate and in the same direction (positive) over the entire range of the other variable. The curved line in Figure 4.2B depicts a *curvilinear* relationship. In this case, the rate of change in one variable is *not* consistent over all values of the other: Variable Y increases more rapidly for low values than for high values of variable X and then reverses direction. We would expect this pattern of relationship to occur between the age and annual earnings of adult workers. Earnings will generally increase with age up to retirement and then will decline.

Tabular and graphic representations of relationships between variables and other statistical analyses are discussed further in chapter 15. There as well as elsewhere in the book we focus on linear relationships since this pattern is most often assumed and analyzed in the social sciences. A common statistical measure of the strength and direction of linear relationships between two quantitative variables is the Pearson product-moment coefficient of correlation, or **correlation coefficient** for short. Symbolized by the letter r, the correlation coefficient may vary between -1.00 and $+1.00$. The signs $-$ and $+$ indicate the direction of the relationship, pos-

A. Linear relationship



B. Curvilinear relationship

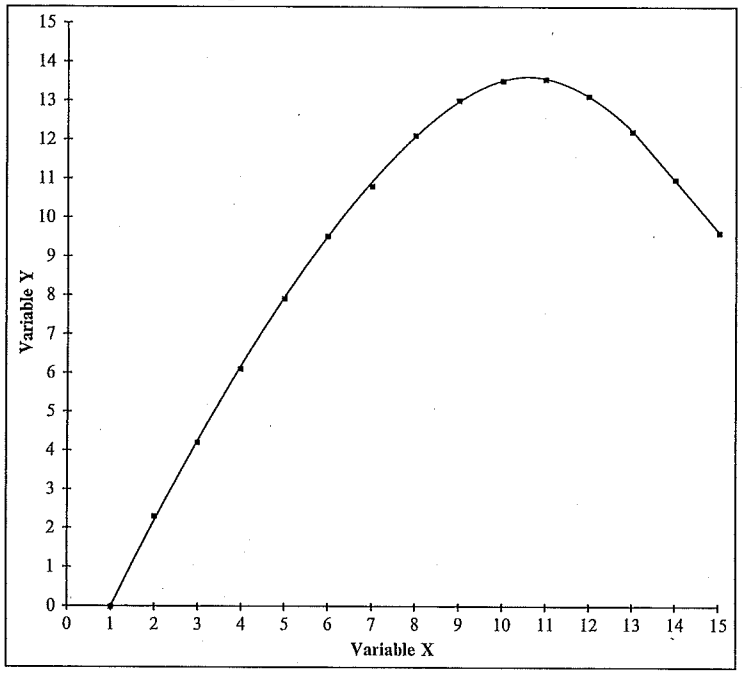


FIGURE 4.2. Linear and curvilinear relationships between two quantitative variables.

itive or negative; the magnitude of the coefficient, ignoring the sign, indicates the strength of association. Again, strength of association can be construed as accuracy of prediction. Relationships between two quantitative variables are represented in graphs such as Figure 4.2; however, when each case is plotted according to its values on the independent and dependent variables, the plots seldom form a perfectly straight line, as in Figure 4.2A. Rather, the plots cluster to varying degrees about the best-fitting line. If we use this line to predict Y values from X values, then the more closely the plots cluster about the line, the more accurate the prediction and the stronger the association. Figure 4.3 shows two possible outcomes. In Figure 4.3A, where the plots cluster fairly tightly around the line, there is a relatively strong correlation of .74 between variable X and variable Y. This is the same correlation that Broh found between students' grades in English and grades in math. In Figure 4.3B, where the plots are very loosely clustered about the line, the correlation is .08. This is the correlation Broh found between number of hours spent on homework and grades in English.

Relationships between a Qualitative and a Quantitative Variable

Another way of assessing relationships is used when both qualitative and quantitative variables are involved. Most often in such cases, especially in experiments, the independent variable is qualitative and the dependent variable is quantitative. This is the type of relationship considered here.

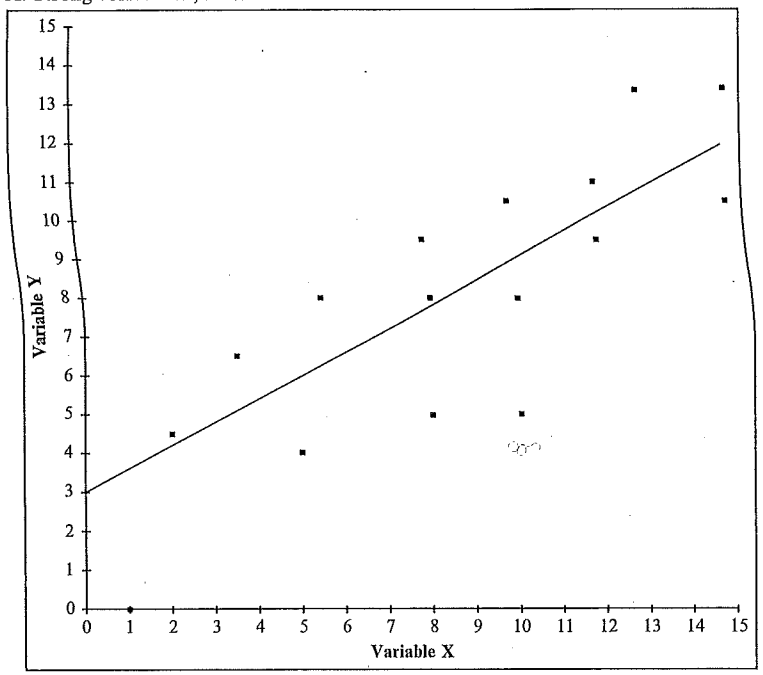
It will be helpful, once again, to think in terms of predictions. A relationship is said to exist if the different categories of the independent variable predict different values for the dependent variable. Thus, if each category of the independent variable is treated as a distinct group, then a relationship can be described in terms of the *differences among groups* of the dependent variable. We might compute an average value of the dependent variable for each group. No differences in these averages across groups would then indicate no relationship. And, in general, the larger the differences, the stronger the relationship.

For example, suppose a researcher were interested in examining the relationship between race and annual income. Since annual income, as measured in dollars earned, is a quantitative variable, average incomes could be computed for each racial group (i.e., for each category of the qualitative variable race), as shown in Table 4.3.[3] Note in the table that there is a difference in the averages across groups: In 2005, whites earned $50,784 on the average and blacks earned $30,858. Thus, we may conclude that a relationship exists between race and income.[4]

Statistically Significant Relationships

Social scientists do not ordinarily consider an association between two variables meaningful unless the association is **statistically significant**. Consider the moderate association between race and party affiliation in Table 4.2B. Note that there are only twenty cases in this table. Now suppose that these twenty cases represent a random sample of the U.S. population. It would not be appropriate to conclude from these data that there is an association between race and party affiliation in the gen-

A. Strong correlation, r = .74



B. Weak correlation, r = .08

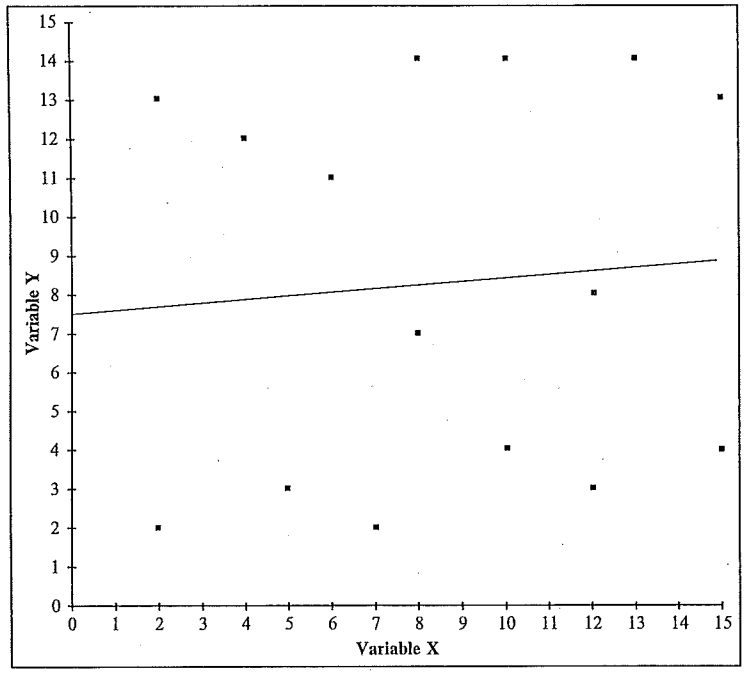


FIGURE 4.3. Strong and weak correlations between two quantitative vaiables.

TABLE 4.3. Relationship between a Quantitative Variable (Median Income in 2005 Dollars) and a Qualitative Variable (Race), 2005

	White	*Black*
Median income	$50,784	$30,858

Source: DeNavas-Walt, C., B. D. Proctor, and C. H. Lee. *Income, Poverty, and Health Insurance Coverage in the United States: 2005.* Current Population Reports, P60-231. Washington, D.C.: U.S. Government Printing Office, table 1.

TABLE 4.4. Race by Political Party Preference,* 2006 General Social Survey

	White	*Black*
Democrat	42.5%	89.3%
Republican	57.5	10.7
	100.0%	100.0%
(*N*)	(1863)	(237)

*Party preference indicates respondents who described themselves as either "strong" or "not very strong" Democrats (Republicans). Based on 3,803 white and black respondents.

eral population. Why? Because in a sample this small the difference in party affiliation between blacks and whites easily could have occurred by chance (i.e., as a result of random selection) even if there is no relationship in the larger population.

Statistically significant associations are those that are *not* likely to have occurred by chance or random processes. Table 4.4 shows the relationship between race and party affiliation in a recent national sample of the adult population. The respondents in this survey were asked if they usually thought of themselves as a Democrat, Republican, or Independent. The table shows only those respondents who described themselves as either Democrats or Republicans. Blacks were far more likely than whites to identify themselves as Democrats (89.3 percent compared to 42.5 percent); given the size of the sample (3803), this difference is highly unlikely to have occurred by chance and therefore is statistically significant. When this property of relationships is reported in the research literature, it is indicated with an italic lowercase p followed by $<$ and a decimal; for example, $p < .05$. This means that the probability is less than 5 in 100 that the relationship could occur by chance assuming that there is no relationship in the larger population from which the sample is drawn. With odds this low, we would conclude that the association did not occur by chance and therefore that a relationship between race and party affiliation exists in the United States. (See chapter 15 for a brief discussion of tests of statistical significance.)

It is important to note that statistical significance is not the same as strength of association. Although strong associations are more likely to be statistically significant than weak ones, a very weak association may be significant provided that the sample is large enough. For example, both of the correlations reported above for the Broh study were statistically significant even though one (.74) was strong and the other (.08) was very weak. The reason that such a low correlation was significant was that it was based on a sample size of over 10,000 students. With samples this large, even relatively weak associations are unlikely to occur by chance and, therefore, likely to differ from a zero association in the population (all high school students) from which the sample was drawn.

KEY POINT

Relationships may be described by their *form* (how changes in one variable vary with changes in the other), *strength* (how accurately values of one variable predict values of the other), and *statistical significance* (likelihood that the relationship occurred by chance or random processes).

So far we have examined some common ways of depicting relationships. Also, we have noted four properties of relationships: strength, directionality, linearity, and statistical significance. The *strength* of a relationship refers to the extent to which variables are associated or correlated. *Directionality* and *linearity* tell us how changes in one variable are related to changes in another. Do the variables change in the same or the opposite direction? Is the rate of change in one variable consistent over all values of the other? Together these two properties describe the *form* of the relationship. Knowing the strength and form of relationships as well as whether it is statistically significant will often satisfy the researcher's curiosity. However, if our interest is in explaining social phenomena, we also will need to know about the causal link between variables. And these statistical properties are never sufficient to establish a cause-and-effect relationship.

The Nature of Causal Relationships

In chapter 2 we noted that, for purposes of explanation and prediction, scientists find it helpful to think in terms of cause-and-effect relationships. But how does one identify such a relationship? What is meant by the term "causality"? At first glance, the task of defining "causality" seems simple. Since we are so accustomed to thinking causally and causal terms are so frequently used in everyday life, the concept would appear to be widely understood. In lay terms, a *cause* is something that makes something happen or change. It seems obvious that a rock thrown against a window will cause the glass to shatter. And the fact that drinking too much soda causes me to get a stomachache is a **causal relationship** that you can comprehend even if you fortunately have not had the same experience.

In contrast to the implicit understanding of causality that seems to exist in everyday life, the meaning of the concept of "cause" has been hotly debated by philosophers and scientists for centuries. Much of this debate stems from the philosopher David Hume's analysis (1748). Hume argued that all that one can observe is a constant or stable association between events. From such association we infer a causal connection; however, there is no way of logically or empirically showing that a causal connection actually exists. A causal relationship exists only in the observer's

mind; it is something inferred from an observed association between events. Following Hume's line of reasoning, some philosophers and some scientists (e.g., Kerlinger, 1973) have maintained that the concept of "cause" ought to be discarded from scientific work. Yet many people regard causal relationships as the heart of scientific understanding. Furthermore, even if such relationships cannot be "proven" empirically (just as no generalization can be proven by scientific evidence), researchers have found it helpful to *think* causally (Blalock, 1964:6) and have found working with causal hypotheses to be a very productive way of doing science.

The important point of Hume's analysis, therefore, is that we should understand the bases for making causal inferences. In other words, what kind of evidence supports the belief that a causal relationship exists? Social scientists generally require at least three kinds of evidence to establish causality. These requisites are association, direction of influence, and nonspuriousness.

Association

For one variable to be a cause of another, the variables must be statistically associated. If the pattern of changes in one variable is not related to changes in another, then the one cannot be producing, or causing, changes in the other. Thus, for instance, if intelligence is unrelated to delinquency—that is, if adolescents of high and low intelligence are equally likely to be delinquent—then intelligence cannot be a cause of delinquency.

Associations, of course, are almost never perfect; so a perfect association between variables is not a criterion of causality. According to logicians, in fact, the very idea of causation implies imperfect associations between events. Causes can have invariable effects only in "closed systems" that exclude all other factors that might influence the relationship under investigation. Many of the laws of physics, for instance, are said to apply exactly only in a vacuum. However, vacuums are not found in nature; neither is it possible in real social situations to eliminate completely the influence of extraneous factors. Perfect associations may be expected, therefore, only under the theoretical condition that "all other things are equal" but not in the "real world" of observations.

Barring "perfect" associations, then, the application of this first criterion necessarily involves a judgment about whether an association implies a meaningful causal relationship. In the social sciences, causal relationships often are implied from comparatively "weak" associations. One reason for this is that many measurements in the social sciences are relatively imprecise. The primary reason, though, is that in explaining human action, multiple causes may independently or jointly produce the same or similar effects. A weak association may mean that only one of several causes has been identified, or it may mean that a causal relationship exists but not under the conditions or for the particular segment of the population in which the weak association was observed. Rather than strength of association, therefore, social scientists rely on tests of statistical significance to determine whether a meaningful, interpretable association exists between variables. Although Broh found a weak correlation between time spent on homework and grades, because this association was statistically significant, it was interpreted as a cause (among others) of academic performance.

Direction of influence

A second criterion needed to establish causality is that a cause must precede its effect, or at least the direction of influence should be from cause to effect. In other words, changes in the causal factor, or independent variable, must influence changes in the effect, or dependent variable, but not vice versa. For many relationships in social research the direction of influence between variables can be conceived in only one way. For example, characteristics fixed at birth, such as a person's race and gender, come before characteristics developed later in life, such as a person's education or political party affiliation; and it is hard to imagine how changes in the latter could influence changes in the former.

Direction of influence is not always so easy to determine. Suppose you found a correlation between racial prejudice and interracial contact showing that the more contact a person has with members of other races, the less prejudiced he or she is apt to be. One possible interpretation is that racial contact increases familiarity and contradicts stereotypes, thereby reducing prejudice. An equally plausible interpretation is that prejudiced people will avoid contact while tolerant people will readily interact with other races, so that racial prejudice influences racial contact. Without any information about the direction of influence between these variables, there is no basis for deciding which variable is the cause and which is the effect. To take another example, a correlation between grades and class attendance may mean that greater attendance "increases the amount learned and thus causes higher grades" or it may mean that "good grades lead students who obtain them to attend class more frequently" (Neale and Liebert, 1986:89–90).

Direction of influence was an issue in the Broh study. An association between sports participation and academic performance could mean that playing sports has educational benefits, but it also could mean that higher-achieving, "good" students are more likely to choose or be selected to play sports than other students. Because Broh hypothesized that sports participation had a positive influence on grades, it was important for her to rule out the possibility that superior academic performance leads to sports participation.

The requirement that causes influence effects has two major implications for research: (1) hypothesized relationships should always specify the direction of influence among variables and (2) whenever the direction cannot be established theoretically, it should be tested empirically. As you will see, this task is relatively easy in experiments but often difficult in other kinds of research.

Nonspuriousness (elimination of rival hypotheses)

If two variables happen to be related to a common extraneous variable, then a statistical association can exist even if there is no inherent link between the variables. Therefore, to infer a causal relationship from an observed correlation, there should be good reason to believe that there are no "hidden" factors that could have created an accidental or spurious relationship between the variables. When an association or correlation between variables cannot be explained by an extraneous variable, the relationship is said to be *nonspurious*. When a correlation has been produced by an

extraneous third factor and neither of the variables involved in the correlation has influenced the other, it is called a **spurious relationship**.

The idea of spuriousness is obvious when we consider two popular examples in the social sciences. The first is a reported positive correlation in Europe between the number of storks in an area and the number of births in that area (Wallis and Roberts, 1956:79). This correlation might explain how the legend that storks bring babies got started, but it hardly warrants a causal inference. Rather, the correlation is produced by the size of the population. Storks like to nest in the crannies and chimneys of buildings; so as the population and thus the number of buildings increases, the number of places for storks to nest increases. And as the population increases, so does the number of babies. We also would expect to find a positive correlation between the number of firefighters at a fire and the amount of damage done. But this does not imply that firefighters did the damage. The reason for the correlation is the size of the fire: Bigger fires cause more firefighters to be summoned and cause more damage.

In these examples the original relationship is an incidental consequence of a common cause: an antecedent extraneous variable. In the first example, the size of the population accounts for *both* the number of storks and the number of births; in the second, the severity of the fire determines *both* the number of firefighters and the amount of damage. These relationships are depicted in Figure 4.4. The examples are intuitively obvious, and the third factor is fairly easy to identify.

In actual research, spurious relationships are much less apparent, and the possibility often exists that an unknown variable may have produced an observed association. For many years, numerous studies have shown that children who are breast-fed tend to have higher IQ scores than those who are not. Proponents of breast-feeding (which does have many other benefits for mother and child) have inferred a causal connection, contending that the effect may be due to a component of breast milk or perhaps to the physical interaction between mother and baby. Recent research suggests, however, that this association is spurious. Both breast-feeding and child intelligence are influenced by the mother's intelligence: Mothers with

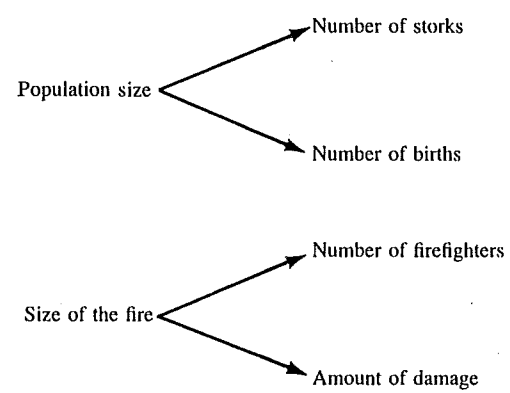


FIGURE 4.4. Examples of spurious relationships.

high IQs are more likely to breast-feed *and* more likely to have children with high IQs (Der, Batty, and Deary, 2006).

To infer that a relationship is nonspurious, researchers must identify and control for extraneous variables that might account for an association. We can see how this works by considering the relationship between cigarette smoking and lung cancer. According to the Centers for Disease Control and Prevention (2008), people who smoke are ten to twenty times more likely to develop lung cancer than nonsmokers. But the reason scientists now believe that there is a causal link between smoking and cancer is not based simply on this association; rather, it is based on the fact that smokers' greater risk of getting cancer remains about the same when the effects of other variables are removed.

Suppose we control for the variable urban–rural residence. This could create a spurious association between smoking and cancer if urban areas have more smokers as well as sources of lung cancer (e.g., greater air pollution) than rural areas. To control for extraneous variables we must remove their effects, such as by computing the incidence of lung cancer among smokers and nonsmokers separately for urban and rural residents. A finding of no difference in the cancer rate between smoking and nonsmoking urban dwellers and between smoking and nonsmoking rural residents would suggest that urban–rural residence produced a spurious association between smoking and cancer. If we found, however, that both urban and rural smokers were more likely to develop lung cancer than nonsmokers in these areas, we would be more confident that the relationship is nonspurious.

KEY POINT

An association between variables does not necessarily imply that the variables are causally related. In addition to association, causal inferences require evidence of direction of influence and nonspuriousness.

To infer nonspuriousness, researchers must show that the relationship is maintained when all extraneous variables are held constant. Circumstances seldom allow for the control of all variables; therefore, researchers attempt to control the effects of as many as possible. The greater the number of variables that are controlled without altering a relationship, the greater the likelihood that the relationship is not spurious. Thus, we would become even more confident in a causal link between smoking and lung cancer if the incidence of lung cancer among smokers and nonsmokers remains the same when examined separately for men and women, people of low and high socioeconomic status, people of different races, and so forth. (Box 4.1 provides another example of the importance of the nonspuriousness criterion. In this case, the causal interpretation that exercise reduces the risk of heart attacks was challenged by several rival explanations.)

In one way or another, tests for spuriousness entail controlling for extraneous variables. The type of statistical control employed in our smoking–cancer example is very common in nonexperimental research. Its major drawback is that one can control statistically only for those variables that have been observed or measured as part of the research. Hence, the effects of any unknown or unmeasured variables cannot be assessed. A stronger test of nonspuriousness is provided in experiments

BOX 4.1 Problems in Causal Interpretation: The Case of Exercise and Heart Attacks

Of the three criteria needed to establish a causal relationship, the most difficult to assess is nonspuriousness. One can never be sure that a causal connection exists between correlated variables. Indeed, mistaken impressions of causality may remain undetected for years. An interesting example of this problem in social science research is related by psychologists Schuyler Huck and Howard Sandler (1979:151, 152, 227).

In recent years there has been much interest in the relative benefits of regular exercise. One controversial claim is that exercise can reduce the risk of heart attacks. An early study by Dr. J. N. Morris of London shows, however, just how difficult this is to establish. Examining drivers and conductors of London's double-decker buses, Morris found that the drivers were far more likely to suffer from heart disease and to die from coronaries than were the conductors. Since the drivers sat in their seats all day while the conductors ran up and down stairs to collect fares, he concluded that it was the differential amount of exercise inherent in the two jobs that brought about the observed differences in health. Before reading further, you might try to think of variables other than exercise that could have produced the difference in heart problems between the drivers and conductors. Morris uncovered one variable in a follow-up study, and Huck and Sandler mention two others.

Some time after the publication of the above results, Morris examined the records maintained on the uniforms issued to drivers and conductors and discovered that drivers tended to be given larger uniforms than conductors. Therefore, he concluded, differences in weight rather than exercise might be the causal factor. That is, heavier men, who were more coronary-prone to begin with, may have chosen the sedentary job of driver, whereas thinner men chose the more physically active job of conductor.

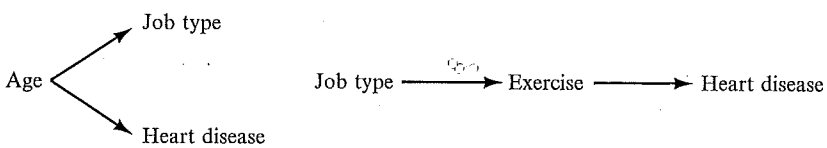
Another explanation is related to the amount of tension associated with the two jobs. As Huck and Sandler (1979:227) point out,

> The conductors probably experienced very little tension as they went up and down the bus collecting fares from the passengers; the worst thing that they probably had to deal with in their jobs was a passenger who attempted to ride free by sneaking around from one seat to another. Normally, however, we suspect that the conductors actually enjoyed their interaction with other people while on the job.
>
> But on the other hand, each driver had the safety of everyone on the bus as his responsibility. And as anyone who lives in or visits a city knows, driving in rush-hour traffic is anything but restful. Having to dodge pedestrians, being cut off by other vehicles, watching for signal changes—these activities can bring about temporary outbursts of anger and chronic nervousness. Imagine how it would affect your heart to be in the driver's seat of a bus for eight hours each working day!

Finally, a third variable that could account for the different rate of heart problems is age. If mobility or seniority or some other function of age were related to job assignment, then employees assigned to the driver jobs may have been older and

those assigned to the conductor jobs younger. And since we would expect more heart attacks among older persons, age rather than the nature of the job could be the causal variable.

In this example it is impossible to tell which cause—exercise, weight, job stress, or age—may have produced the observed differences in health between drivers and conductors. Since both weight and age are antecedent to job type and heart disease, either of these uncontrolled extraneous variables could have created a spurious relationship. However, if exercise or job stress were the correct interpretation, then the original relationship would not be spurious, since exercise and job stress specify intervening variables through which the job itself can make a person more or less susceptible to heart problems. The following diagram shows the difference in these two outcomes.



Of course, it is possible that two or more of these variables are operating jointly to produce the health differential between the two groups. The only safe conclusion is that we really do not know which interpretation is correct.

In general, correlation does not imply causation. All correlations must be interpreted; like any fact, they do not speak for themselves. To infer a causal relationship from a correlation, an investigator must detect and control for extraneous variables that are possible and plausible causes of the variable to be explained. The fatal flaw in Morris's study is that relevant extraneous variables were not controlled; without directly assessing the effects of such "hidden" causes, we cannot tell which interpretation is valid.

through a process called "*randomization*" that makes it theoretically possible to control for all extraneous variables. Experimental controls are discussed in chapter 7, and causal analysis techniques involving statistical manipulation of nonexperimental data are discussed in chapter 16.

Causation, intervening variables, and theory

Recall that extraneous variables may be either antecedent or intervening in relation to the independent and dependent variables. An antecedent variable, which is causally prior to both the independent and dependent variables, can produce a spurious relationship. By contrast, the identification of an intervening variable or mechanism linking the independent and dependent variables strengthens the causal

inference. Indeed, this is sometimes advocated as a fourth criterion—in addition to association, direction of influence, and nonspuriousness—for establishing that one variable causes another (see Hyman, 1955). For example, one may argue that the belief that smoking causes lung cancer will be enhanced considerably if and when it is established that certain chemical agents from cigarettes produce cancerous cells. Knowing the causal process through which smoking produces cancer would provide one last shred of evidence against a spurious correlation. However, once the criterion of nonspuriousness is firmly established, a causal relationship is generally inferred even if the intervening mechanisms are not known. Few scientists today doubt that smoking causes cancer. Thus, while specifying the intervening variables in a relationship may lead to a better theoretical understanding and more accurate prediction, it is "not part of the minimum requirements for demonstrating causality. Holding a match to a pile of leaves is a cause of their bursting into flame, even if one cannot describe the intervening chemical reactions" (Hirschi and Selvin, 1967:38).

KEY POINT

Antecedent variables may create spurious associations; *intervening variables* clarify the causal connection between variables.

Though not a necessary causal criterion, intervening variables are nonetheless an essential part of scientific inquiry. Often, in fact, this is what the development and testing of theory is all about. For example, Durkheim's theory of suicide stipulated the intervening causal mechanism—social integration—for a number of relationships. The reason that fewer suicides are found among Catholics than among Protestants and that there are fewer suicides among married than among single people, according to the theory, is that being a Catholic and being married each engenders a greater sense of social integration, which in turn reduces the likelihood that anyone within the group will commit suicide. Broh tested the effects of several intervening variables. Having found that participation in interscholastic sports benefits students' academic performance, she also showed that this relationship was best explained by developmental and social capital theories. High school athletes develop a greater work ethic and sense of control over their lives than do nonathletes, and their athletic participation generates stronger ties among students, parents, and teachers, all of which have positive effects on academic performance.

Yet, theory plays a much larger role in causal analysis than specifying intervening variables. Theories not only render a more complete understanding of the causal processes that connect events but also provide the general framework for investigating the nature of all relationships. Theories tell the researcher which relationships to observe, what extraneous variables are likely to affect the relationships, and the conditions under which a causal relationship is likely to exist. It is only in terms of some theory, in short, that the researcher can determine how to assess the meaningfulness of a "weak" association and how to test for direction of influence and nonspuriousness. Thus, we see again the importance of the interplay between theory and research in science. Theory guides research, and research provides the findings that validate and suggest modifications in theory.

Formulating Questions and Hypotheses

Having introduced the language of units of analysis, variables, and relationships, we are now ready to return to formulations of research questions. According to Glenn Firebaugh (2008), research questions should meet two fundamental criteria: They should be researchable and interesting. A researchable question is one that is answerable through empirical research. As we saw in the preceding chapters, some questions cannot be answered because they are beyond the realm of science (e.g., Does God exist? Is capital punishment morally wrong?) or because ethical considerations rule them out. Other questions may be too grand. Novice as well as experienced researchers often begin with questions that are so broad in scope that they provide little or no immediate direction for research. For example, Stanley Schachter, well known in part for his work on affiliation, was interested in what motivates people to be around others (Evans, 1976:159), a question nearly as encompassing as the entire field of social psychology. Before Schachter could begin to do research, he had to reformulate this question so that it pointed to identifiable variables and relationships. After reading and speculating about "affiliative tendencies," Schachter eventually focused on the question of how fear affects the desire to be with others.

Firebaugh's second criterion for good research questions is that they should be interesting to you and, especially, to others. What makes a question interesting? Primarily it should contribute to ongoing conversations in a scientific field of study. The only way to know if your question speaks to such conversations is to learn about the latest research through a thorough review of the literature. Here is where you will find the major theories, key findings, and unresolved issues in the field. Becket Broh's literature review revealed several studies which found a positive association between high school students' participation in sports and grades. Relatively few studies, however, had examined the effect of other extracurricular activities, and even fewer had attempted to test theoretical explanations of how sports participation enhances educational achievement. After considering various theories, Broh formulated three unanswered research questions:

- Why does sports participation boost students' achievement? Does sports participation benefit students' development and social networks, and are these the mechanisms that link participation to educational outcomes?
- Are the educational benefits of sports participation unique to sports, or do nonsports extracurricular activities also promote achievement?
- Do nonsports extracurricular activities benefit students' development and social networks? (Broh, 2002:73)

Broh's research questions were "interesting" because they extended prior research in new directions. (Also of interest, albeit much less common, are research questions that challenge well-established findings.) Her questions extended research in two ways: by broadening the question to include participating in nonsports activities as well as interscholastic sports and by identifying an intervening variable (social capital) that accounts for the effect of sports participation on grades. Two other ways of extending prior research suggested by Firebaugh (2008) are investigating whether a finding applies to a new population (imagine Broh replicating her study outside the United States) and whether a finding applies to a new time period (e.g., would a similar association be found in the first half of the twentieth century?).

The tentative answers to research questions are called "hypotheses." Formally defined, a **hypothesis** is an expected but unconfirmed relationship between two or more variables. Hypotheses come from a variety of sources, including everything from theory to direct observation to guesses and intuition. Sometimes the formulation of hypotheses is the principal outcome of research. At other times, hypotheses are never made explicit, even though they implicitly guide research activities. However, whenever the research objective is clearly one of testing relationships among variables, hypotheses should be stated formally and precisely so that they carry clear implications for testing the stated relations (Kerlinger, 1973:18).

While stated in a variety of ways, all hypotheses should speculate about the nature and form of a relationship. An adequate hypothesis statement about two variables should indicate which variable predicts or causes the other and how changes in one variable are related to changes in the other. For example, if we thought that education generally increases tolerance, we might hypothesize that "an increase in education will result in a decrease in prejudice." This statement implies two features about the relationship: first, which variable causes, explains, or predicts the other (education predicts prejudice); and second, how changes in one variable are related to changes in the other (as education increases, prejudice decreases). Hypotheses that specify the form of the relationship are said to be testable because it is possible, assuming each variable has been measured adequately, to determine whether they are true or false, or at least whether they are probably true or probably false.

There are several forms of expressing testable hypotheses, such as the one relating education and prejudice:

1. *"If–then" (conditional) statements.* These statements say that if one phenomenon or condition holds, then another will also hold. An example would be, "If a person has a high level of education, then he or she will have a low level of prejudice." Alternately, one could say, "If a person has a low level of education, then he or she will have a high level of prejudice." In logic, such statements are called **conditionals.** A conditional consists of a connection between two simple statements, each pointing to a condition or category of a variable. As used in science, the connection asserted by a conditional is that the condition (or variable category) following the "if" *causes* the condition (or variable category) following the "then." Social research hypotheses seldom are specified in this form. Still, it is always possible to restate testable hypotheses as conditionals; and by using this standard logical form, some scientists contend that it is easier to ascertain the kinds of inferences that legitimately can be made from research findings to the hypothesis (McGuigan, 1993:37).

2. *Mathematical statements.* Many hypotheses may be stated in the form of the standard mathematical formula $Y = f(X)$, which reads "Y is a function of X." An example is Einstein's famous formula $E = mc^2$ (i.e., energy equals mass times the speed of light squared). Of course, though once a hypothesis, $E = mc^2$ is now called a scientific "law" because it has been confirmed repeatedly. Mathematical formulas represent precise formal statements of hypotheses. Because variables in social research generally

are measured with less precision than in the physical sciences, social scientists seldom state hypotheses in this form. Nonetheless, mathematical formulas are the ideal in science because they yield precise predictions and express complex relationships parsimoniously. The mathematical form is also equivalent to the conditional. $Y = f(X)$ merely says "If (and only if) X is this value, then Y is that value" (McGuigan, 1993:38).

3. *Continuous statements.* Hypotheses of the form "The greater the X, the greater (or lesser) the Y" indicate that increases in one variable (X) are associated with increases (or decreases) in another variable (Y). For example, as education increases, prejudice decreases; or, expressed in slightly different form, the higher the level of education, the lower the prejudice.

4. *Difference statements.* Statements in this form assert that one variable differs in terms of the categories of another variable. For example, people with high education are less prejudiced than people with low education. Whether "continuous" or "difference" statements are used to express a hypothesis will depend on whether the variables in the hypothesis are quantitative or qualitative. If both variables could be "quantified," then the relationship could be stated in the continuous form. But if either variable consisted of discrete categories, such as "high" and "low" prejudice, then the relationship would need to be stated in the difference form. In either case, it is easy to transform the hypothesis statement into a conditional. For example, the statement "The higher the level of education, the lower the level of prejudice" becomes "If education is high, then prejudice will be low."

Both continuous and difference statements clearly specify the *form* of the relationship. However, both types of statements are ambiguous about the causal connection between variables. Ordinarily, a statement of the form "The greater the X, the greater the Y" is meant to imply that X causes Y. But it also can mean that Y causes X or that X and Y cause one another. As you saw in our discussion of causation, it is important to know which variable is presumed to cause the other; in addition, since most hypotheses take the form of continuous or difference statements, the causal linkage can be problematic. Fortunately, in most research articles and reports, researchers make this connection clear in their discussions of the hypothesis.

At this point we should reiterate an important tenet of scientific research. Even though all of the above forms of expression would appear to assert that the relationship is absolutely true or false, hypotheses in science can only have probabilistic, not exact, confirmation. Thus, while the statement "If X, then Y" logically can only be true or false, it is assumed that observations will show it to be "probably true" or "probably false."[5] One often sees this kind of assumption built into statements of hypotheses in the form of qualifiers such as "tends to" or "in general." For example, one might say that "increased education will *tend* to reduce prejudice." Such statements acknowledge that tests of hypotheses are always restricted by the limited accuracy of measures and the inability to specify and control all the variables affecting events.

The idea that hypotheses can only be judged to be *probably* true or *probably* false gives rise to another type of hypothesis associated with tests of statistical significance. Significance tests indicate the probability that an observed relationship between two variables is due to chance variation; if a test is significant, this means

that there is a low probability, say less than 5 in 100, that the relationship occurred by chance, and we therefore conclude that the relationship exists or is real. In statistics, the hypothesis that a relationship is a chance occurrence, which is the opposite of the hypothesized relationship, is called the **null hypothesis.** In other words, evidence supports the research hypothesis only insofar as the null hypothesis is rejected with a known probability of error. Researchers do not always make reference to the null hypothesis when testing relationships; however, it is central to the logic of significance testing and is always implicit.

How a hypothesis is expressed in a given study will depend on several factors: the researcher's discretion, the present state of knowledge about the research problem, and whether qualitative or quantitative variables are involved. Regardless of how hypotheses are expressed, however, they should indicate at least the form of the relationship between variables and, ideally, should specify the causal linkage between variables; for ultimately it is causal relationships in which scientists are interested.

Research Purposes and Research Design

In the previous pages we have emphasized the role of relationships in social research. The sense that social scientists make of the social world is expressed in terms of relationships among variables. Furthermore, although they may be only vaguely defined or implicit and unknown to the researcher, anticipated relationships structure the researcher's every activity, from deciding which variables to measure to deciding how observations should be made and interpreted. Not all research, however is conducted for the immediate purpose of testing relationships. Research is undertaken for three broad purposes: (1) to *explore* a phenomenon such as a group or setting in order to become familiar with it and to gain insight and understanding about it, frequently in order to formulate a more precise research problem for further study; (2) to *describe* a particular community, group, or situation as completely, precisely, and accurately as possible; and (3) to examine and formally to *test relationships* among variables. Whether a study is conducted primarily for the purpose of exploration, description, or testing of relationships is important to know, for these three functions have different implications for research design.

Exploratory studies are undertaken when relatively little is known about something, perhaps because of its "deviant" character or its newness. Falling into this category of research would be observational studies of street gangs and radical political and religious movements; clinical case histories of persons, groups, and events; and anthropological accounts of entire cultures. When exploring a topic or phenomenon about which one knows very little, one necessarily begins with a general description of the phenomenon. This sounds easy but in fact is probably the most difficult kind of study for the novice researcher to undertake. There are no clearly delineated independent and dependent variables and, therefore, no preset categories of observation and analysis. The researcher may have few, if any, guidelines to help determine what is important, whom to interview, or what leads to follow. For these reasons, the research plan in an exploratory study is more open than in other kinds of research. Decisions are made about the kinds of instruments needed (e.g., pho-

tographic equipment, tape recorders) and the key persons with whom one will need to speak at first, but the paths down which these initial steps may lead are almost impossible to foresee. In chapter 11, on field research, we discuss some data-gathering approaches that frequently are devoted to exploration.

The objective of a **descriptive study**, as the name implies, is to describe some phenomenon. The nature of the description, however, differs considerably from exploratory research. A descriptive study is much more structured. Basically a fact-finding enterprise, it focuses on relatively few dimensions of a well-defined entity and measures these dimensions systematically and precisely, usually with detailed numerical data. The information is gathered from a set of cases that are carefully selected to enable the researcher to make estimates of the precision and generalizability of the findings. Examples of descriptive studies are the various censuses conducted periodically at the local and national levels. A census may provide information about everything from the age and racial composition of a community to its employment and housing costs. Also having a descriptive purpose are the ubiquitous opinion polls that attempt to estimate the proportion of people in a specified population who hold certain opinions or views or who behave in certain ways. For example, how many favor capital punishment? How many say they will vote for candidate A in the next election?

The third purpose for which research is conducted is to test relationships. Studies with this purpose are sometimes called **explanatory studies** because they formally seek the answers to questions and hypotheses. Since all research involves description at some level, the primary difference between descriptive and explanatory studies lies in the scope of the description. Purely descriptive research operates at a lower level of description by merely seeking information about isolated variables, whereas explanatory research goes beyond this step to a description of relationships among variables.

Both descriptive and hypothesis-testing research are highly structured and must be carefully planned. With these kinds of research it is important, therefore, to have a complete, detailed strategy worked out before the data are collected. This preliminary strategy or outline is what we have called the **research design**. Basically, it consists of a clear statement of the research question as well as plans for gathering, processing, and interpreting the observations intended to answer the question. To formulate a research design is to anticipate the entire research process, from beginning to end. To do this, one needs to have an adequate knowledge of every stage of social research. Examining these stages now will serve not only to clarify the key components of research design but also to introduce the reader to the remainder of the book.

Stages of Social Research

Figure 4.5 outlines the major steps in the research process. Although most research follows this pattern, steps may be omitted, depending on the type of research, and there may be numerous feedback loops. In short, this is an idealized model; some research is much less orderly than it implies.

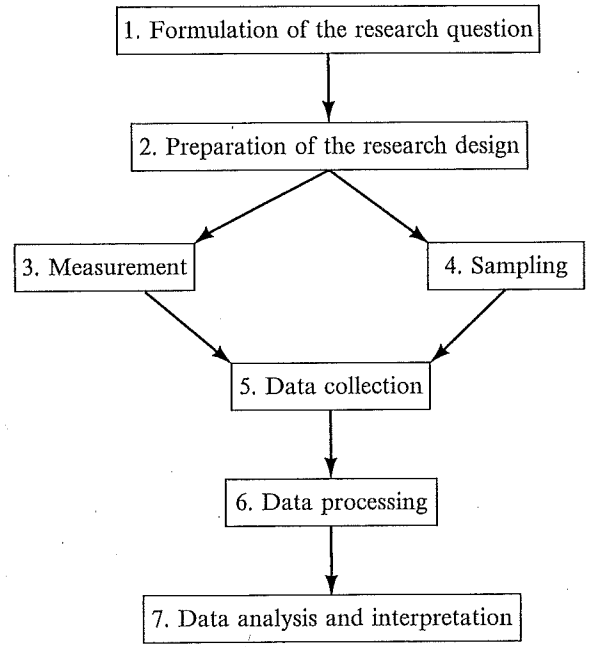


FIGURE 4.5. The stages of social research.

Stage 1: Formulation of the Research Question

Research begins with a question. Questions initially chosen almost always require more precise formulation to be amenable to research. From a general idea, one must decide more specifically what one wants to know and for what purpose one wants to know it. The best ideas for formulating questions are likely to be found in the scientific literature. Besides helping to narrow and refine the research question, a review of the literature may reveal its broader theoretical significance. Literature reviews also may help to identify relevant control variables and should suggest pertinent methods and procedures by indicating how other researchers have addressed the question.

Stage 2: Preparation of the Research Design

Once the question has been clearly formulated, the researcher must develop an overall plan or framework for the investigation. To do this, he or she must, in effect, anticipate all of the subsequent stages of the research project. Preliminary decisions have to be made about what sort of observations are needed to answer the research question or to provide an adequate test of the hypothesis. The researcher must then select an appropriate strategy for making the observations—experiment, survey, field research, or use of available data. Each of these approaches has its unique strengths and weaknesses that determine its suitability for given questions.

Often, the best strategy, as we argue in chapter 13, is a combination of approaches. Within the context of selecting an overall strategy, decisions also must be made on the unit of analysis, on which variables to observe and control and how they should be measured, and on how best to analyze the data. Thinking through all of these issues in advance should prevent serious mistakes and omissions in a study. Of course, not all problems can be foreseen, especially in exploratory research, and many of the decisions at the design stage will be arbitrary and subject to change.

Two issues—measuring variables and selecting units of analysis—warrant special attention and are ordinarily worked out in detail after the basic research design is complete. These are therefore considered separate, concurrent stages.

Stage 3: Measurement

Part of the research plan involves devising operations that will link specific concepts to empirically observable events. This process of operationalization and measurement will be taken up in chapter 5, a substantial part of which describes techniques for assessing the quality or goodness of measures.

Stage 4: Sampling

Besides deciding on the unit of analysis, the researcher must also determine how many units should be selected and how to go about choosing them. The problems and methods related to sampling are addressed in chapter 6.

Stage 5: Data Collection

Once the research design is completed, the researcher is ready to make the observations—to collect the data. Since the general approach to data collection affects decisions about both measurement and sampling, the approach to be taken is determined early in the design phase. As we have pointed out, there are four basic choices: experiment, survey, field research, and use of available data. The underlying logic of each of these approaches, as well as their distinctive problems, advantages, and disadvantages, are covered in chapters 7 through 12.

Stage 6: Data Processing

Having made the observations, the researcher is ready to analyze and interpret them. In most cases, however, the form of the data does not allow easy interpretation. The data must be transformed or processed for analysis. Most data processing involves, at some point, the hardware and software of computers. Chapter 15 begins with a discussion of data processing, giving special attention to the preparation of data for computer analysis.

Stage 7: Data Analysis and Interpretation

Once the data are ready for analysis, they must be manipulated further so that their meaning and bearing on the problems and hypotheses that initiated the inquiry can

be extracted. There are several types of analysis, many of which involve statistical tests, a topic that is beyond the scope of this textbook. Without any training in statistics, however, one can learn how to read and to present data properly in tables and graphs and when it is appropriate to apply certain statistical procedures. Also, it is possible to develop a solid understanding of the logic of analyzing causal relationships. These topics are also covered in chapters 15 and 16. Having analyzed the data, one draws conclusions about the hypotheses and theory that guided the research and, if appropriate, assesses the practical implications of the findings. Finally, one reports the results.

Summary

Our purpose in this chapter has been to introduce the basic terms in the technical language of social research as these terms relate to question formulation and research design. There are many sources of ideas and factors affecting topic selection, including theoretical and practical relevance, social premiums, and the personal values and resources of the researcher.

General topics are refined and research questions emerge in the course of working out the overall study plan, or research design, the key elements of which are units of analysis, variables, and relationships. Units of analysis are the entities about whom or which the researcher gathers information. These include individual people, groups and organizations of all kinds, communities, nations, and social artifacts. Researchers often aggregate information about individuals in order to describe the social unit that the individuals comprise. In general, conclusions should be restricted to the unit of analysis to which the data pertain. Thus, conclusions about individuals should not be drawn from group-level data; otherwise, one risks committing an ecological fallacy.

Variables are the characteristics of units that may vary in successive observations. Research focuses on explanatory variables while attempting to eliminate the influence of extraneous variables. The dependent variable is the explanatory variable that the researcher tries to explain or predict; an independent variable is a presumed cause of the dependent variable. Other variables may be either antecedent or intervening in relation to specific independent and dependent variables, and either controlled or uncontrolled during the course of the study. Quantitative variables have categories that express numerical distinctions, whereas qualitative variables involve differences in kind rather than in number.

Anticipated and predicted relationships guide all social research. Relationships occur when the changes in two or more variables form a predictable pattern. This pattern has two properties—strength and form—which are depicted differently according to whether the variables are qualitative or quantitative. In addition to statistical measures of strength of association, such as the correlation coefficient, tests of statistical significance indicate whether a relationship is likely to exist or whether it is likely to be the product of random processes. Social scientists are particularly interested in causal relationships. To establish that variable X causes variable Y, one must show that X and Y are statistically associated, that the direction of influence is from X to Y, and

that the association between X and Y is nonspurious. Theories that identify intervening variables also may strengthen the inference that X causes Y.

Research questions should be scientifically answerable and also interesting by contributing to ongoing research. Questions ask about a relation between variables, whereas a hypothesis is a conjecture about the nature and form of a relationship. In social research, a hypothesis may take the form of a conditional, a mathematical equation, a continuous statement, or a difference statement. Although much social research is undertaken to test hypotheses, it also may be done for the purposes of exploration and description.

Key Terms

- unit of analysis
- aggregate data
- ecological fallacy
- variable
  - explanatory variable
  - extraneous variable
  - dependent variable
  - independent (predictor) variable
  - antecedent variable
  - intervening variable
  - control variable
  - quantitative variable
  - qualitative variable
- relationship
  - positive (direct) relationship
  - negative (inverse) relationship
  - correlation coefficient
  - statistical significance
- causal relationship
  - association
  - direction of influence
  - spurious relationship
  - nonspuriousness
- hypothesis
  - conditional
  - mathematical statement
  - continuous statement
  - difference statement
- null hypothesis
- descriptive research
- exploratory research
- hypothesis testing (explanatory) research
- research design

Exercises

1. Suppose an anthropologist studies several villages in southern India. For each village she has data on the number of people and the average age of men and women, and she has computed an index of the monetary value of various consumer goods in the village (e.g., the value of the total number of refrigerators, televisions, and air conditioners divided by the size of the population). An analysis of these data reveals that the villages with populations having the lowest average age also have the highest average dollar value of modern consumer goods (Bernard, 1994). What is the unit of analysis in this study? Can the anthropologist conclude from this association that young people purchase more modern consumer goods than older people? Why or why not?

2. For each of the following research questions or hypotheses, identify the *unit of analysis*, *independent variable*, and *dependent variable*.

   a. Boys whose parents are divorced or separated have more behavior problems than boys living with both biological parents.
   b. The number of sexually explicit magazines (such as *Hustler* and *Playboy*) sold in each of the fifty states is positively correlated with the number of reported rapes.
   c. Residents of rural communities are less tolerant of people holding controversial views than urban residents.
   d. How does the level of economic development of a country affect the level of human services provided to its people?

3. The contact hypothesis, described recently as "one of the most durable ideas in the sociology of race and ethnic relations," has been investigated extensively over the past 50 years. In its simplest form, the hypothesis is that contact between members of different races fosters positive racial attitudes. Imagine you want to test this hypothesis. Specifically, you hypothesize that prior interracial contact in schools and neighborhoods will lead to greater racial diversity in one's current circle of friends. Assume you are going to do a survey at your college (or a local college).
   a. Suppose you find a correlation of .23 between prior interracial contact and racial diversity of friends. What does this indicate about the relationship?
   b. Now suppose the correlation was significant at $p < .01$. What does this information reveal about the relationship?
   c. An important control variable in your study would be race. Restate the hypothesis with race as a control variable.

4. In a campus survey, a researcher finds a statistically significant correlation of −.22 between frequency of exercise and depression; that is, the more students exercise, the less depressed they are. Can he conclude from these data that lack of exercise is one cause of depression? Why or why not? Now carefully explain how gender could create a spurious association between frequency of exercise and depression.

5. Based on a national survey, Christopher Ellison and Kristin Anderson (2001) examined the relationship between religious involvement and domestic violence among U.S. couples. Several previous studies had shown that regular religious attendance is inversely related to abuse among both men and women. The researchers analyzed several extraneous variables including respondents' age, education, employment status, and psychological depression.
   a. Select one of these variables and explain how it could create a *spurious association* between religious involvement and partner violence.
   b. The researchers hypothesized that "religious communities may reduce the risk of abuse by enhancing the levels of social integration and support enjoyed by their members." Are they hypothesizing that the relationship between religion and abuse is spurious? Explain.

6. Formulate three hypotheses, each relating two variables from the following list. For each hypothesis, specify the independent and dependent variables.
   Gender (male/female)
   Level of education (highest grade completed)
   Marital status (married/widowed/divorced/separated/never married)
   Party identification (Republican/Democrat/Independent/none)

Belief in life after death (yes/no)
Attitude toward capital punishment (favor/oppose)

7. Express one of the hypotheses formulated in question 6 in the form of a (a) conditional, (b) continuous statement (if possible), and (c) difference statement.

Notes

1. When Frans van Poppel and Lincoln Day (1996) tested Durkheim's theory with individual-level data from the Netherlands at roughly the same period as Durkheim, they found that the Catholic–Protestant difference in suicide rates could be accounted for by how the causes of deaths were recorded. Catholic deaths were far more likely than Protestant deaths to be recorded as "sudden death" or "death from ill-defined or unspecified cause"; these deaths would have been categorized as suicides had they occurred among Protestants. Although Durkheim's theory may be supported by other data, this analysis shows the pitfalls of using group-level data to make inferences about individuals.

2. We caution the reader not to confuse this distinction with the distinction between quantitative and qualitative research or quantitative and qualitative measurement. Both quantitative researchers—typically survey researchers and experimenters who rely on numbers and statistical methods—and qualitative researchers—who do intensive interviews, participant observation, and depth analyses of historical materials and rely on discursive methods—use both kinds of variables in their analyses (see King, Keohane, and Verba, 1994).

3. The table reports a type of "average" called the "median," which is equal to the fiftieth percentile: Half of the group earn less and half earn more.

4. Determining the level of confidence that a relationship exists, as well as estimating the strength of the relationship, is the job of statistics and is therefore beyond the scope of this textbook. In general, confidence and strength will be greater to the extent that (a) the observed differences between groups are large, (b) the computed averages are based on a large number of people, and (c) people in the same group are quite similar to one another and different from those in other groups.

5. Determining whether a statement is probably true or probably false, like assessing the existence and strength of a relationship, is a matter of applying the laws of probability, which is the subject of statistics.

5

Measurement

Measurement is the process of assigning numbers or labels to units of analysis in order to represent conceptual properties. This process should be quite familiar to the reader even if the definition is not. For example, we measure every time we "rate" something, such as a movie, restaurant, or blind date: "The movie was 'pretty good'"; "the new restaurant definitely merits a four-star rating—the decor, service, and food are excellent, and the price is right"; "on a scale of 1 to 10, I would give him a 2—not the worst date I have ever had but close to it." In a somewhat more refined manner, you may have "measured" someone's "intelligence" by his or her grade-point average, and you probably have measured your weight on a bathroom scale. Each of these examples contains the essentials of measurement: Labels or numbers ("pretty good," "four-star," a "2," a particular GPA, a pointer reading on a scale) are assigned to objects (movies, restaurants, people) to represent properties (the overall quality of a movie, restaurant, or date; intelligence; weight).

There is a difference, of course, between these everyday examples of measurement and the process of measurement in social research. In the examples, the rules for assigning labels or numbers to objects are more or less intuitive, whereas in social research these rules must be spelled out in detail. Scientific norms require that we fully describe our methods and procedures so that others can repeat our observations and judge the quality of our measurements. In this chapter we outline the measurement process, provide several examples of measurement in social research, and then discuss three criteria for evaluating the nature and quality of measurements.

The Measurement Process

The measurement process begins as the researcher formulates his or her research question or hypothesis. Every question or hypothesis contains terms—concepts or variables—that refer to aspects of reality in which the researcher is interested. Measurement involves thinking about what these terms mean in both an abstract and an empirical sense. The ultimate goal of measurement is to specify clearly observable referents of the terms contained in one's hypothesis, but to get to this point one must first consider the abstract meaning of the terms. Thus, the entire measurement process consists of moving from the abstract (concepts) to the concrete (measures of concepts).

Conceptualization

Recall Beckett Broh's (2002) study of extracurricular involvement and academic achievement, which we introduced in chapter 4. Broh theorized that extracurricular activities increase social capital, which in turn enhances academic achievement. This theory relates three terms: "extracurricular activities," "social capital," and "academic achievement." The terms are merely labels for concepts. To understand fully what the theory means and to arrive at an appropriate set of observations for testing it, one must know the meaning of these concepts. Thus, the initial step in measurement is to clarify the concepts embedded in one's theories and hypothesis with words and examples, ultimately arriving at conceptual (also called "theoretical") definitions.

Broh considered the meaning of each of her key concepts. She relied on common understandings of "extracurriculum" and "academic achievement." *Academic achievement* generally refers to "cognitive learning outcomes which are products of [school] instruction or aimed at by [such] instruction" (Helmke and Schrader, 2001:13552); the *extracurriculum* pertains to school activities outside the set of courses offered at a school. *Social capital*, according to Broh (2002:72), "is generally recognized as the ability to accrue benefits through membership in social networks."

Theoretical definitions of this sort direct the search for appropriate measures of concepts, establish a basis for judging the quality of one's measures, and enable others to evaluate the meaning of one's research findings. Such definitions, however, are not worked out anew with each research project, nor are they the arbitrary invention of each individual investigator making sure that others understand what he or she means. If this were so, there would be no basis for developing a shared body of knowledge. Rather, the process of formulating and clarifying concepts, called **conceptualization**, is linked to theory testing and construction. This ongoing process may occur prior to any particular empirical investigation, and it usually continues through research as theories and their constituent concepts are refined and elaborated.

Broh derived her theoretical definition of "social capital" from analyses of this concept by, among others, sociologists James Coleman and Alejandro Portes. Coleman (1990:304) distinguished social capital, "embodied in the relations among persons," from human capital, "embodied in the skills and knowledge acquired by an individual." He also argued (Coleman, 1988) that the family is a primary site of social capital, while Portes (1998) analyzed important extrafamilial networks. Taking into account these theoretical analyses, Broh (2002:72) concluded that "participation in sports and other extracurricular activities may serve to create social capital within the family by providing opportunities for increased social interaction between the parents and the child"; these activities further create social capital outside the family "by offering opportunities for the formation and intensification of social ties among students, parents, and the school." Operating through these networks, social capital facilitates academic achievement in two ways: It exerts social control over students by encouraging them to comply with school norms and values, and it provides channels for students to acquire important educational information and resources.

The conceptualization of complex concepts such as "social capital" often involves making careful distinctions among similar ideas and breaking down the concept into various components or dimensions. Thus, in her conceptualization, Broh distinguished between "social capital" and "human capital," and she identified different social ties—parent–child, parent–teacher, student–teacher, and student–student—in which social capital is grounded. With regard to the extracurriculum, she carefully noted the numerous activities in which students participate, such as sports, drama, music, and vocational clubs; and she further distinguished between types of sports activities, such as interscholastic sports, intramural sports, and cheerleading. This sort of analysis further clarifies the meaning of concepts and generates more precise statements of problems and hypotheses. Beginning with the general theoretical relationship between the extracurriculum and social capital, Broh concentrated on the less abstract relationship between participation in interscholastic athletics and social ties among parents, students, and teachers.

Two aspects of concepts are especially relevant to the measurement process. First, a concept may signify a single category, such as "male" or "B student," or a concept may imply several categories or values, such as "gender" or "grade in English." Measurement assumes the possibility of assigning *different* values or categories to units of analysis; hence, we measure concepts that vary, which we refer to as "variables." Second, many social science concepts are not directly observable. For example, we cannot "see" social capital or academic performance in the same sense that we can see a table or a horse or the color red. On the other hand, while we cannot see school performance, we can observe how many questions students answer correctly on standardized tests and the grades they receive in core academic subjects such as English and math. After conceptualization, the next step is to identify such manifestations of one's concepts, and it is at this point that we move from a language of concepts to a language of variables.

Just where this shift in language occurs is difficult to pinpoint, and researchers often use the terms "concept" and "variable" interchangeably. Still, it is important to realize that these terms connote different levels of abstraction. Once the researcher begins to speak in terms of a variable, he or she generally has some observable events in mind that represent the underlying concept.

Operationalization

Once the meaning of a concept has been clarified and the concept is construed as a variable, the researcher begins the process of **operationalization**. The counterpart of a conceptual definition is an operational definition. An *operational definition* describes the research operations that specify the values or categories of a variable. Broh operationalized the variables in her study by using responses to survey questions from the National Educational Longitudinal Study (NELS), which administered questionnaires to students in the 8th, 10th, and 12th grades. In both the 10th and 12th grades, for example, students were asked whether they had participated on an interscholastic sport team—that is, "your school team competes with other schools' teams." Based on these questions, Broh (2002:74) operationally defined *participation in interscholastic sports* as "whether a student participated in

interscholastic sports during both the 10th and 12th grades (1 = participated in both years, 0 = did not participate in both years)."

To understand better the notion of operational definitions, let us consider a simple illustration from everyday life. Suppose your friend bakes you a delicious carrot cake. You ask your friend how he made it because you would like to make one. Your friend says, "Oh, you take some carrots, flour, sugar, eggs, and so forth, add some nuts, bake it, and voilà!—you have a carrot cake." Would you be able to make an identical cake with these directions? Not likely. What you need is an operational definition of your friend's concept of "carrot cake." You would need to have the complete directions—that is, details such as all of the ingredients, the amount of each ingredient to use, the steps necessary to combine the ingredients, the oven temperature, and baking time. In short, your friend's operational definition should look like an ordinary recipe. Using the recipe (operational definition), you should be able to produce a very similar cake.

KEY POINT

Measurement begins with *conceptualization*, the clarification of the meaning of a concept, and ends with *operationalization*, a procedure detailing the observational categories or values representing the concept.

Many operational definitions are possible; social scientists must choose or develop one that they believe corresponds reasonably well to the concept in question. To return to the carrot cake example, how could you be certain that your friend's recipe represented an "authentic" carrot cake? Are nuts really an essential ingredient? Suppose you substituted whole-wheat flour for white; would you still have a carrot cake? If you compared his recipe with others, you would no doubt find some differences. How are you to conclude which is the correct recipe? In the end you would find that there is no correct recipe, but you would still have to decide for yourself whether your friend's operational definition (recipe) "really" corresponded to your idea of what a carrot cake is. As you can see, operational definitions, so essential to social research, are somewhat arbitrary and restricted expressions of what a concept "really" means.

When creating operational definitions, a researcher may consider many different empirical representations or indicators. An **indicator** consists of a single observable measure, such as a single questionnaire item in a survey. If each indicator classified units in exactly the same way, the choice would be wholly arbitrary. However, no two indicators measure a given concept or variable in the same way, and no one indicator is likely to correspond perfectly to its underlying concept. Indicators provide imperfect representations of concepts for two reasons: (1) they often contain errors of classification and (2) they rarely capture all the meaning of a concept. Consider one indicator of social capital in Broh's study: whether students talk to their teachers outside class about their school work. Whether this promotes social capital will depend on why students speak with their teacher. Are they doing so for disciplinary reasons at the insistence of the teacher or because they seek guidance on an assignment? And, of course, there are many sources of social capital other than student–teacher relations, including students' relations with other students and with parents and parent–teacher relations.

Because of the imperfect correspondence between indicators and concepts, researchers often choose to rely on more than one indicator when operationalizing a concept. Sometimes several indicators of a given concept are analyzed separately, yielding multiple tests or cross-checks of a hypothesis. At other times, indicators are combined to form a new variable, as when answers to several questions, each a distinct indicator, are combined to create the variable "IQ score." Researchers generally use several indicators to operationalize a complex concept like social capital. To measure social capital between students and parents, Broh summed the responses to three questions which asked students how often they talked to their parents about school courses, programs and activities, and studies. With simpler concepts like "participation in interscholastic sports," a single indicator will suffice. James Davis (1971:18) suggests this rule for deciding whether to use single or multiple indicators to represent a concept: "If you have to ponder about the best way to measure a key

BOX 5.1 Operationalizing the Concept of "Religiosity"

Sociologist Ronald Johnstone (1983:289–304) provides an excellent illustration of the processes of concept clarification and measurement with respect to the concept of "religiosity." Johnstone notes, first, that the conceptualization of religiosity should begin with a definition of *religion*: "Religion is a set of beliefs and practices, centered around a belief in the supernatural and an orientation toward the sacred, that are shared by members of a group" (p. 290). Having defined religion, one could say that religiosity is the concept of "being religious." This conceptual definition suggests several ways of operationalizing religiosity:

1. The group affiliation approach focuses on the religious group to which a person belongs. In this case, a person is religious if he or she professes to be a member of some religious group, such as a Protestant or Catholic, while nonmembers are nonreligious. The General Social Survey (GSS) (Davis, Smith, and Marsden, 2007) asks respondents, What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion? Those who answer "no religion" would be devoid of religiosity, but this fails to capture the intensity of belief or degree of religious interest, which is closer to the concept of individual religiosity.
2. Single indicators of individual religiosity typically have emphasized "ritual participation"—in particular, the frequency of attendance at formal religious services. The GSS includes such a measure:

   How often do you attend religious services?

   [ ] Never
   [ ] Less than once a year
   [ ] About once or twice a year
   [ ] Several times a year
   [ ] About once a month
   [ ] 2–3 times a month
   [ ] Nearly every week
   [ ] Every week
   [ ] Several times a week

Related questions tap frequency of prayer or extent of involvement in the total organizational life of the congregation. The problems with such measures are that they (1) are subject to bias and (2) assume that religiosity is a one-dimensional phenomenon. Attendance may reflect, for example, family or other group pressures, habit, or the desire to socialize with friends after services, rather than religious commitment. Moreover, people who rank high on one dimension of religiosity may rank low on another, so different conclusions about the impact of religion could be reached, depending on which measure of religiosity is used. For example, Stephen Ainlay and James Hunter (1984) found that while church attendance declined with age among persons over 50, nonchurch participation measures such as Bible reading and listening to religious radio programs increased with age.

3. A frequently cited example of the multidimensional approach is Charles Glock's (1962) identification of five dimensions of religiosity: experiential, ritualistic, ideological, intellectual, and consequential. Joseph Faulkner and Gordon DeJong (1966) have developed sets of questions for measuring each of Glock's dimensions. For example, the experiential dimension concerns the degree of emotional attachment to the supernatural. Items developed by Faulkner and DeJong to measure this dimension include the following:

   Would you say that one's religious commitment gives life a certain purpose which it would not otherwise have? (1) Strongly agree; (2) Agree; (3) Disagree; (4) Strongly disagree.

   All religions stress that belief normally includes some experience of "union" with the Divine. Are there particular moments when you feel "close" to the Divine? (1) Frequently; (2) Occasionally; (3) Rarely; (4) Never.

4. The final two approaches identified by Johnstone conceive of religiosity in terms of (1) open-ended "ultimate concerns" and (2) intrinsic–extrinsic religious orientation. The ultimate-concern approach attempts to define religiosity in terms of basic religious feelings, without reference to traditional religious forms or institutionalized religious groups. The intrinsic–extrinsic concept distinguishes between extrinsically motivated persons for whom religion represents a self-serving instrumental conformity to social conventions and intrinsically motivated persons for whom religion provides a framework for living and understanding life. Here are two sample items from Joe Feagin's (1964) intrinsic–extrinsic religious orientation scale:

   The church is most important as a place to formulate good social relationships. (1) I definitely disagree; (2) I tend to disagree; (3) I tend to agree; (4) I definitely agree.

   My religious beliefs are what really lie behind my whole approach to life. (1) This is definitely not so; (2) Probably not so; (3) Probably so; (4) Definitely so.

[concept], it is worth measuring in two or more different ways. . . ." (See Box 5.1 for examples of different ways to operationally define the concept of "religiosity.")

Operational Definitions in Social Research

There are two general kinds of operational definitions in social research: manipulated and measured. *Manipulation* operations are designed to change the value of a

variable, whereas *measurement* operations estimate existing values of variables. Both types are nicely illustrated in a study of people's reactions in a common situation—a car failing to proceed immediately when a traffic signal turns green. Anthony Doob and Alan Gross (1968) used this setting to investigate unobtrusively the relations among status, frustration, and aggression. They hypothesized that frustration in traffic caused by a higher-status person would elicit fewer aggressive responses than would frustration caused by a lower-status frustrator. The independent variable, status of the frustrator, was *manipulated* by blocking traffic with two different automobiles: a new luxury automobile for the high-status condition and an older vehicle for the low-status condition. To operationalize the dependent variable, aggression, they used a *measured* definition—number and frequency of horn-honking responses by drivers blocked by the experimental vehicle. The hypothesis was confirmed: For example, 84 percent of the blocked drivers in the low-status condition honked at least once in 12 seconds, compared with only 50 percent of those blocked in the high-status condition.

The operations involved in the manipulation of frustrator's status are spelled out in detail by Doob and Gross. These include such fine detail as the models and ages of the experimental cars, the clothes worn by drivers of these cars, and the amount of time the drivers remained stopped after the signal changed to green. Similarly, precise details are given on the measured operations (observers, stopwatches, and tape recorder) used to determine horn honking. Unless investigators explicitly spell out the details of their manipulations and measurements, other researchers will not be able to replicate or judge the quality of the research.

Manipulation of an independent variable is by definition experimental, and we will have a good deal more to say about this in chapter 7. For now, let us examine some of the various approaches to operationalization of measured variables.

Verbal Reports

By far the most common form of social measurement is the **verbal** or **self-report**. Such reports consist of replies to direct questions, usually posed in interviews or on questionnaires. Self-reports provide simple and generally accurate measures of background variables such as age, gender, marital status, and education. They also are used extensively to measure subjective experiences, such as knowledge, beliefs, attitudes, feelings, and opinions.

All the variables in Beckett Broh's study were operationalized by means of self-reports. She used a single item from the first year of the NELS survey to measure parents' income: What was your total family income from all sources in 1987? (possible responses ranged from 1 = none to 15 = $200,000 or more).

Composite measures

In self-report attitude measurement, responses to several questions frequently are combined to create an **index** or **scale**. As we noted earlier, it is best to use several indicators to measure complex concepts. Scales and indexes (the terms often are

used interchangeably) condense or reduce the data generated by multiple indicators into a single number or scale score. This not only simplifies the analysis but also increases precision and provides a means of assessing the quality of the measurement.

Broh created several composite measures in her study, combining from two to seven questions. For example, she added together the answers to two parts of one question to measure the amount of time students spent on homework each week:

Overall, about how much time do you spend on homework EACH WEEK in and out of school?

	In school each week	Out of school each week
( ) None	00	00
( ) Less than 1 hour	01	01
( ) 1–3 hours	02	02
( ) 4–6 hours	03	03
( ) 7–9 hours	04	04
( ) 10–12 hours	05	05
( ) 13–15 hours	06	06
( ) 16–20 hours	07	07
( ) Over 20 hours	08	08

To operationalize self-esteem, Broh drew on seven items from Morris Rosenberg's (1965) widely used self-esteem scale. Respondents were asked to indicate whether they strongly agree, agree, disagree, or strongly disagree with each of the following seven statements. Values between 1 and 4 were assigned to the response categories, with a 4 representing strong agreement with a positive statement about the self (or, conversely, strong disagreement with a negative statement).

1. I feel good about myself. (Strongly agree = 4)
2. I feel I am a person of worth, the equal of other people. (Strongly agree = 4)
3. I am able to do things as well as most other people. (Strongly agree = 4)
4. On the whole, I am satisfied with myself. (Strongly agree = 4)
5. I feel useless at times. (Strongly disagree = 4)
6. At times I think I am no good at all. (Strongly disagree = 4)
7. I feel I do not have much to be proud of. (Strongly disagree = 4)

An individual's responses to these seven questions were added together to produce a single scale score that could range from 7 (low self-esteem) to 28 (high self-esteem).

As with most scale construction, the development of the Rosenberg self-esteem scale involved some rather sophisticated statistical techniques that are beyond the scope of this book. Because of the difficulty of creating scales and because literally thousands of scales and indexes now exist, we recommend that the beginning researcher use well-validated scales to measure attitudes, opinions, and other subjective experiences whenever possible. Chapter 13 contains a more advanced discussion of the underlying logic of scaling. Also, at various points, we will note how composite measures increase the quality and precision of measurement.

KEY POINT

Many operational definitions (or indicators) are possible, but rarely does any one perfectly represent a concept; therefore, it is usually preferable to use multiple indicators.

Verbal reports vary widely with respect to question wording and response formats. The number of response categories ranges from two to seven for most attitude measures, but the researcher may provide many more categories or none at all. Instead of labeling all response categories, as in the self-esteem scale, investigators may ask respondents to place a check mark along a line whose end points represent opposite ends of a continuum.

Verbal reports also are elicited with pictures and diagrams, which can simplify complex issues and are particularly useful with children. To measure attitudes toward residential integration, for example, Maria Krysan and Reynolds Farley (2002) presented respondents five cards with diagrams depicting neighborhoods containing fourteen homes. Each card showed a different interracial mixture ranging from all-black to all-white with three racially mixed neighborhoods. Respondents "were asked to imagine that they had been looking for a house and found a nice one they could afford." Then they were asked to rank the cards from the most to least attractive and to indicate if there were any of the neighborhoods that they would not want to move into.

Because slight changes in phrasing can drastically alter the meaning of a question, wording is very important. We will discuss question formats and wording at length in chapter 10, which deals with interview and questionnaire construction.

Observation

Doob and Gross measured drivers' levels of aggression by direct observation of behavior: They observed how quickly and how often the drivers honked at a blocking vehicle. Lesley Joy and her colleagues (1986) developed an observational measure of children's aggression to compare a Canadian town before and after it first received television reception in 1974 with two similar "control" towns that had one or several television channels prior to this time. Trained observers recorded grade school children's physically aggressive behavior (hitting, pushing, chasing, etc.) and aggressive words (threatening, arguing, insulting, etc.) on school playgrounds during recess. Summary measures of physical and verbal aggression (average number of acts per minute of playground observation) increased for both sexes following the introduction of television.

Observation provides direct and generally unequivocal evidence of overt behavior, but it also is used to measure subjective experiences such as feelings and attitudes. For example, one could measure interpersonal attraction by observing and recording the physical distance that two people maintain between themselves, with closer distances

indicative of greater liking. Similarly, Zick Rubin (1970) operationally defined "romantic love" by observing the length of time couples spent gazing into one another's eyes.

Besides firsthand observation, hardware such as videotapes, audiotapes, and counters are commonly used for recording purposes. To measure which television programs people watch, the Nielsen Organization attaches an electronic monitoring device to the television set of each person in their sample, which automatically records, minute by minute, the channel to which the set is tuned.

Archival Records

Archival records, which refer to existing recorded information, provide another invaluable source of measurement. The various types of archival data, discussed in chapter 12, include statistical records, public and private documents, and mass communications.

Using public data from the U.S. 1980, 1990, and 2000 censuses, John Logan, Brian Stults, and Reynolds Farley (2004) found modest declines in black–white residential segregation in metropolitan areas during this 20-year period, with declines centered in the West and South. "Residential segregation" was operationally defined by the index of dissimilarity, which indicates the percentage of whites (or blacks) who would have to change their census block of residence to produce zero segregation. Among the 255 metropolises that were analyzed in 2000, the index of dissimilarity ranged from 85 for Detroit (the most segregated) to 20 for Missoula, Montana (the least segregated). This means that in Detroit "either 85% of whites or 85% of blacks would have to move to eliminate residential segregation and thereby reduce the index to zero" (p. 6).

Another example of an operational definition derived from archival records comes from a study testing Riesman's hypothesis that there was a trend in the twentieth century from "inner-" to "other-directedness." Sanford Dornbusch and Lauren Hickman (1959) analyzed advertisements in a popular magazine, *Ladies' Home Journal*, sampled over a 67-year period. An advertisement was operationally defined as "other-directed" if it used endorsements by persons or groups (e.g., "Billie Burke wears Minerva Sweaters," "Housewives like the Singer Sewing Machine") or claimed that use of a product would benefit interpersonal relations ("He'll like you better if you use Revlon"). Beginning about 1920, the authors found a dramatic shift upward in the frequency of other-directed advertising themes, thus supporting Riesman's hypothesis.

Selection of Operational Definitions

Given that you have a concept in mind, how do you decide on an appropriate operational definition? First of all, your decision will be made in the context of an overall research strategy, the choice of which depends to a degree on the specific research question or hypothesis. As we shall see, each of the four basic approaches has its distinctive strengths and limitations, which must be taken into account. Some hypotheses, for example, contain variables that may be impractical or unethical to

manipulate and hence, to study experimentally. No one would propose inducing varying amounts of emotional suffering in people to study effects on physical health.

Each of the different approaches favors certain types of operational definitions. For the most part, survey research involves verbal reports; field research entails observational measurement; and the use of available data, by definition, includes archival records. Experiments, on the other hand, use a combination of measures—verbal reports and/or observation in addition to manipulation procedures. Field researchers often supplement their observations with verbal reports. And survey researchers sometimes use observational measures; for example, an interviewer may observe the type of household and neighborhood in which an interviewee resides as a measure of social class.

With an overall research strategy in mind, the most basic requirement is to select an operational definition that fits the concept well. While we have said that no operational definition can capture a concept's meaning perfectly or completely, this does not license the researcher to select just any measure. It is still desirable to get the best possible fit between concept and measure, and the best way to do this is by carefully considering the meaning of the concept, especially as it relates to the theory in which it is embedded. The National Television Violence Study, for example, developed a coding scheme for measuring television violence based on the theory that television violence is more harmful to the viewer when the depicted perpetrator is physically attractive, portrayed as a hero, and goes unpunished and when the victim exhibits no pain cues or evidence of harm (Kunkel et al., 1995).

Another example of a study in which theory guided the selection of an appropriate operational definition is Albert Pierce's test (1967) of Durkheim's (1951) hypothesis that suicide rates increase in periods of rapid economic change independently of the direction of change (boom or bust). Durkheim theorized that marked economic changes can disturb the existing goals and norms toward which people orient their lives, can thrust individuals into new social settings in which they are ill-suited to manage, and hence can increase the probability of suicide. Data for white males during the peacetime years 1919 to 1940 were examined. At first, Pierce correlated the suicide rate with various *objective* measures of economic change, based on income, percentage of labor force unemployed, and housing construction, with indecisive results. Finally he struck upon the notion of using a measure—the "index of common stock prices"—that would reflect the *public definition* of the economic situation, which is more in tune with Durkheim's theory. That is, rapid fluctuations of stock-market prices may be viewed as indicators of public economic uncertainty, resulting in disruption or discontinuities in perceived goals and norms. Pierce's analysis revealed that suicide rates correlated highly with the rate of change in the public definition of economic conditions as operationally defined by the index of stock-market prices.

Beyond attending to the basic research strategy and the concept's meaning, the choice of operational definitions is largely a matter of creativity, judgment, and practicality. One's selection also should be aided by considering three characteristics that describe the quality of information provided by operational definitions: levels of measurement, reliability, and validity.

Levels of Measurement

Speaking now in terms of variables, we can define *measurement* as "the assignment of numbers or labels to units of analysis to represent variable categories." One characteristic of measurement emerges from the nature of the variable categories. There are different empirical rules for sorting cases into categories, and each rule signifies a different interpretation of the numbers that may be assigned to the categories. Unlike the numbers derived from the measurement of length, time, mass, and so forth in the physical sciences, the numbers applied to social measurement do not always have a simple and straightforward interpretation. The various meanings of these numbers, which reflect basic empirical rules for category assignment, are what is meant by **levels of measurement**. The four general levels usually identified are nominal, ordinal, interval, and ratio.

Nominal Measurement

The lowest level, **nominal measurement**, is a system in which cases are classified into two or more categories on some variable, such as gender, race, religious preference, or political party preference. Numbers (or more accurately, numerals)[1] are assigned to the categories simply as labels or codes for the researcher's convenience in collecting and analyzing data. For example, political party preference might be classified as

1. Democrat
2. Republican
3. Independent
4. Other
5. No preference

Since we are merely using numbers as labels, no mathematical relationships are possible at the nominal level. We cannot say that $1 + 2 = 3$ (Democrat plus Republican equals Independent) or that $1 < 2$ (Democrats are "lower" on political preference than Republicans). We can say, however, that all 1s share the same political preference and that 1s differ from 2s in their political preference.

With nominal measurement, the empirical rule for assigning units to categories is that cases placed in the same category must be equivalent. Also, the categories of the variable should possess two characteristics: They must be both exhaustive and mutually exclusive. To be **exhaustive** means that there must be sufficient categories so that virtually all persons, events, or objects being classified will fit into one of the categories. The following set of categories for the variable "religious preference" does not meet this criterion:

1. Protestant
2. Catholic

You can probably think of other categories needed to make this measure exhaustive, especially if you happen to have no religious preference or if you are Jewish or Muslim or a member of some other religion. Even if one expected few non-Catholic or non-Protestant respondents, one should at least add the categories "None" and "Other" to cover all the possibilities.

The criterion of **mutual exclusivity** means that the persons or things being classified must not fit into more than one category. Suppose that a researcher hastily came up with the following categories for the variable "place of residence":

1. Urban
2. Suburban
3. Rural
4. Farm

You can see that some persons would fit into both categories 3 and 4. The following set of categories would be an improvement:

1. Urban
2. Suburban
3. Rural, farm
4. Rural, nonfarm

Ordinal Measurement

In **ordinal measurement**, numbers indicate the rank order of cases on some variable. Psychologist S. S. Stevens (1951), who developed the idea of measurement level, used hardness of minerals as one example of ordinal measurement. We can determine the hardness of any two minerals by scratching one against the other: Harder stones scratch softer ones. By this means, we could number a set of stones, say five, from 1 to 5 according to their hardness. The numbers thus assigned, however, would represent nothing more than the order of the stones along a continuum of hardness: "1" is harder than "2," "2" is harder than "3," and so on. We could not infer from the numbering any absolute quantity, nor could we infer that the intervals between numbers are equal; in other words, we could not say how much harder one stone is than another.

Another example of ordinal measurement would be an individual's ranking of certain leisure activities in terms of the pleasure derived from them. Suppose you ranked three activities as follows:

1. Playing tennis
2. Watching television
3. Reading sociology

From this ordering we could not make any statements about the intervals between the numbers; it may be that you enjoy watching television almost as much as playing tennis but that reading sociology is not nearly as pleasurable as watching TV. In this sense, ordinal measurement is like a very elastic tape measure that can

be stretched unevenly; the "numbers" on the tape measure are in proper order, but the distances between them are distorted.

One virtue of ordinal measurement, as Julian Simon (1978:231) notes, is "that people can often make an accurate judgment about one thing *compared to another*, even when they cannot make an accurate *absolute* judgment." Simon (1978:231) illustrates the accuracy of comparative judgments with a familiar example:

> [Y]ou can often tell whether or not a child has a fever—that is, when the child's temperature is two or three degrees above normal—by touching the child's face to yours and comparing whether her skin is warmer than yours. But you would be hard-put to say whether the temperature outside is 40° or 55°F, or whether a piece of metal is 130° or 150°F.

Similarly, in the realm of social measurement one can probably say with some certitude whether security or chance for advancement is the more important job characteristic, without being able to say just how important either characteristic is.

The ability of human observers to make such comparative judgments permits a wide range of reasonably accurate social measurements at the ordinal level—for example, measures of socioeconomic status, intelligence, political liberalism, various preference ratings, and attitude and opinion scales. On the other hand, ordinal measurement is still rather crude. At this level we cannot perform most mathematical (statistical) operations in analyzing the data. We cannot add, subtract, multiply, or divide; we can only rank things: $1 < 2$, $2 < 3$, $1 < 3$.

Interval Measurement

Interval measurement has the qualities of the nominal and ordinal levels plus the requirement that equal distances or intervals between "numbers" represent equal distances in the variable being measured. An example is the Fahrenheit temperature scale: The difference between 20°F and 30°F is the same as the difference between 90°F and 100°F—10 degrees. We can infer not only that 100°F is hotter than 90°F but also how much hotter it is. What enables us to make this inference is the establishment of a standard measurement unit, or *metric*. For Fahrenheit temperature, the metric is degrees; similarly, time is measured in seconds, length in feet or meters, and income in dollars.

When numbers represent a metric, the measurement is "quantitative" in the ordinary sense of the word. Thus, we can perform basic mathematical operations such as addition and subtraction. However, we cannot multiply or divide at the interval level. We cannot say, for example, that 100°F is twice as hot as 50°F or that 20°F is one-half as hot as 40°F. The reason is that interval measures do not have a true or absolute zero but an arbitrary one. That is, the zero point on the scale does not signify the absence of the property being measured. Zero degrees Fahrenheit does not mean that there is no temperature; it is simply an arbitrary point on the scale. Its arbitrariness is illustrated by comparison with another interval scale designed to measure the property of temperature: 0°F equals about −18°C (Celsius or centigrade), and 0°C equals 32°F.

While social researchers often aim to create interval measures, most of what passes for this level of measurement is only a very rough approximation. IQ score, for example, is sometimes treated as an interval-level measure, even though it makes no sense to add IQ scores or to infer that equal numerical intervals have the same meaning. (Is the difference between IQ scores of 180 and 190 equal to the difference between 90 and 100?) The empirical rule defining interval measurement is to create equal intervals between numbers. Some attitude scaling techniques attempt to do this, although most scales (e.g., Rosenberg's self-esteem scale) have ordinal-level measurement.

Ratio Measurement

The fourth level, called **ratio measurement**, includes the features of the other levels plus an absolute (nonarbitrary) zero point. The presence of an absolute zero makes it possible to multiply and divide scale numbers meaningfully and thereby form ratios. The variable income, measured in dollars, has this property. Given incomes of $20,000 and $40,000, we can divide one into the other (i.e., form a ratio) to signify that one is twice (or one-half) as much as the other.

Many measures in social research have a well-defined metric and a zero point that meaningfully signifies none of the property being measured. Besides income, other examples are age in years, number of siblings, and years of employment. Ratio-level measures often are obtained by simply counting—for example, number of horn honks, number of siblings, number of people in a social network. Also, aggregate variables, which characterize collectivities of people, frequently are measured at this level by counting and then dividing by a population base—for example, crude birth rate (number of births per 1000 people in the total population), divorce rate (number of divorces per 1000 existing marriages), percentage of labor force unemployed, percentage Democrat.

KEY POINT

Level of measurement indicates the kinds of inferences that can be made when comparing different categories or values of a variable: *nominal* allows one to infer differences, *ordinal* to infer rank order, *interval* to infer distances, and *ratio* to infer ratios.

Discussion

The level of measurement achieved depends upon the empirical procedures used to operationalize a concept. A field researcher, for example, may obtain an ordinal measure of age by using people's appearance, manner, and other observed characteristics to classify them as "children," "young adults," "middle-aged," and "seniors." Or date-of-birth information might be used to measure age as a ratio scale.

It is interesting to note that the four levels of measurement themselves form an ordinal scale with regard to the amount of information they provide. Each level has the features of the level(s) below it plus something else. Table 5.1 illustrates this.

In most social science research, the distinction between interval and ratio levels of measurement is not very important compared with the differences among interval, ordinal, and nominal measures. Many older statistical techniques (including

TABLE 5.1. Information Provided by the Four Levels of Measurement

Information provided	*Nominal*	*Ordinal*	*Interval*	*Ratio*
Classification	X	X	X	X
Rank order		X	X	X
Equal intervals			X	X
Nonarbitrary zero				X

Pearson's correlation coefficient) assume interval measurement and are therefore inappropriate for lower levels of measurement. Some newer techniques are well suited for drawing meaningful inferences about nominal and ordinal measures.[2]

Finally, researchers may use words rather than numbers to represent gradations in conceptual properties (Sorokin, 1937:21–22). If a historian describes the extent of political unrest in various epochs as "low," "high," and "extremely high," an ordinal scale is being used to measure political unrest. If a field researcher classifies the homeless as "mentally ill," "drug-dependent," and "other," the categories represent a nominal scale. The use of words rather than numbers to represent variable categories does not imply poorer or better measurement (King, Keohane, and Verba, 1994:151–52). Whether words or numbers are used, the quality of a measure is judged in terms of reliability and validity, as explained below.

Reliability and Validity

Level of measurement provides a framework for interpreting the categories of a variable. It tells us what sort of inferences we can make about cases assigned to different categories. (Are they merely different? Can we say that one is greater or lesser than the other? And so on.) But what about the adequacy of the set of categories as a whole and of the operationalization procedure for assigning cases to categories? In other words, how does one evaluate the goodness of specific operational definitions?

We have seen that for any concept a large number of operational definitions are possible and that creative insight, good judgment, and relevant theory aid in the development of operational definitions. Admittedly, these aids are rather subjective in nature; however, once an operational definition is selected, there are more objective ways to evaluate its quality. Social scientists use the terms "reliability" and "validity" to describe issues involved in evaluating the quality of operational definitions.

Reliability is concerned with questions of stability and consistency. Is the operational definition measuring "something" consistently and dependably, whatever that "something" may be? Do repeated applications of the operational definition under similar conditions yield consistent results? If the operational definition is formed from a set of responses or items (e.g., scores on the self-esteem scale), are the component responses or items consistent with each other? An example of a highly reliable measuring instrument is a steel tape measure. With such an instrument, a piece of wood 20 inches long will be found to measure, with negligible variation, 20 inches every time a measurement is taken. A cloth tape measure would be somewhat less reliable in that it may vary with humidity and temperature and in that we can expect some variation in measurements depending on how loosely or tightly the tape is stretched.

Measurement **validity** refers to the congruence or "goodness of fit" between an operational definition and the concept it is purported to measure. Does this operational definition truly reflect what the concept means? Are you measuring what you intend to measure with this operational definition? If so, you have a valid measure. An example of a valid measure is amniocentesis, a technique for determining various genetic characteristics of an unborn child, including sex. It is a valid measure of biological sex because it can determine with virtually perfect accuracy whether the unborn will be a boy or a girl. At one time, a number of invalid "measures" of the unborn's sex existed in the form of folk wisdoms. One belief, for example, involves tying a string to the pregnant woman's wedding band and holding the band suspended over her abdomen. If the band swings in a circle, the baby will be a girl; if the band swings back and forth, the child will be a boy.

A highly unreliable measure cannot be valid—how can you measure something accurately if the results fluctuate wildly? But a very reliable measure still may not be valid; that is, you could be measuring very reliably (consistently) something other than what you intended to measure. To take a facetious example, let us suppose we decide to measure the "intelligence" of students by standing them on a bathroom scale and reading the number off the dial (Davis, 1971:14). Such an operational definition would be highly reliable as repeated scale readings would yield consistent results. However, this obviously would not be a valid measure of an individual's intelligence.

KEY POINT

Reliability is synonymous with "consistency"; *validity* is synonymous with "accuracy." A valid measure is necessarily reliable, but a reliable measure may or may not be valid.

Sources of Error

When we apply an operational definition to a set of cases, we use different labels, ranks, ratings, scores, and so forth to represent differences or variation among the cases. There are three potential sources of variation. By examining these, we can better understand the ideas behind reliability and validity assessment. The three sources of variation in any measurement are expressed in the following equation:

$$\text{Observed value} = \text{true value} + \text{systematic error} + \text{random error}$$

The first source of variation is *true differences* in the concept the operation is intended to measure. One would hope that this would account for most of the variation in the measurements; after all, this is what validity is all about! In the ideal situation, with a perfectly valid operational definition, *all* of the measured variation would reflect differences in the concept under study. Observed differences in IQ scores obtained with an IQ test, for example, ought to reflect only true differences

in intelligence and nothing else. However, since perfect measurement is unobtainable, a realistic approach is to be aware of other possible sources of variation and to try to eliminate or reduce their effects as much as possible.

Social scientists refer to sources of variation other than true differences in the variable being studied as "error." The main problem in interpreting the observed variation is to determine what part of it can be explained by true differences and what part is due to one or more sources of error. There are two basic types of measurement errors: systematic and random.

Systematic measurement error results from factors that systematically influence either the process of measurement or the concept being measured. Assuming interval-level measurement, systematic error would be reflected in ratings or scores that are consistently biased in one direction—either too high or too low. A cloth tape measure that has stretched with wear would create error in the form of constant over estimates of length. An example of systematic error in social measurement is the cultural bias of IQ tests. Most IQ tests contain problems and language that tend to favor particular groups in society. Given the same "true" intelligence level, the person familiar with the test problems and language will always score higher than the person who is unfamiliar with the test problems or who speaks a different language from the one in which the test is communicated. Thus, differences in IQ scores may reflect a systematic error introduced by the cultural bias of the test, as well as differences in intelligence.

Many of the systematic errors that contaminate social measurement arise from respondents' reactions to participating in research. When the respondent's sensitivity or responsiveness to a measure is affected by the process of observation or measurement, we refer to this as a **reactive measurement effect** (Webb et al., 2000:12). Just as people behave differently alone than in front of an audience, or with friends than with strangers, they tend to react differently when in a research setting. Awareness of the presence of a social scientist "observer," for example, can increase or decrease the incidence of some observed behaviors.

It has been shown that people are less willing to admit to holding undesirable positions and attitudes when they are aware of being "tested." This is why verbal report measures generally underestimate (a systematic error) the prevalence of socially unacceptable traits, behaviors, and attitudes such as psychiatric symptoms, deviant behaviors, and racial prejudice. A good example is John McConahay's (1986) Old-Fashioned Racism Scale. This measure asks respondents whether they agree or disagree with palpable racist statements: For example, "Blacks are generally not as smart as whites"; "If a black family with about the same income and education as I have moved next door, I would mind it a great deal." In this post–civil-rights era, however, it is no longer socially acceptable to support acts of open discrimination or to endorse blatant racist beliefs about black intelligence, ambition, and other stereotyped characteristics. "This form of racism," as McConahay (1986:93) says, "is as out of style in trendy circles as are wide ties and spats." Consequently, many white Americans, knowing the socially desirable answers, may not express their true feelings when responding to the Old-Fashioned Racism Scale or similar measures. Indeed, it is precisely for this reason that researchers have questioned evidence of a decline in racial prejudice in the late twentieth century. That is, the apparent decline may be a function of social desirability effects rather than of real changes in attitudes (Crosby, Bromley, and Saxe, 1980).

Besides this **social desirability effect**, there are other response tendencies that can introduce systematic measurement error. Respondents are more likely to agree than to disagree with statements irrespective of their content (called the "acquiescence response set"); also, when sequences of questions are asked in a similar format, respondents tend to give stereotyped responses, such as endorsing the right-hand or left-hand response (Webb et al., 2000). Such tendencies produce systematic error when they correspond to variation in the measured property. For example, if a question measuring political liberalism is worded such that agreement indicates a liberal view, then the researcher could not be sure whether a person's agreement indicated a liberal view or simply a tendency to agree with statements regardless of their content.

Systematic errors differ from the true value of a variable by a constant amount. Therefore, they bias measurements in a particular direction, underestimating or overestimating the true value, which affects their accuracy or validity. Because of their constancy, however, systematic errors do not adversely affect reliability. Reliability is undermined by inconsistencies in measurement that arise from random errors.

Random measurement error is unrelated to true differences in the concept being measured. It is the result of temporary, chance factors, such as transitory upswings and downswings in the health and mood of research subjects, temporary variations in the administration or coding of a measure, or momentary investigator fatigue. A tired or bored respondent, for example, may give erroneous responses by not attending carefully to the questions asked. Similarly, an ambiguously worded question will produce random errors by eliciting responses that vary according to respondents' interpretations of the question's meaning.

Such error is random because its presence, extent, and direction are unpredictable from one question to the next or from one respondent to the next. Thus, random errors in a measure of the variable "age" would not be consistently high or low but rather would fall on either side of respondents' real ages so that the average error would be zero. Random error could be demonstrated by asking someone whose visual acuity is impaired (perhaps by drunkenness) to measure the length of an object several times. It is likely that this person's measurements will vary about the object's true length. Sometimes the errors will vary in one direction (overestimating length) and sometimes in the other (underestimating length); sometimes they will be large and sometimes small. Random errors produce imprecise and inaccurate measurements, affecting reliability; however, because they are unsystematic, random errors tend to cancel each other out with repeated measurements. Thus, they do not bias the measure in a particular direction.

KEY POINT

A completely valid measure is free of both systematic and random error; a completely reliable measure is free from random error but may contain systematic error.

In Figure 5.1, a target is used to illustrate the relationship between reliability, validity, and the two sources of measurement error. Measurement is an attempt to hit the bull's eye, which represents the theoretical definition of the concept. A tight pattern, irrespective of its location on the target, reflects a reliable measure because it

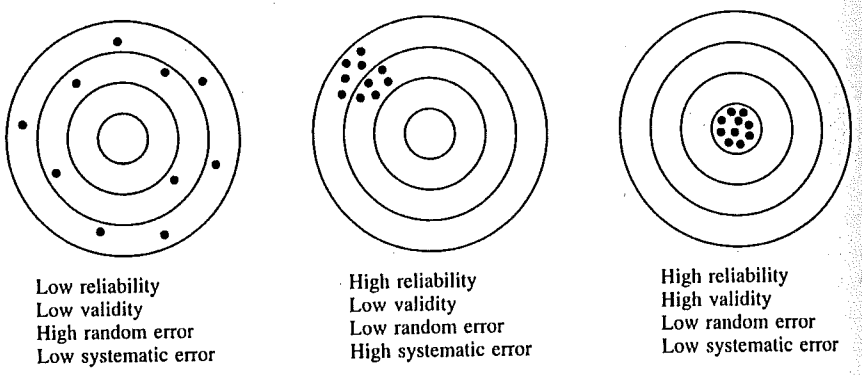


FIGURE 5.1. Analogy of target to reliability, validity, and measurement error. Adapted from Babbie (1995).

is consistent. Validity is a reflection, however, of how closely the shots cluster about the bull's eye. Random error affects the tightness of the pattern as well as distance from the center, hence both reliability and validity; systematic error affects only the distance of shots from the bull's eye, hence only validity.

Reliability Assessment

So far we have said that reliability indicates consistency, or the extent to which a measure does not contain random error. Because we can never know for certain the precise true value of that which we measure, measurement errors can be examined only indirectly. In fact, we infer random error from the degree of consistency observed across measurements. If a measure yields the same result time after time, then it is free of random error; furthermore, the greater the variation in repeated measurements, the greater the random error. Reliability assessment is essentially a matter of checking for such consistency—either over time (as when the same measurements are repeated) or over slightly different but equivalent measures (as when more than one indicator or more than one observer/interviewer/recorder is used).

Test–Retest Reliability

The simplest method for assessing reliability, the **test–retest** procedure, involves testing (i.e., measuring) the same persons or units on two separate occasions.[3] For example, a researcher might administer the self-esteem scale to the same group of students on consecutive days. One then calculates the statistical correlation between the sets of "scores" obtained from the two measurements, and the resulting value serves as an estimate of reliability. Such correlations range from 0 (indicating a completely unreliable measure with total random error) to 1.00 (indicating a perfectly reliable measure subject to no random error). For the test–retest procedure, the correlation tends to be high, with anything less than .80 considered dangerously low for most measurement purposes. Test–retest reliability checks of the self-esteem scale have produced correlation coefficients of .82 to .88 for 1- and 2-week intervals (Rosenberg, 1979; Gray-Little, Williams, and Hancock, 1997); however, test–retest reliability coefficients for this scale are much lower for longer periods of 6 months (.63) and 1 year (.50) (Gray-Little, Williams, and Hancock, 1997).

While simple in principle, the test–retest method has several problems that limit its usefulness as an estimate of reliability. First, either the persons responding to questions or the persons recording observations may remember and simply repeat the responses they gave the first time, thereby inflating the reliability estimate. Second, real change in the concept being measured may occur in the interim between the two "tests." In attitude measurement, new experiences or new information may result in a shift in attitude. For example, positive or negative experiences between administrations may raise or lower a respondent's self-esteem; the loss of a job may change one's attitude toward unemployment insurance or social welfare programs. Because such true changes are inseparable from random errors in test–retest correlations, they falsely lower the reliability estimate. Third, the first application of a measure may itself bring about conceptual changes in the persons under study. Suppose, for example, that a scale designed to measure traditional sex-role attitudes is administered on two occasions to the same group of persons. If, after the first administration, the persons began to think through some of their assumptions about women and as a consequence changed some of their beliefs, a subsequent administration of the scale would yield different scores and a lowered estimate of reliability.

To a certain extent, these problems are manageable. For example, one can try to time the second measurement optimally so that responses to the first testing will have been forgotten but little real change in the concept will have had time to occur. Also, the above difficulties may be more or less problematic depending on the type of measure under study. Using a test–retest reliability estimate for an attitude measure would be fraught with problems because such measures tend to be highly reactive and genuine changes in attitudes are likely to take place over time. On the other hand, the test–retest procedure provides a good evaluation of the reliability of many relatively stable concepts such as characteristics of organizations or political units and individual background variables.

Split-Half and Internal Consistency Reliability

Rather than obtain a *stability* estimate based on consistency over time, as in test–retest reliability, a second set of procedures for assessing reliability estimates the agreement or *equivalence* among the constituent parts or items of a multi-item measure. If we assume that each component of a measure represents the same underlying concept, a lack of agreement among the components would indicate a high degree of random error, hence low reliability. Like the test–retest estimate, all of the statistical estimates of equivalence yield coefficients that run from 0 to 1.00.

A commonly used equivalence estimate is the **split-half method**. In this procedure, a scale or index (i.e., a measure containing several items) is applied once to a sample of cases, after which the items of the scale are divided into halves, usually by random selection; each half is then treated as a subtest with the results of the two subtests correlated to obtain an estimate of reliability. The higher the correlation, the more equivalent the halves and the greater the reliability of the measure.

The split-half technique assumes the existence of equivalent subsets of items. From there it is a short step to the assumption that every item in a scale is equivalent to every other item. This gives rise to another technique for assessing reliability, called **internal consistency**. With this approach, the researcher examines the relationships among all the items simultaneously rather than arbitrarily splitting the items. The basic question is, To what extent are the items homogeneous—that is, to what extent do they measure the same concept? Homogeneity or internal consistency may be estimated via a number of statistical procedures that are beyond the scope of this text. The most common estimate of internal-consistency reliability is Cronbach's alpha, which is based on the average of the correlations among the responses to all possible pairs of items. Numerous studies have reported Cronbach's alpha for the self-esteem scale, ranging from a low of .72 for a sample of men 60 years or older to a high of .88 for a group of college students (Gray-Little, Williams, and Hancock, 1997).

Intercoder Reliability

A second type of equivalence measure examines the extent to which different interviewers, observers, or coders using the same instrument or measure get equivalent results. Assuming that the different users have been properly trained, a reliable operational definition must yield comparable results from user to user. Estimates of equivalence are calculated by comparing the records of two or more researchers who independently apply the same operational definition. For example, in the aforementioned study of other-directedness (Dornbusch and Hickman, 1959), two coders were directed to judge whether magazine advertisements contained some form of other-directed appeal. A reliability check was made by comparing the number of such appeals independently analyzed by each coder in two issues of the magazine. The level of agreement, or **intercoder reliability**, was very high.

KEY POINT

Reliability assessment examines consistency (1) across repeated applications of a measure over time, (2) among the constituent parts or items comprising the measure, or (3) among different observers or coders applying the measure.

It is the norm in social research to check for intercoder reliability whenever measures are derived from systematic observation or from archival records. With regard to the other types of reliability, the split-half and internal consistency techniques are the most frequently used, even though they are limited to multi-item measures. The test–retest approach is less common, not only because of the problems mentioned above but also because of the impracticality of applying the same measure twice to the same sample of cases.

Improving Reliability

At some point in this discussion of reliability assessment you might have wondered, What can be done if your measure turns out to have low reliability? Should you discard it and start over with another measure? In some cases you may decide to do just that. However, there are ways to raise the reliability of an operational definition to an acceptable level.

1. Exploratory studies, preliminary interviews, or pretests of a measure with a small sample of persons similar in characteristics to the target group are ways to gain crucial information about whether the measure is clearly understood and interpreted similarly by respondents. The need for preliminary work with actual respondents before the final form of an instrument is completed cannot be overstated. Indeed, it is a topic we will consider again in relation to experiments and survey research.
2. Simply adding items of the same type to a scale will usually increase reliability. Other things being equal, a composite measure containing more items will normally be more reliable than a composite measure having fewer. There are two reasons for this. First, as we noted earlier, random errors deviate on either side of the "true" value. Thus, with repeated measurements or additional items, such errors will tend to cancel each other out, yielding a more stable and accurate measure of the true value. Second, since any given set of items represents a sample of the possible measures of a concept, adding items increases sample size. As you will see in the next chapter, a basic principle of sampling is that the larger the sample, the more reliable the estimate.
3. An item-by-item analysis will reveal which items discriminate well between units with different values on a particular variable. Those items that do not discriminate appropriately should be omitted. For instance, on a test measuring knowledge of research methods, students with the highest scores should be more likely to answer a given question correctly than students with the lowest scores. If both those scoring high and those scoring low on the test are equally likely to answer the question incorrectly, the question may be ambiguous or misleading. By keeping only those items or questions that correlate highly with the total score, reliability may be greatly improved.
4. Clues for improving reliability may be found in the instructions to respondents. Are they clear, or is there some room for misinterpretation? One should also examine the conditions under which the instrument is used or administered. Are they consistent? Finally, one might question whether the users of the instrument have been adequately and uniformly trained.

Of course, it is important to bear in mind that although a highly unreliable measure cannot be valid, it is possible to have highly reliable but invalid measures. Therefore, unless validity has been demonstrated, caution should be used in drawing conclusions about the "goodness" of even the most reliable measure.

Validity Assessment

Reliability assessment is relatively simple; the major forms that we have outlined use straightforward procedures that yield precise estimates of consistency and random error. These procedures are independent of the theories under investigation; that is, they can be applied and interpreted without regard to what is actually being measured. Validity assessment, by contrast, is more problematic. Systematic errors, which affect validity but not reliability, are more difficult to detect than random errors. And the issue of measurement validity generally cannot be divorced from larger theoretical concerns; sooner or later you must ask what the nature of your concept is, what it means, and whether your operational definition faithfully represents this meaning or something else.

Validity cannot be assessed directly. If it could—if we knew a case's true value on a variable independent of a given measure—then there would be no need for the measure. Or if we had perfectly valid, established measures of concepts, assessing validity would simply be a matter of checking to see if the application of a new operational definition corresponded with the application of an existing one. If we invented a new measure of length, for example, we could easily check its validity by determining whether we get the same results with standard instruments—tape measure, yardstick, transit, and so on—for measuring length. But there are very few established operational definitions of concepts in the social sciences. Therefore, to assess validity, one must either (1) subjectively evaluate whether an operational definition measures what it is intended to or (2) compare the results of the operational definition with the results of other measures with which it should or should not be related. As you will see, the relevant kinds of subjective judgments and objective evidence depend on the purpose of measurement.

Subjective Validation

There are two methods of validity assessment based on subjective evaluation of an operational definition: face validity and content validity. **Face validity** refers simply to a personal judgment that an operational definition appears, on the face of it, to measure the concept it is intended to measure. In some cases this claim alone would seem reasonable to establish a measure's validity. Few would dispute the face validity of common indicators of variables such as age, gender, and education. Many observational measures of behavior have similar palpable validity—for example, "hitting another person" as an indicator of aggression and "offering assistance to a stranger" as an indicator of helping. As a method of validity assessment, however, face validity is generally not acceptable. Most operational definitions have it. After all, why would an instrument be offered as a measure of some concept if it did not appear to be valid? Face validity is based solely on personal judgment rather than objective evidence. Furthermore, it suggests that validity is an all-or-none matter when, in fact, measures have degrees of validity. For example, several operational definitions of age are possible. One could ask respondents directly what their age is, ask for their age at their last birthday, or ask them when they were born. While these all have face validity, they are not equally accurate.[4] One would not know which operationalization is most accurate without resorting to a validation method other than face validity.

Content validity concerns the extent to which a measure adequately represents all facets of a concept. A more acceptable form of subjective evaluation, this type of validation is used most often in psychology and education, where it is applied to measures of skill, knowledge, and achievement. An instructor testing the reader's knowledge of this chapter, for example, ought to be concerned with the test's content validity—that is, with whether it includes questions on all sections of the chapter. Such a test would not have content validity if it omitted questions on reliability and validity and contained only questions on the measurement process and levels of measurement.

Psychologists and educators speak of the performance or content "domain" in relation to content validity. To demonstrate content validity, one must be able to identify clearly the components of the total domain and then show that the test items adequately represent these components. This is not too difficult for most tests of knowledge. With respect to knowledge of the present chapter, one could list all the major topics and subtopics and then develop test items for each of these, making sure that the number of items per topic was proportionate to the breadth of coverage. However, such a process is considerably more complex when measuring the abstract concepts typical of the social sciences. The domain of concepts like social capital, alienation, and social status is not easily specified; therefore, it is difficult to determine how adequately the domain has been tapped by specific indicators.

To some extent, the problems associated with face and content validity are not unique. All forms of validation are subjective in the sense that judgments on the validity of an operational definition ultimately rest with the verdict of the scientific community. However, social scientists generally do not find content validity evidence as persuasive as the kinds of "external" evidence provided by the validation procedures examined in the next two sections. Evidence that is external to the investigator is less subject to unintentional distortion and is easier to verify.

Criterion-Related Validation

Criterion-related validity applies to measuring instruments that have been developed for some practical purpose other than testing hypotheses or advancing scientific knowledge. One may wish to devise measures that will identify children with learning disabilities, determine a person's ability to fly an airplane or drive a car, or predict success in college. Under these circumstances, the investigator is not interested in the content or apparent meaning of the measure but in its usefulness as an indicator of a specific trait or behavior. The trait or behavior is called a *criterion,* and validation is a matter of how well scores on the measure correlate with the criterion of interest. The higher the correlation, the more valid the measure with respect to that criterion. As Jum Nunnally (1970:34) notes, with this form of validation the criterion variable is the only necessary standard of comparison: "If it were found that accuracy in horseshoe pitching correlated with success in college, horseshoe pitching would be a valid measure for predicting success in college."

The selected measure may indicate either an individual's *present* or *future* standing on the criterion variable. For example, a mental health inventory de-

signed to identify those in need of psychiatric care could be given to a sample of "well" persons and a sample of persons currently under psychiatric care. Its validity would be indicated by the degree to which the inventory distinguishes between the present two groups. Or the measure may predict a person's future standing on the criterion variable. A college entrance examination, for example, might be validated by comparing the exam scores of high school students with a criterion measure of their later success in college, such as grade-point averages or whether or not they graduate.

Since criterion-related validity rests on the correspondence between a measure and its criterion, it is only as good as the appropriateness and quality of the criterion measure. Unfortunately, this may present major difficulties. By what standards do you choose the criterion? What if no reasonable criterion exists? What if the criterion exists but practical problems prevent using it? For example, how would you demonstrate the validity of a county civil service test developed to assist in the hiring of probation officers? Logically, you could suggest that the county hire high-, average-, and low-scoring persons; then at a later time you could compare some measure of their job performance, such as supervisors' ratings, with their scores on the civil service test. Most likely, however, the county will hire only the top scorers and, regardless of their performance on the job, you would not know how it might have compared with the job performance of those who scored average or low. Thus, you could not assess the criterion-related validity of the measure.

Despite such problems, evidence of criterion-related validity is crucially important when a test or measure serves a specific, practical end. Thus, if a test is designed to screen and select candidates for certain jobs or to place students in ability tracks or special programs, then it is important to know how well it works for the given purpose. Except for applied areas of psychology and education, however, social science measures are not developed to help solve practical problems of this sort. Operational definitions are created to reflect the meaning of certain concepts, and there is seldom a clear and adequate criterion variable for evaluating validity.

Construct Validation

When neither a pertinent criterion of prediction nor a well-defined domain of content exists for determining validity, investigators turn to construct validation. (The term "construct" is interchangeable with the term "concept"; a concept developed for scientific purposes is sometimes called a "construct.") **Construct validation** emphasizes the meaning of the responses to one's measuring instrument. How are they to be interpreted? Is the instrument measuring the intended concept (or construct), or can it be interpreted as measuring something else? While this sounds like face validity, the orientation is quite different; construct validity is based upon an accumulation of research evidence and not on mere appearances.

According to the logic of construct validation, the meaning of any scientific concept is implied by statements of its theoretical relations to other concepts. Thus, the validation process begins by examining the theory underlying the concept being measured. In light of this theory, one formulates hypotheses about variables that should be related to measures of the concept. At the same time, one

considers other variables that should *not* be related to measures of the concept but might produce systematic error. Then one gathers evidence to test these hypotheses. The more evidence that supports the hypothesized relationships, the greater one's confidence that a particular operational definition is a valid measure of the concept.[5]

An example of construct validation is Morris Rosenberg's validation (1965) of his self-esteem scale. *Self-esteem* refers to an individual's sense of self-respect or self-worth: Those with high self-esteem have self-respect; those with low self-esteem lack it. Rosenberg reasoned that, if his "scale actually did measure self-esteem," then scores on it should "be associated with other data in a theoretically meaningful way" (p. 18). Thus, because clinical observations indicated that depression and neurosis often accompany low self-esteem, people who score low on the scale should report more depressive feelings such as unhappiness and gloom, appear more depressed to outside observers, and reveal more symptoms of neurosis. Also, given the sociological proposition that an individual's self-esteem is determined largely by what others think of him or her, students with high self-esteem scores should be chosen more often as leaders by classmates and described more often as commanding the respect of others. Evidence confirmed these and other theoretical expectations, thereby supporting the construct validity of the self-esteem scale.

Table 5.2 depicts the differences between construct validation and criterion-related validation. The test of criterion-related validity is the ability of the measure to classify, group, or distinguish persons (or other units of analysis) in terms of a single criterion. What the measure means aside from its ability to make such distinctions is of little concern. What is important is the strength of the correlation between the predictive measure and a measure of the criterion; the stronger the correlation, the higher the validity.

In construct validation, on the other hand, one is less interested in the accuracy of a prediction per se than in what a relationship reveals about the meaning of the concepts being measured. One does not necessarily expect the correlation between measures of two theoretically related concepts to be extremely high because the two concepts do not mean exactly the same thing. That is why it is important to exam-

TABLE 5.2. Comparison of Criterion-Related and Construct Validity

	Criterion-related validity	*Construct validity*
Validity a matter of:	ability to classify units with precision	ability to capture the meaning of the concept
The measure is used for:	practical application	theoretical application
Assessment a matter of comparison with:	a single, agreed-upon independent criterion	no clear, single criterion
Degree of validity a matter of:	predictive accuracy	a consistent pattern of relationships

ine the relation of the measure in question to several other variables rather than base the assessment of validity on a single prediction. All the tested relations contribute to the evaluation of construct validity; it is their cumulative effect that supports or disputes the validity of the measure.

Evidence of construct validity consists of any empirical data that support the claim that a given operational definition measures a certain concept. Because such evidence may be derived from a wide variety of sources, construct validation is not associated with a particular approach or type of evidence. We will now consider four of the more common types of evidence used to establish construct validity. Remember, though, that no single study or piece of evidence is sufficient; the construct validity of a concept is only as compelling as the amount and diversity of evidence supporting it.

1. *Correlations with related variables.* If a measure is valid, it should correlate with measures of other theoretically related variables. Rosenberg (1965) validated his self-esteem scale in this way by showing that scores on it were correlated with symptoms of depression and neurosis and with peer ratings.

2. *Consistency across indicators and different methods of measurement.* Different measures of the same concept should be correlated, and because each methodological approach (e.g., self-reports, observation, archival records) is subject to different sources of systematic error, measures of concepts should not be tied to a particular method. Thus, one of the most convincing signs of construct validity is the correspondence of results when a concept is measured in different ways. This is called **convergent validity** because the results converge on the same meaning, namely, that conveyed by the underlying concept.

Charles Glock and Rodney Stark (1966) provided such evidence for their Index of Anti-Semitic Beliefs. Scores on the index were determined by level of agreement with six statements about Jews (e.g., "Jews believe they are better than other people," "Jews are more likely than Christians to cheat in business"); hence, respondents were forced to agree or disagree with characteristics imputed by the investigators. To measure anti-Semitism in a different way, Glock and Stark had respondents furnish their own images of Jews by requiring them to complete the following sentences: "It's a shame that Jews . . ." and "I can't understand why Jews . . ." When their answers were divided into those indicating negative, neutral, or positive images of Jews, it was found that 73 percent of those scoring high on the Beliefs index as opposed to 27 percent of the low scorers projected a negative image.

3. *Correlations with unrelated variables.* A valid operational definition should separate the concept being measured from other concepts from which it is intended to differ. In other words, there are some variables (representing systematic errors) with which a measure should not be highly correlated. This is called **discriminant validity**.

A good example is Zick Rubin's validation (1970) of his thirteen-item Love Scale. A valid measure of love should differentiate love from liking as these two concepts are empirically related but conceptually distinct. Rubin therefore developed a parallel scale of liking. When both the love and liking scales were administered, he found that their scores were only moderately correlated. Also, whereas respondents liked their dating partners only slightly more than they liked their friends, they loved their dating partners much more than their friends. Additional evidence showed that the Love Scale tapped an attitude toward a specific other person rather than a general response tendency. For example, Love Scale scores were uncorrelated with scores on the Marlowe–Crowne Social Desirability Scale, designed to measure the tendency to give socially desirable responses.

4. *Differences among known groups.* When certain groups are expected to differ on the measure of a concept, one source of validating evidence would be a comparison of the groups' responses. Richard Contrada and colleagues (2001) used this approach to test the validity of a measure of perceived ethnic discrimination. Respondents were instructed to indicate on a seven-point scale ranging from 1 (never) to 7 (very often), how often over the past 3 months particular forms of ethnic discrimination had been directed at them. One item asked, "How often have others implied that you must be dishonest because of your ethnicity?" Another asked, "How often have others had low expectations for you because of your ethnicity?" When this measure was administered to students at Rutgers University, non-whites, as expected, were significantly more likely to report ethnic discrimination. (See Box 5.2 for an example of construct validation that uses a variety of evidence.)

Box 5.2 Validation of the Attitudes toward Feminism Scale

At the height of the Women's Liberation Movement in the mid-1970s, Eliot Smith, Myra Marx Ferree, and Frederick Miller (1975) developed a twenty-item scale to measure attitudes toward feminism. Each scale item consists of a profeminist or antifeminist statement; for example, "Women have the right to compete with men in every sphere of activity," and "A woman who refuses to bear children has failed in her duty to her husband." Respondents indicate their level of agreement with each statement, numbers are assigned to the answer categories, and then scale scores are calculated by adding up responses to the twenty items. The numerical range of the answer categories is 1 to 5, with the highest number representing a profeminist response; so the scale has a possible range of 20 (extreme antifeminism) to 100 (extreme profeminism). (See Box 13.1 for a description of how the authors created this scale.)

There is unusually good evidence validating this measure, called the FEM Scale (an acronym based on the authors' first names). The creators of the FEM Scale tested its reliability as well as its construct validity with data from 100 Harvard summer school students. Royce Singleton and John Christiansen (1977) also assessed the scale's validity with data from a larger, more heterogeneous sample of respondents. Following is a summary of the evidence regarding the scale's reliability and validity.

Reliability

Both of these studies obtained a reliability estimate, based on internal consistency, of .91, which is quite acceptable for attitude measurement.

Intercorrelations and Convergent Validity

Smith, Ferree, and Miller tested the validity of the FEM Scale by correlating it with three other variables: measures of identification with the Women's Movement, activism in the movement, and the Rubin–Peplau Just World Scale. High scores on the latter scale reflect a belief that the world is a just place and that people generally get what they deserve. Smith, Ferree, and Miller reasoned that feminists are unlikely to view the world in general as a just place since they tend to perceive that women have been treated unjustly. Consistent with expectations, the results showed that FEM Scale scores were positively correlated with the identification and activism measures (evidence of convergent validity) and negatively correlated with scores on the Just World Scale.

Singleton and Christiansen correlated scores on the FEM Scale with measures of dogmatism, antiblack prejudice, and identification with the Women's Movement. They reasoned that dogmatism, a tendency to adopt traditional views, would include an antifeminist orientation. Also, prejudice toward blacks would reflect a general tendency toward prejudice with which antifeminist attitudes would be consistent. The correlations of individuals' scores on the FEM Scale with their scores on these other measures were moderately high, as expected.

Discriminant Validity

In addition to the above measures, Smith, Ferree, and Miller administered the Rotter I–E Scale to their respondents. "I–E" stands for internal–external locus of control. High "internals" are persons who believe that they have a great deal of personal control over their lives; high "externals" are persons who tend to believe that their lives are controlled by forces outside themselves. Because there was reason to expect that feminism would be mildly correlated with both of these tendencies, Smith, Ferree, and Miller expected no correlation with the FEM Scale when internality and externality were treated as a single scale. Once again, the results supported their prediction.

Known-Groups Validity

Singleton and Christiansen also administered the FEM Scale to members of two groups with opposing views on women's issues: the National Organization for Women (NOW) and Fascinating Womanhood. NOW was the largest and most prominent organization in the Women's Movement; Fascinating Womanhood was an antifeminist organization of the early to mid-1970s that strongly advocated a traditional, dependent role for women. FEM Scale scores of the members of these two organizations were widely divergent, as expected. Recall that the scale has a range of 20 to 100. NOW members had an average score of 91, whereas members of Fascinating Womanhood had an average score of 51

While construct validation is now considered the model validation procedure for most social measurement, it is not without its problems. Obviously, it is cumbersome and requires abundant evidence; more important, however, it can lead to inconsistent and equivocal outcomes. If a prediction is not supported, this may mean that the measure lacks construct validity. On the other hand, such negative evidence may mean that the underlying theory is in error or that measures of other variables in the analysis lack validity. Only if one's theoretical predictions are sound and the measures of other variables are well validated can one confidently conclude that negative evidence is due to lack of construct validity (Zeller and Carmines, 1980).

KEY POINT

Objective validity assessment examines the degree to which the operational definition is correlated with a current or future criterion (*criterion-related validity*) or forms expected relationships with other measures and variables (*construct validity*).

A Final Note on Reliability and Validity

The procedures for assessing validity and reliability may seem so complex and cumbersome that one may wonder if investigators ever pass beyond this stage of research.[6] Fortunately, many investigators avoid the issue by borrowing, from previous studies, measures that have established records of validity and reliability. Also, few researchers apply more than one of the simpler procedures to ascertain the reliability or validity of their new measures. What usually happens is that once a new measure of some sort is introduced (perhaps with minimal evidence of reliability or validity), the measure is widely used for awhile until invalidating features are found (such as the middle-class bias of IQ tests) and the instrument is revised or replaced. Thus, validity and reliability assessment is not confined to intrastudy or intrainvestigator efforts but is an ongoing process that extends across studies and investigators and over a considerable length of time.

It is also true that the most elaborate procedures were developed in response to the difficulties of operationalizing some concepts, such as attitudes, that are relatively unstable and pose reactivity and other measurement problems. More stable and less reactive measures are not as problematic.

In speaking about the issues validity raises, James Davis (1971:14–15) aptly observes that "[a]t the extreme [validation] constitutes a philosophical thicket which makes a dandy hiding place from which antiempirical social scientists can ambush the simple-minded folk who want to find out what the world is like rather than speculate about it." Although it is true that such difficulties can immobilize diehard perfectionists, for most social scientists the validation problem presents challenging opportunities to exercise their creativity. No measure is perfect, but an imperfect measure is better than none at all. Or, as Davis further notes, "Weak measures are to be preferred to brilliant speculations as a source of empirical information" (p. 15).

Summary

Measurement is the process of assigning numerals or labels to units of analysis in order to represent conceptual properties. This process involves *conceptualization* (the development and clarification of concepts) followed by *operationalization* (the description of the research procedures necessary to assign units to variable categories to represent concepts). Because single indicators do not correspond perfectly to concepts, researchers often use multiple indicators, which may be combined to create an index or scale. Operational definitions may be formed either by experimentally manipulating a variable or through nonmanipulative procedures such as verbal reports, observations of behavior, and archival records. One selects operational definitions in the context of an overall research strategy with an eye toward obtaining the best possible fit with the concept being measured.

Operational definitions are described in terms of their level of measurement and are evaluated with respect to their reliability and validity. Measurement level alerts us to the various ways that we can interpret the numerals assigned to different variable categories. In nominal measurement, the numbers are simply labels that signify differences of kind. Variables measured at this level form a classification system that should be exhaustive and mutually exclusive. In ordinal measurement, different numbers indicate the rank order of cases on some variable. In interval measurement the numbers form a metric so that different numbers imply not only rank order but also countable distances. Ratio measurement contains these features plus an absolute zero point, making it possible to form ratios of the numbers assigned to categories.

Reliability refers to the stability or consistency of an operational definition, whereas *validity* refers to the goodness of fit between an operational definition and the concept it is purported to measure. A valid measure is necessarily reliable, but a reliable measure may or may not be valid. One can see this in relation to the three sources of variation in all measures: true differences and systematic and random measurement errors. A completely valid measure reflects only true differences, which means that it is free of both systematic and random error. A completely reliable measure is free from random error but may reflect true differences and/or systematic error.

We assess reliability by calculating the correlation between (1) repeated applications of the measure (test–retest reliability) and (2) responses to subsets of items from the same measure (split-half reliability); by (3) examining the consistency of responses across all items (internal consistency); or by (4) observing the correspondence among different interviewers, observers, or coders applying the same measure (intercoder reliability). Of these approaches, the first is used least often. One can increase reliability by adding items to a scale, by omitting items that do not discriminate well, and by clarifying questions and instructions contained in one's operational definition.

We assess validity by subjectively evaluating an operational definition, by checking the correspondence between the operational definition and a specific criterion, or by determining whether the operational definition of a given construct correlates in expected ways with measures of several other constructs. Subjective validation involves judgments of either whether an operational definition appears to be valid (face validity) or whether it adequately represents the domain of a concept (content validity). Content validity is important in educational testing, but otherwise subjective judgment per se is unacceptable as a source of validation. Criterion-related validation applies to measures (or "tests") that are intended to indicate a person's present or future standing on a specific behavioral criterion. It is especially important to assess when a measure is a practical, decision-making tool. Construct validation is based on an accumulation of research evidence. This evidence may include differences among groups known to differ on the characteristic being measured and correlations with related variables, with different measures of the same concept (convergent validity), and with measures from which the concept should be differentiated (discriminant validity).

Key Terms

- conceptualization
- operationalization
- indicator
- verbal (self-) report
- index/scale
- level of measurement
  - nominal
  - ordinal
  - interval
  - ratio
- exhaustive
- mutually exclusive
- reliability
- validity
- random measurement error
- systematic measurement error
- reactive measurement effect
- social desirability effect
- test–retest reliability
- split-half reliability
- internal consistency reliability
- intercoder reliability
- subjective validation
  - face validity
  - content validity
- criterion-related validation
- construct validation
  - convergent validity
  - discriminant validity

Exercises

1. Look up an article in a social science journal that reports an empirical study that investigates one of the concepts listed below, as either an independent variable or a dependent variable. Do the researchers provide a *theoretical definition* of the concept? What is the researchers' *operational definition*? What is the source of their operational definition (self-report, observation, or records)?

alienation	liberalism/conservatism
fear of crime	political participation
group cohesion (or cohesiveness)	racial prejudice

2. In chapter 1 we introduced the topic of altruism. "Altruism" refers to helping behavior that is motivated purely by a desire to benefit others without anticipation of personal rewards. Helping behavior provides some benefit to or improves

the well-being of another person. Assuming you were going to conduct a campus survey, how would you operationalize altruistic behavior? Give examples of at least two indicators.

3. Indicate the level of measurement of each of the following variables.
   a. *Seriousness of criminal offense*: measured by having judges rank offenses from the most to the least severe
   b. *Political activism*: measured by the total number of politically related activities in which an individual participates
   c. *Ethnic group membership*: measured by having respondents check one of these categories: black, Hispanic, Oriental, Caucasian, other
   d. *Educational attainment*: measured by asking respondents to check one of the following categories: 8th grade or less; 9–11 years; high school graduate; some college; college graduate
   e. An item measuring an attitude or opinion that uses the following response format: strongly agree, agree, undecided, disagree, strongly disagree

4. Suppose that members of a class in research methods were given an examination on this chapter. Assuming that the examination is designed to measure students' knowledge of the material in the chapter, describe three possible sources of measurement error (either systematic error or random error) in the set of examination scores. For each source of error, explain whether it is likely to affect measurement validity, reliability, or both.

5. One problem with many studies of domestic violence, including child abuse, is that they have relied upon self-reports. Because domestic abuse is socially stigmatized and can result in criminal charges, individuals may be tempted to underreport abusive conduct. What kind of measurement error (random or systematic) does this introduce? What effect would this have on the reliability and/or validity of such self-report measures?

6. Below are three of the items you are considering for your campus survey of altruism (see question 2). First, using your personal judgment and keeping in mind that all measures are subject to some degree of error, evaluate the validity of each item; that is, indicate why you believe the item is likely to be a valid or invalid measure of altruism. Second, let's assume you create a composite measure of altruism in which you ask respondents whether they engaged in several different altruistic actions during the past year (besides the possible items below, other items might include giving money to charity, helping carry a stranger's belongings, and giving food or money to a homeless person). How would you assess the reliability and validity of your composite measure?
   a. Have you ever donated blood?
   b. Have you ever helped someone pack and move?
   c. Have you ever been a "big brother" or "big sister" to a child who was not a sibling?

Notes

1. At times we will use the word "number" where technically the term "numeral" is more accurate. The difference, simply put, is that numbers are abstract concepts (such as the number 2), whereas numerals are the squiggly lines used to represent the concepts. In the nominal level of measurement, numerals do *not* represent the number concept but rather are arbitrarily assigned to categories for coding purposes.

2. Choosing an appropriate statistical technique, however, is not always easy as an actual measure may fall between two levels of measurement. One might think of educational attainment (years of schooling), for example, as providing more information than an ordinal scale but not attaining the interval-scale level.

3. As the terminology suggests, most of the procedures for determining reliability were developed in connection with psychological testing.

4. Studies have shown that if a single question is used, it is best to ask respondents in what year they were born. However, the way to get the most accurate measure is to ask both date of birth and age at last birthday, check one against the other, and then inquire about any discrepancies (Sudman and Bradburn, 1982).

5. As you can see, this process is quite similar to general hypothesis testing. In fact, construct validation "is not simply a matter of validating a [measure]. One must try to validate the theory behind the [measure]" (Kerlinger, 1973:461).

6. The foregoing discussion of reliability and validity does not exhaust the use of these concepts in the social sciences, for in addition to judging the adequacy of operational definitions, the terms are applied in the evaluation of other aspects of a research study. For example, the term "validity" is used in reference to the adequacy of a research design (chapters 7 and 8). The concept of "reliability" is used frequently in judging the quality of a sample (chapter 6).

6

Sampling

Sampling, whether we are aware of it or not, is part of everyday life. For instance, after eating at one of a chain of restaurants, a person may decide that all the restaurants in the chain serve poorly prepared food and provide poor service. A student's decision to take a given course may be based upon the opinions of friends who have already taken it. Having met a few people from New York, someone may conclude that New Yorkers are pushy and aggressive. In each of these examples, inferences about a whole class of objects are made from observations of a subset of such objects. The examples implicitly contain the basic idea behind sampling:

1. We seek knowledge or information about a whole class of similar objects or events (usually called a **population**).
2. We observe some of these (called a **sample**).
3. We extend our findings to the entire class (Stephan and McCarthy, 1958:22).

While simple in principle, sampling can be fraught with difficulties in practice. Consider some possibilities. The restaurant patron may have picked a day when the head cook had quit and had been replaced by the busboy, or she may have patronized the worst franchise in the chain. The student talked to friends, who are likely to have highly similar opinions that may differ markedly from the opinions of other students who have taken the course. Finally, characterizations of groups of people, like the one of New Yorkers, seldom apply to even a majority of the group. Generalizations such as these, drawn from casual observation, are likely to err because they are based on inadequate samples of information. In this chapter we discuss some rigorous yet surprisingly simple techniques that are designed to reduce the hazards of generalizing from incomplete information.

Chapter 5 dealt with the measurement of variables. Before actual measurements are taken, researchers must select an appropriate set of units (or cases). Selecting a set of cases raises several crucial questions. The first of these was addressed in chapter 4: What is the unit of analysis—individual people, married couples, newspaper editorials, cities? The answer to this question, recall, hinges largely on the research topic, which determines what the researcher seeks to describe and compare. Thus, research on the relationship between a nation's literacy rate and its economic development would use nations as units; a study of the influence of years of schooling on prejudice would concern individuals. The decision about *what type of unit* to examine will not be considered further in the present chapter; we will assume that the appropriate type of unit has been determined. The

interests here will be with *how many* such units of *what particular description* should be *chosen by what method.* But before examining each of these issues, let us take a brief look at some of the reasons that sampling in social research, as in everyday life, is a practical necessity for informed decisions.

Why Sample?

Scientists seek to establish the broadest possible generalizations, applicable to infinitely large classes of events. Yet, it is obviously impossible to observe all relevant events. In the physical sciences, the nature of the elements studied simplifies this problem. Particular physical and chemical elements are assumed to be identical or nearly identical with respect to pertinent properties. Thus, any sample of one or more cases suffices; one test tube of hydrogen, distilled water, or nitroglycerine will be like any other. As we move from the physical to the biological and social sciences, however, this assumption of homogeneity becomes increasingly risky. Because people and other social objects vary widely on nearly every imaginable property, studying any one case simply will not suffice as a basis for generalizing. The heterogeneity of cases requires careful procedures designed to ensure that the range of variation in the population will be represented adequately in one's sample of observations.

Beyond the need to represent population variability adequately, practical considerations usually necessitate sampling. Researchers often want to know something about a specific social group or population that, for reasons of size, time, cost, or inaccessibility, cannot be studied in its entirety. Indeed, unless the researcher is willing to restrict his or her inferences to a very small set of narrowly defined cases or unless one has unlimited time and funds, there is no getting around sampling.

To interview the entire 20,000-student population at a university, a population that is relatively small by survey research standards, would require a long period of time or a large number of interviewers. However, time is often of the essence. A lengthy period of data collection would render some data, such as attitudes about current issues or voter preferences, obsolete by the time the information was completely in hand. Moreover, the responses of persons interviewed late in the investigation could not be considered comparable to those of persons interviewed early. Responses could be affected by current events, or rumors and scuttlebutt about the ongoing survey may influence the responses of persons interviewed late. A large staff of interviewers also would present problems. Not only would this add to the total costs of the study, but it would be more difficult to train and supervise interviewers adequately. To employ a large number of interviewers, one may also be forced to hire some who are not well qualified, thereby reducing the accuracy of the data.

This last point brings us to a third reason for sampling. Paradoxically, the attempt to observe all cases may actually describe a population less accurately than a carefully selected sample of observations, especially in survey studies of large populations. The reason is that the planning and logistics of observation are more manageable with a sample. Greater attention can be given to the design of interviews or

questionnaires, to procedures for locating difficult-to-find respondents, and, once again, to the hiring, training, and supervision of a competent staff of interviewers, all of which will increase the quality of the data collected.

A case in point is the decennial Census, a complete enumeration of the U.S. population made every 10 years. The Census is an enormous undertaking; the 2000 Census, which counted more than 280 million people, cost about $6.5 billion to collect, tabulate, and publish; and the Bureau of the Census hired more than 500,000 interviewers just to track down those who did not return their census forms in the mail (General Accounting Office, 2001a, 2001b). Yet, despite its efforts, the Census Bureau estimated a net undercount of over 3 million people; an estimated 7.6 million people may have been missed and 4.3 million more may have been counted twice (Schmitt, 2001). As James Gleick (1990) said about the 1990 Census, there were simply "too many doors to knock on, and too many people living without doors . . . [and too high] a rate of noncooperation." These obstacles are much easier to overcome in a sample. That is why data from some sample surveys, such as the 50,000-household Current Population Survey, are considered more accurate for many purposes than the Census. A post enumeration sample survey is used to estimate the accuracy of the Census and to make statistical adjustments to the count. And to increase accuracy and reduce costs, many statisticians recommend that the Census be replaced with a survey of a large sample.[1]

We hasten to add that only in the case of extremely large populations would we expect a sample to yield more accurate information about a population than a survey of the whole population. But the justification for sampling need not rest on a claim of greater accuracy. More to the point, carefully selected samples are an efficient way of producing accurate information. You can see this for yourself on the next election day. Note the accuracy with which the major news networks predict certain election outcomes, long before all the votes have been tabulated. This accuracy is made possible by the careful sampling procedures of expert pollsters. (The polls have not always been so accurate, however, as revealed in Box 6.1.)

BOX 6.1 Sources of Errors in Survey Sampling: Two Notorious Failures in Political Forecasting

Every 4 years in the United States, the accuracy of public opinion polls is put to the test. Why? Because unlike most surveys, in which there are no external criteria for assessing the validity of sample estimates, pre-election poll forecasts can be judged against the actual outcome of the election (Martin, Traugott, and Kennedy, 2005). Based on samples, polls cannot be expected to predict an outcome perfectly; but given random sampling, they should be within an expected margin of sampling error. Taking this into account, social scientists have devised various methods for assessing poll accuracy (see Martin, Traugott, and Kennedy, 2005). These analyses indicate that most polls do rather well, with estimates off by no more than the expected margin of error—a few percentage points. In 2008, nineteen final polls, all conducted within a few days of the election, had Barack Obama winning the presidential election by 5 to 11 percentage points; Obama's actual margin of victory was 7 points. Election polls, however, have not always been so accurate, as two infamous examples from the annals of sample surveys reveal.

Perhaps the biggest polling debacle occurred in the presidential election of 1936, when *Literary Digest* magazine predicted that Alfred Landon would be a landslide winner, by 57 to 43 percent, over Franklin Roosevelt. At this time, the *Literary Digest* was well known for its polls on public issues and enjoyed considerable prestige, having predicted the winner in every presidential election since 1920. Its 1932 poll erred by less than 2 percentage points. However, in 1936 it missed calling the election by a wide margin. What went wrong? With over 2 million individuals polled, the *Digest*'s sample was certainly large enough. Indeed, it was enormous by current standards. In 2008, a sample of only 2000 to 3000 people was required to predict accurately the voting behavior of over 130 million voters.

There were two major reasons for the failure of the *Literary Digest* poll. The first was the manner in which the sample was selected, which produced *coverage error* (see p. 185). The *Digest* conducted its polls by mailing ballots to millions of people. The names and addresses of these people came from a variety of sources, including phone books, automobile registration lists, and the *Digest*'s own subscription list. Because such sources tended to exclude the poor, who could not afford telephones or automobiles, the sample did not provide a representative cross section of American voters. With rich and poor voting similarly before 1936, this bias against the poor apparently had little effect on the predictions. But in 1936, the rich tended to vote for Landon, the poor voted overwhelmingly for Roosevelt, and the bias inherent in the *Digest*'s initial sample contributed to the error.

A second problem with the *Digest* poll has to do with what is called *nonresponse bias* (see p. 185). With any survey, there will be a certain percentage of the selected sample who do not respond or who cannot be contacted. Studies have shown that these nonrespondents tend to differ in important ways from respondents. For example, both lower-class and upper-class people are less likely to respond to questionnaires or personal interviews than middle-class people. Respondents are also likely to have a stronger interest in the survey topic than nonrespondents. Although the *Digest* mailed 10 million sample ballots to prospective voters, it based its predictions on 2.4 million responses. Peverill Squire (1988) has shown that this low response rate combined with a nonresponse bias produced a substantial error in the 1936 poll. Feeling more strongly about the election than did the pro-Roosevelt majority, the anti-Roosevelt minority was more likely to respond to the *Digest* poll.

Another notorious polling failure occurred in 1948, known as "the year the polls elected Thomas E. Dewey President." All three major polls—Gallup, Crossley, and Roper—that covered the 1948 campaign projected that Dewey would be the winner with 50 percent or more of the popular vote, but he ended up with just over 45 percent. Several factors combined to produce this failure. For one thing, those who were undecided about their election choice when the polls were taken went predominantly for Truman. Critics of the 1948 polls, however, agree that the way the samples were selected also contributed to the prediction error.

By this time the pollsters were using a procedure called "quota sampling" (see p. 174). With quota sampling the population is divided into subgroups (e.g., men and women, blacks and whites), and interviewers are assigned a fixed quota of persons to interview in each of the subgroups. For example, an interviewer might be required to interview three men under 40 years of age and four over 40, as well as four women under 40 and four over 40. Quotas are determined so as to guarantee that the sample will be representative of the voting population with respect to important characteristics thought to affect voting behavior. One shortcoming of this method is that the quotas are theoretical; they are an indirect means of creating a representative sample of the voting population *before the election*. However, demographic characteristics of the voting population cannot be known with certainty until after the election is held. Another problem is that the interviewers, who are free to choose anybody they like within the assigned quota, may be subject to selection bias. In fact, Gallup and Crossley poll results from 1936 to 1948, based on quota sampling, revealed that interviewers consistently preferred Republicans (Katz, 1949). With Democratic candidate Roosevelt possessing a large lead in 1936, 1940, and 1944, Gallup and Crossley were able to predict the winner in spite of the Republican bias of their polls. But this was too much to overcome in an election as close as that of 1948.

Population Definition

After determining the unit of analysis, the first task in sampling is to define the population of interest—to describe the particular collection of units that make up the population. Sociologist Kenneth Bailey (1982:86) notes that the experienced researcher always gets a clear picture of the population before selecting the sample, thus starting from the top (population) and working down (to the sample). "In contrast, novice researchers often work from the bottom up." Rather than making explicit the population they wish to study, they select a predetermined number of conveniently available cases and assume that the sample corresponds to the population of interest. Consider a sample consisting of "randomly" chosen passersby at a shopping center on a Saturday afternoon. What could this sample possibly represent? There is simply no way of knowing. It almost certainly would not be representative of people in the surrounding community. For one thing, weekend shoppers would probably overrepresent people who work during the week and underrepresent the retired and unemployed. And if the shopping area happens to be a luxurious shopping mall, the sample is likely to be heavily weighted in favor of the financially well-off. At best, this sample would represent Saturday afternoon shoppers at the shopping center in question on the day the data were collected, which is a strange definition of the population and surely not the one in which the researcher would be interested.

Defining the population is a two-step process. First, one must clearly identify the **target population**, that is, the population to which the researcher would like to generalize his or her results. To define the target population, the researcher must specify the criteria for determining which cases are included in the population and which cases are excluded. The relevant criteria depend on the type of unit and the research topic. With individual people, some combination of locale, age, and selected demographic variables such as gender, race, employment status, and age is ordinarily used (Sudman, 1976:12). For example, in a study of aging and religion, Stephen Ainlay, Royce Singleton, and Victoria Swigert (1992) defined the target population as "noninstitutionalized residents of Worcester, Massachusetts, who were 65 years of age or older."

For the National Education Longitudinal Study (NELS), which Beckett Broh (2002) used to analyze the relationship between extracurricular activities and academic achievement, the target population for the base year, 1988, consisted of students enrolled in "all public and private schools containing eighth grades in the fifty states and District of Columbia" (Spencer et al., 1990:6). This definition was clarified further by taking into account several pragmatic considerations, as the researchers note:

> Excluded . . . are Bureau of Indian Affairs (BIA) schools, special education schools for the handicapped, area vocational schools that do not enroll students directly, and schools for dependents of U.S. personnel overseas. The student population excludes students with severe mental handicaps, students whose command of the English language was not sufficient for understanding the survey materials (especially the cognitive tests), and students with physical or emotional problems that would make it unduly difficult for them to participate in the survey. (p. 6)

If the unit is groups or organizations, then organizational type and size as measured by number of members or employees are often specified. For instance, to study factors influencing the adoption of maternity-leave policies, Erin Kelly and Frank Dobbin (1999) limited the population to public, nonprofit, and private-sector establishments with fifty or more workers in three states: California, New Jersey, and Virginia. Studies of families also invoke unique definitional criteria: for example, presence or absence of children, number of generations, and intact or single parents. Which set of criteria is used depends on the researcher's purposes.

Two defining characteristics that always implicitly or explicitly define the target population are its geographic and time referents. Notice that each of the examples of target populations has a clear political or geographic boundary (city of Worcester, United States, three states) and each has a distinctive time frame (1989, 1987–88 academic year, 1986).

The researcher next must find a way of making the target population operational. This involves the second phase of population definition: constructing the sampling frame. The **sampling frame** denotes the set of all cases from which the sample is actually selected. Since the term can be misleading, please note that the sampling frame is *not a sample*; rather, it is the *operational definition of the population* that provides the basis for sampling.

There are two ways of constructing a sampling frame, which correspond to the two ways to define a set (or subset) in logic: (1) listing all cases and (2) providing a rule defining membership. For example, in a city telephone survey, the sampling frame could consist of the city phone book (a listing) or the set of all telephone numbers with certain telephone exchanges (a rule). In survey research, establishing a sampling frame often amounts to obtaining an adequate listing—either of the population as a whole or of subgroups of the population. But listing is not always possible or preferable. As long as cases can be identified, a rule procedure can usually be devised for finding and selecting cases. Suppose that you wanted to observe the behavior of inmates in a mental institution. Listing all acts before they occur is obviously impossible. However, if we consider that all acts within the institution occur at specified places, days, and times, then this would provide the basis for a three-dimensional rule (site, time, and day) for identifying and selecting cases. Similarly, if you wanted to interview people attending a rock concert, the fact that everyone must arrive at the concert at a particular time allows you to establish a rule based on time of arrival.

KEY POINT

To draw a sample, first specify the target population and then find a sampling frame, usually a list, identifying all cases within the population.

It is the researcher's hope that the sampling frame and the target population will be identical. Unfortunately, this usually occurs only for very small, geographically concentrated populations. Because it is often impossible or impractical to create an accurate list of the target population, one must rely on existing lists; and such lists are inevitably incomplete. For the aging and religion study, the sampling frame consisted of the 1989 Worcester city census, which provided names, addresses, and birth dates. However, not only was it likely that many people were not counted in the census, but through deaths and migration in and out of Worcester, the target population was bound to have changed somewhat by the time the interviews were actually conducted. Thus, the sampling frame could not have matched precisely the population of interest.

Strictly speaking, inferences should be made only about the population represented by the sampling frame. Yet, it is the target population to which we wish to generalize. Therefore, one should always evaluate the extent to which cases in the target population have been omitted from the sampling frame. How many omissions are there? How are they likely to differ from cases included in the sampling frame? Because they frequently differ in a systematic way from those included in the frame, excluded cases can seriously bias findings. Telephone directories, for example, exclude the poor, who cannot afford telephones, and the more wealthy, who tend to have unlisted numbers. Whenever such biases are known to exist, the researcher should decide whether to use the list or to find an alternative sampling frame. Regardless of the frame that is used, however, the nature of possible excluded cases should be acknowledged and carefully taken into account when making inferences about the target population.

Obviously, the problem of omitted cases is solved best by finding a good list to begin with. Despite its flaws, the annual Worcester city census provided a reasonably accurate frame. For sampling individuals, many small and mid-size U.S. cities have population and household directories. If the target population consists of a

local professional association, union, school, church, or some other institution, membership directories are usually available and provide excellent frames. But at the national level few lists exist. An explicit list of all 8th-grade students in the United States, for example, is simply not available. To sample national populations, researchers generally break the target population down into natural groupings for which lists are available or can be constructed at low cost. For the NELS, the researchers obtained a list of public and private schools in the United States, a sampling frame that comprised an implicit list of all 8th-grade students. We discuss this sampling strategy further under "Sampling Designs."

Often, there is no single list that adequately enumerates the population. This is particularly true for organizations as a unit of analysis. Although national lists exist for special populations such as hospitals, newspapers, and businesses, no one has tallied American voluntary associations (Knoke, 1990) or all public organizations, let alone all types of organizations. The sampling frame in the Kelly and Dobbin study of maternity-leave policy was compiled from several sources (see Dobbin et al., 1993). First, thirteen sectors were chosen to represent different parts of the economy, including publishing, banking, retail trade, and transportation. Then, the most complete published list of organizations for each sector was selected. Among the frames were Dun's *Million Dollar Directory* (for various private-sector organizations), the directory of the American Hospital Association (for hospitals), and the *Encyclopedia of Associations* (for nonprofits). To enumerate organizations in the public sector, the investigators used commercial telephone directories as well as official state and federal government directories.

Equally problematic is finding or constructing sampling frames for studies of rare or special target populations, such as Vietnam War veterans, people with a particular illness or disability, or people with incomes above or below a given amount. If a list of the general population is available, it may be possible to screen the rare population from the larger sampling frame. In the Worcester study of the elderly, the sampling frame for the general population was culled to include only people over 65 years of age. Alternatively, members of the special target population, such as people over age 65, could be identified at the data-collection stage.

Finally, some populations defy identification and listing. Imagine trying to identify for sampling purposes members of a terrorist group or other radical political groups, such as Aryan Nations, or deviant populations such as drug abusers, criminals, or delinquents. For obvious reasons, political radicals and deviants tend to resist identification by anyone outside their own group. In studying such populations, therefore, the process of population definition breaks down because an adequate sampling frame is unobtainable.

Sampling Designs

Ideally, in sampling we would like to obtain a sample that will be representative of the target population. To be "representative" means to provide a close approximation of certain characteristics of the target group. If the population consists of the students at Alpha College, then a perfectly representative sample would be like Alpha College students in all respects. The sample would contain the same proportion

of first-, second-, third-, and fourth-year students, the same proportion of commuter students, the same proportion of sociology majors, and so forth, as is contained in the student body as a whole.

While intuitively appealing, however, the concept of "representativeness" has nearly disappeared from the technical vocabulary of sampling (Kish, 1965). Not only is it extremely unlikely that one will be able to draw a *perfectly* representative sample, it is rarely possible to evaluate a specific sample in terms of its overall representativeness. Because the populations we study are not known in all respects (which is, of course, the reason we study them), there is no way of knowing just how representative a given sample is. Therefore, we use the term "representative" in reference to specific, known population characteristics but not in reference to the overall quality of the sample.

The quality of a sample must be judged in terms of the procedure that produced it, that is, in terms of its sampling design. **Sampling design** refers to that part of the research plan that indicates how cases are to be selected for observation. Sampling designs are generally divided into two broad classes: probability and nonprobability. **Probability sampling** is scientifically more acceptable, although it is not always feasible or economical. Its essential characteristics are that all cases in the population are randomly selected and have a known probability of being included in the sample. In **nonprobability sampling**, the chances of selecting any case are not known because cases are nonrandomly selected.

Probability sampling designs offer two major advantages over nonprobability sampling designs. The first advantage is that they remove the possibility that investigator biases will affect the selection of cases. The second advantage is that, by virtue of random selection, the laws of mathematical probability may be applied to estimate the accuracy of the sample. With probability sampling, one knows to which population the sample may be generalized as well as the limits of generalizability; but with nonprobability sampling, the population itself is undefined and the laws of probability do not apply.

Probability Sampling

Random Selection

Probability sampling always involves the process of **random selection** at some stage. In popular usage, "random" describes something that occurs or is done without plan or choice; random events or choices are haphazard or nondeliberate. As used in sampling, however, "random" has a more specific, technical meaning. It refers to a process that gives each element in a set, such as each case in a population, an equal chance of being selected. If for any reason the selection process favors certain cases or if the selection of one case increases or decreases the likelihood that another case will be selected, then the selection is biased. By this definition, your circle of friends would clearly be biased as a sample of the student population at your college. So also would a variety of other, more or less haphazardly chosen samples, such as the students enrolled in introductory sociology classes, students who happen to be in the library on a weekday night, or those students who pass in

front of the campus center at a given time. In each of these samples, it is a good bet that certain types of students are more likely to appear than others; therefore, they are not random samples of the student body.

KEY POINT

The quality of a sample rests on the procedure for selecting cases; good "representative" samples always involve *random selection.*

To satisfy the condition of randomness, the investigator cannot simply pick cases haphazardly or in any hit-or-miss fashion; subtle and often unconscious biases will invariably enter into the selection process. Rather, mechanical or electronic aids should be used to ensure that chance alone dictates selection. One popular "mechanical" procedure is the lottery method. All cases are numbered; the numbers are written on equal-size objects, such as balls or disks; the objects are thoroughly mixed in a container; and then the requisite number of "cases" are selected one by one. This method is cumbersome and impractical, however, when applied to populations of social scientific interest.[2] In actual practice, social researchers usually draw random samples by using a computer program that generates random numbers or selects cases randomly from a computerized list of the population.

Simple Random Sampling

The basic probability sampling design that is incorporated in all the more elaborate probability sampling designs discussed below is called a **simple random sample.** The defining property of a simple random sample is that every possible *combination* of cases has an equal chance of being included in the sample. For example, in a population of four cases, numbered 1, 2, 3, 4, there are six possible samples of size two: (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), and (3, 4). In a simple random sample, each of these pairs of cases would have the same chance of selection. To guarantee this property, two requirements are necessary: (1) a complete list of the population and (2) the random selection of cases to be included in the sample.

With the concepts of "random selection" and "simple random sampling" now in mind, we are in a position to examine the principles of probability sampling theory. Although a complete understanding of the theory would require advanced training in statistics, the reader can gain a good intuitive grasp of the essential concepts through the following small-scale hypothetical example.[3]

Suppose that you are interested in conducting a study of alcohol consumption on your campus. Your interest perhaps stems from recent reports that problem drinking is increasing among college-age youth. Very limited time and resources prevent you from studying the student population as a whole, so you focus instead on all the students who live in your off-campus condominium. There are eight residents in all, which makes it easy to create a list of the target population. You realize, of course, that you would not need to draw a sample of such an incredibly small population. But to understand how samples vary, let us look at some characteristics of samples drawn from this imaginary population. Your dependent variable is alcohol consumption, which you measure by administering a questionnaire in which you ask each resident how many drinks he or she consumes on a typical weekend

night. (You also tell your respondents to consider one drink as a bottle or glass of beer, a glass of wine, a wine cooler, a shot glass of liquor, or a mixed drink.)

The following tabulation contains hypothetical data for the population of eight students:

Variables	*Ann* A	*Bea* B	*Cora* C	*Dee* D	*Ed* E	*Fred* F	*Greg* G	*Hal* H
Gender	F	F	F	F	M	M	M	M
Class standing	U	G	U	G	U	G	U	G
Number of drinks	3	1	0	2	3	4	6	5

Notice that, in addition to alcohol consumption, information is given on gender and undergraduate versus graduate student status. There are four females (F) and four males (M); four undergraduates (U) and four graduate (G) students. These two variables will be considered later.

Characteristics of a population are called **parameters**. For our population of eight students, we are interested in the average (or mean) number of drinks consumed. We determine this parameter by adding up all of the reported drinks (24) and dividing by the number of cases (8), which equals 3.0. Of course, in the real research situation, the parameter would not be known, and our task would be to obtain an estimate from a sample of observations. Sample estimates of population parameters are called **statistics**. Just as we calculated the mean number of drinks consumed in the population, we can estimate the population mean by calculating the mean number of drinks consumed in a sample. To get an idea of how such sample estimates will vary, let us examine all possible simple random samples of two students and of four students that could be drawn from the population. By seeing how the sample statistics vary in relation to the population parameter, we also can see the kinds of generalizations about a population that are possible from a single sample.

There are twenty-eight possible combinations of two cases each that can be drawn randomly from eight cases; thus, there are twenty-eight simple random samples of size two. You can check this for yourself by systematically arranging the possibilities in the following manner: AB, AC . . . AH, BC, BD . . . GH. In fact, we have done this for you in Table 6.1, which also gives the mean number of drinks for each combination.

Similarly, there are seventy simple random samples of size four. To see how the estimates of alcohol consumption will vary from sample to sample, let us calculate the mean number of drinks consumed for each possible sample of size two and size four. Calculating a statistic, such as the mean, for all possible samples of a given size results in a **sampling distribution**. The sampling distributions of the means for samples of size two and size four are presented in Figure 6.1 and in columns two and four of Table 6.2. Note first the sample of two cases: Only one combination (BC) yields a sample mean of 0.5 (Bea has 1 drink and Cora has 0, for an average of 0.5); only one combination (CD) has a sample mean of 1.0; three combinations (AC, BD,

TABLE 6.1. **List of All Possible Samples of Size Two**

Combination of cases	*Mean number of drinks*	*Combination of cases*	*Mean number of drinks*
AB (Ann, Bea)	2.0	CE (Cora, Ed)	1.5
AC (Ann, Cora)	1.5	CF (Cora, Fred)	2.0
AD (Ann, Dee)	2.5	CG (Cora, Greg)	3.0
AE (Ann, Ed)	3.0	CH (Cora, Hal)	2.5
AF (Ann, Fred)	3.5	DE (Dee, Ed)	2.5
AG (Ann, Greg)	4.5	DF (Dee, Fred)	3.0
AH (Ann, Hal)	4.0	DG (Dee, Greg)	4.0
BC (Bea, Cora)	0.5	DH (Dee, Hal)	3.5
BD (Bea, Dee)	1.5	EF (Ed, Fred)	3.5
BE (Bea, Ed)	2.0	EG (Ed, Greg)	4.5
BF (Bea, Fred)	2.5	EH (Ed, Hal)	4.0
BG (Bea, Greg)	3.5	FG (Fred, Greg)	5.0
BH (Bea, Hal)	3.0	FH (Fred, Hal)	4.5
CD (Cora, Dee)	1.0	GH (Greg, Hal)	5.5

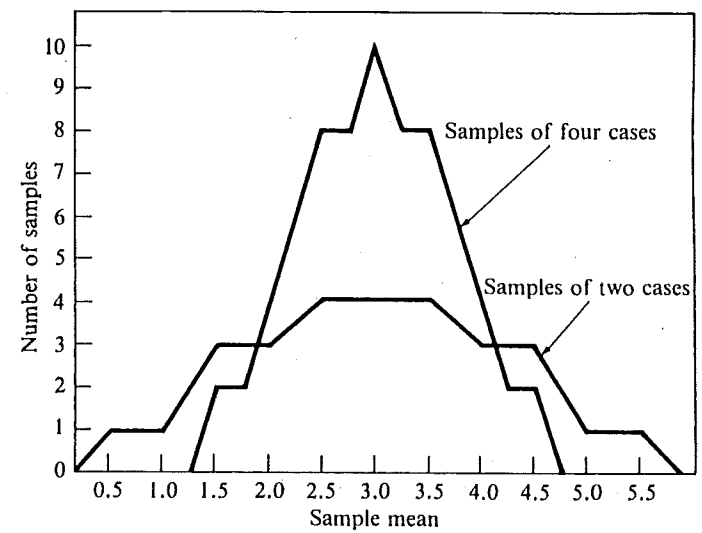


FIGURE 6.1. Frequency distributions of sample means for simple random samples of two cases and four cases drawn from a hypothetical population of eight cases with a mean of 3.0.

CE) have sample means of 1.5; and so on. Now, examining samples of four cases, two combinations (ABCD and BCDE) yield sample means of 1.5, two combinations (ABCE and BCDF) yield sample means of 1.75, and so on.

From the sampling distributions one can calculate the probability that any given sample will be selected. (Probabilities, which range from 0 to 1, correspond

to percentages. Thus, an event with a probability of .40 is likely to occur 40 percent of the time.) Recall that with simple random sampling, each possible combination of cases is equally likely to be selected. This means that the likelihood, or probability, of selecting any particular combination of two cases (e.g., BC) from the above population will be equal to 1 divided by the total number of combinations (or 1/28 = .036).[4] This is also the probability of drawing a sample with a mean of .5 or of drawing a sample with a mean of 1.0 since each of these means is the result of one and only one combination of two cases. To get the probability of selecting a sample with a mean of 1.5, we simply add the probabilities of each combination that yields a mean of 1.5 (1/28 + 1/28 + 1/28 = .107), which, as you can see, equals the number of such combinations (3) over the total (28). The remaining probabilities are given in columns three and five of Table 6.2. From the table,

TABLE 6.2. Means and Probabilities of Simple Random Samples of Two Cases and Four Cases Drawn from a Hypothetical Population of Eight Cases with a Mean of 3.0

	Sample of two cases			*Sample of four cases*		
SAMPLE MEANS	NUMBER OF SAMPLES	PROBABILITY		NUMBER OF SAMPLES	PROBABILITY	
0.50	1	.04				
0.75						
1.00	1	.04				
1.25						
1.50	3	.11		2	.03	
1.75				2	.03	
2.00	3	.11		4	.06	
2.25				6	.09	
2.50	4	.14		8	.11	
2.75				8	.11	
3.00	4	.14	.64	10	.14	.89
3.25				8	.11	
3.50	4	.14		8	.11	
3.75				6	.09	
4.00	3	.11		4	.06	
4.25				2	.03	
4.50	3	.11		2	.03	
4.75						
5.00	1	.04				
5.25						
5.50	1	.04				
Total number of samples	28			70		
Mean of sample means	3.0			3.0		

notice that the probability of obtaining a particular sample mean increases as the estimate approaches the actual population mean of 3.0 drinks. By the same token, the more a sample mean deviates from the population mean, the lower its probability of selection.

It is also possible to calculate the probability of obtaining sample estimates within a given range of the population mean. For example, there are eighteen samples (count them) of two cases with means between 2.0 and 4.0; thus, the probability of obtaining a sample mean within this range is 18/28 = .64. Similarly, with samples of four cases, the probability of getting a sample mean within the range of 2.0 to 4.0 is 62/70 = .89; and with samples of six cases (not shown), the probability is 1.0 (i.e., all samples have means between 2.0 and 4.0). This illustrates a very important principle in sampling theory: The larger the sample, the closer its mean is likely to be to the population mean.

The amount that a given sample statistic deviates from the population parameter it estimates is known as **sampling error**. For example, in our hypothetical population, a sample mean of .5 has a sampling error of 2.5 because this is how much it differs from the known population mean of 3.0. Similarly, a sample mean of 1.5 has a sampling error of 1.5 (3.0 − 1.5 = 1.5). The statistical measure of the "average" of such errors for an entire sampling distribution is called the **standard error**. Using this concept, then, another way of stating the principle about sample size is, the larger the sample, the smaller the standard error.

KEY POINT

As sample size increases, the *standard error* (a measure of the "average" sampling error) decreases.

So far, we have demonstrated that knowledge of the sampling distribution of the mean enables one to make the following kinds of statements regarding the accuracy of a sample mean: The probability is .89 that the mean of a randomly selected sample of four cases will deviate from the population mean by 1 or less (i.e., will fall within the range 2.0–4.0). Of course, in the above example, the sampling distributions were constructed by calculating means for all possible random samples from a *known* population. In actual research, we select only one sample in order to estimate an *unknown* population characteristic. The key to being able to make the same kind of statement about sample accuracy in the actual research situation is provided by the mathematical theory behind probability sampling. According to this theory, the sampling distributions of various statistics (such as the mean) form stable, predictable patterns. Consequently, even though the population parameter is not known, the theory indicates *how* sample estimates will be distributed (by specifying the sampling distribution) and provides a statistical formula for calculating the standard error, a measure of *how much* the sample estimates will tend to vary.

More rigorous accounts of sampling theory and sample estimation procedures can be found in the inferential statistics section of most statistics textbooks. In actual research, sample inferences about a population parameter are arrived at in the following manner. From a single sample, drawn randomly:

1. The sample statistic (e.g., the sample mean) is considered the best single estimate of the population parameter (e.g., the population mean).
2. The extent to which the sample statistic tends to deviate from the population parameter, indicated by the standard error, is calculated on the basis of variation within the sample.
3. Theoretical knowledge about the sampling distribution is utilized to attach a probability, or level of confidence, to a calculated range within which the population characteristic should fall.

This range, called a **confidence interval**, is established around the sample statistic. From this procedure we end up with statements such as the following: "We are 95 percent confident that the population mean is between 2.0 and 4.0." With the mean of 3.0 as the midpoint of this interval, the margin of random sampling error—the amount by which a sample mean is likely to deviate from the population mean—is plus or minus 1. In presenting the results of opinion polls, newspapers and magazines usually give the margin of error for an unstated 95 percent level of confidence; for example, when reporting the percentage of respondents who gave a particular opinion, they may note that the "margin of error is plus or minus 4 percentage points, based on a sample of 750 adults."

KEY POINT

The *confidence interval* indicates the range of values within which the estimated population value is likely to lie. The *level of confidence* is the probability (e.g., 95 percent) that the interval contains the population value.

A major advantage of probability sampling is that the size of the confidence interval (i.e., sample precision), or the amount by which the sample statistic is likely to differ from the population parameter, can be estimated *before* a study is carried out. Recall that the possible error in sample statistics is a function of sample size: The larger the sample, the smaller the standard error. Therefore, one can get a more precise sample estimate by selecting a larger sample. Sample precision also may be increased by using a more complex sampling design known as "stratified random sampling."

Stratified Random Sampling

In **stratified random sampling**, the population is first subdivided into two or more mutually exclusive segments, called **strata**, based on categories of one or a combination of relevant variables. Simple random samples then are drawn from each stratum, and these subsamples are joined to form the complete, stratified sample.

To illustrate, let us return to our hypothetical data. Note that the population of eight cases can be broken down into two strata—male and female—according to categories of the variable "gender" or into two strata—undergraduate and graduate—according to categories of the variable "student status." A stratified random sample of four cases could be drawn by randomly selecting an equal proportion of males and females. In other words, one could draw independent random samples of two cases from each stratum, male and female, and then combine these into a

single sample. This would not produce a simple random sample since there are several combinations of cases that are not possible—for example, the combinations of all females (ABCD) or all males (EFGH). There are thirty-six possible stratified random samples of four cases with equal proportions of males and females. By computing the mean number of drinks consumed for each of these possible samples, we again obtain a sampling distribution of the mean. This distribution can be compared with the sampling distribution based on simple random sampling to demonstrate the effect of stratifying on sample precision.

Table 6.3 presents the sampling distributions for samples of four cases based on simple random sampling, stratified random sampling with gender as the stratification variable, and stratified random sampling with student status as the stratification variable. Notice by examining the last two rows of the table that, compared with simple random sampling, sampling with gender stratification yields a distribution of sample means that cluster more tightly about the population mean. Thus, it is possible for a stratified sample to give a better estimate than a simple random sample. On the other hand, stratifying by student status offers no such improvement; indeed, the percentage of sample means between 2.5 and 3.5 (i.e., that deviate from the population mean by as little as .5) is lower—44 percent—than in simple random sampling—60 percent.

TABLE 6.3. Means of Samples of Four Cases Drawn by Simple Random Sampling and Stratified Random Sampling with Gender and Student Status as Stratifying Variables

Sample means	*Simple random samples*	*Samples stratified by gender*	*Samples stratified by student status*
1.50	2		2
1.75	2		
2.00	4	1	2
2.25	6	2	6
2.50	8	5	2
2.75	8	6	2
3.00	10	8	8
3.25	8	6	2
3.50	8	5	2
3.75	6	2	6
4.00	4	1	2
4.25	2		
4.50	2		2
Total number of samples	70	36	36
Mean of sample means	3.0	3.0	3.0
Percentage of sample means between 2.5 and 3.5	60	83	44
Percentage of sample means between 2.0 and 4.0	89	100	89

Clearly, as these comparisons reveal, stratifying can contribute to sampling efficiency. Whether it does depends on whether the stratifying variable is related to the study variable. In our hypothetical population, gender is related to alcohol consumption but student status is not; that is, men tend to drink more than women, whereas undergraduate and graduate students drink about the same amount. Consequently, stratifying by gender increases sample precision but stratifying by student status does not.

Technically speaking, stratifying by variables correlated with the study variable increases the precision of estimates because it systematically controls relevant sources of variability (or heterogeneity) in the population. The effect of this is to eliminate a source of sampling error—for example, the error that could occur if our sample of students contained all men or all women. Thus, when there are differences across strata, stratified sampling ensures that these differences are accounted for and are not free to vary within the sample. In fact, in the extreme instance, when all the variation in the dependent variable is across strata and each stratum is completely homogeneous, the researcher needs to sample only one case per stratum for a perfectly accurate estimate. In our hypothetical population, this would occur if, say, each woman had one drink and each man drank five drinks. The population mean would be 3.0, and a sample of one woman and one man would yield a mean estimate of 3.0 [$(5 + 1)/2 = 3.0$)].

If the stratifying variable is related to the variable under study, a stratified random sample will be more precise and, hence, more efficient than a simple random sample of the same size. But since we also can increase sample precision by increasing sample size, the greater efficiency of stratifying depends on whether the cost of stratifying is low relative to the cost of sampling more cases. Obtaining a stratified sample necessitates classifying cases according to categories of the stratifying variable *prior to drawing the sample.* Therefore, the sampling frame must contain information on the stratifying variable. Otherwise, you cannot draw separate samples from each stratum. Often, however, the information needed to stratify is not available and is either impossible or too costly to obtain, which limits the use of stratified random sampling.

In addition to increasing efficiency, stratified random sampling may be used to guarantee that variable categories with small proportions of cases in the population are adequately represented in the sample. Suppose, for example, that we wanted to examine ethnic differences in alcohol consumption on a campus where 90 percent of the students are white, 5 percent are black, and 5 percent are Asian. A simple random sample of one hundred would yield, on the average, ninety whites, five blacks, and five Asians. However, the number of cases in the latter two sampled groups is so small that statistical estimates derived from them would be very unreliable. To get around this problem, we could obtain a stratified random sample in which we select a greater proportion of blacks and Asians for our sample than are found in the population. We might select forty whites, thirty blacks, and thirty Asians. This is called **disproportionate stratified random sampling** because the proportion of cases in each stratum of the sample does not reflect the proportion in the population.

When strata are sampled proportionate to the population composition, each case has an equal probability of being selected and one can generalize directly from sample to population. But with disproportionate sampling, the probability of selection varies from stratum to stratum and one must make a statistical adjustment before estimating population parameters.[5] Disproportionate stratified random sampling still constitutes a probability sample because the probability of case selection is *known.* Furthermore, probability sampling theory applies since the subsamples drawn from each stratum are simple random samples.

Cluster Sampling

Both simple random sampling and stratified random sampling assume that a complete list of the population is available. What do you do, however, when your target population is so large that it is either impossible or impractical to list all its members? The only lists of the U.S. population—income tax forms and census forms filed every 10 years—are confidential and not available to the public. Lists of most city and all state populations simply do not exist, and it would be prohibitively expensive to compile lists of these and other very large populations. In such instances, the researcher often is able to obtain a sample in stages, using a method called **cluster sampling.**

In cluster sampling, the population is broken down into groups of cases, called "clusters." The clusters generally consist of natural groupings, such as colleges and churches, or geographic areas, such as states, counties, cities, and blocks. Unlike stratified sampling, which draws cases from each stratum, cluster sampling draws cases only from sampled clusters. To draw a cluster sample, one first selects at random a sample of clusters. Then one obtains a list of all cases within each selected cluster. If all the cases in each sampled cluster are included in the sample, the design is called a *single-stage cluster sample* in that sampling occurs once—at the cluster level. More frequently, cluster sampling involves sampling at two or more steps or stages, hence the term **multistage cluster sampling.** An example of a two-stage cluster sample would be a random selection of 4-year colleges (first stage), followed by a random selection of students within each sampled college (second stage). The sampling units in the first stage of a multistage sample are termed **primary sampling units** (or PSUs); units in the second stage are termed *secondary sampling units,* and so on.

While stratified random sampling is used either to increase sample precision or to provide a sufficient number of cases in small strata, the principal reason for cluster sampling is to reduce the costs of data collection. In face-to-face interview studies of large, widely scattered populations, two major costs are interviewer travel and the listing of population elements. Simple random and stratified random methods produce isolated interviews at dispersed localities, which can create substantial travel costs to reach respondents. However, clustering concentrates interviews within fewer and smaller geographic areas, thereby spreading the travel costs over several cases and reducing the costs of any one interview.

Since the listing of population elements is a prerequisite for simple random or stratified random methods, clustering also can reduce costs by limiting the compilation of lists of cases to selected clusters rather than compiling a list of the entire population. Moreover, multistage cluster sampling may be the only viable method

of drawing a probability sample when an exhaustive list of population elements is unavailable or impractical to create. Often, when this occurs, it is possible to obtain lists of subpopulations (clusters) into which the population is naturally grouped or can be subdivided. For example, although a single list of all 8th-grade students does not exist and would be difficult and extremely costly to compile, a complete and accurate list of public and private schools in the United States is maintained by Quality Education Data, Inc. The NELS used this database to draw a two-stage cluster sample for the 1988 base year survey. From the nearly 40,000 schools nationwide, the NELS obtained a sample of 1057 schools; then, from each participating school, twenty-six students were randomly selected (Spencer et al., 1990). (Box 6.2 provides an example of a multistage sampling design for a national survey of the adult population.)

KEY POINT

Stratified sampling increases efficiency by reducing sampling error (provided that the strata are related to the study variable). *Multistage cluster sampling* increases efficiency by reducing costs.

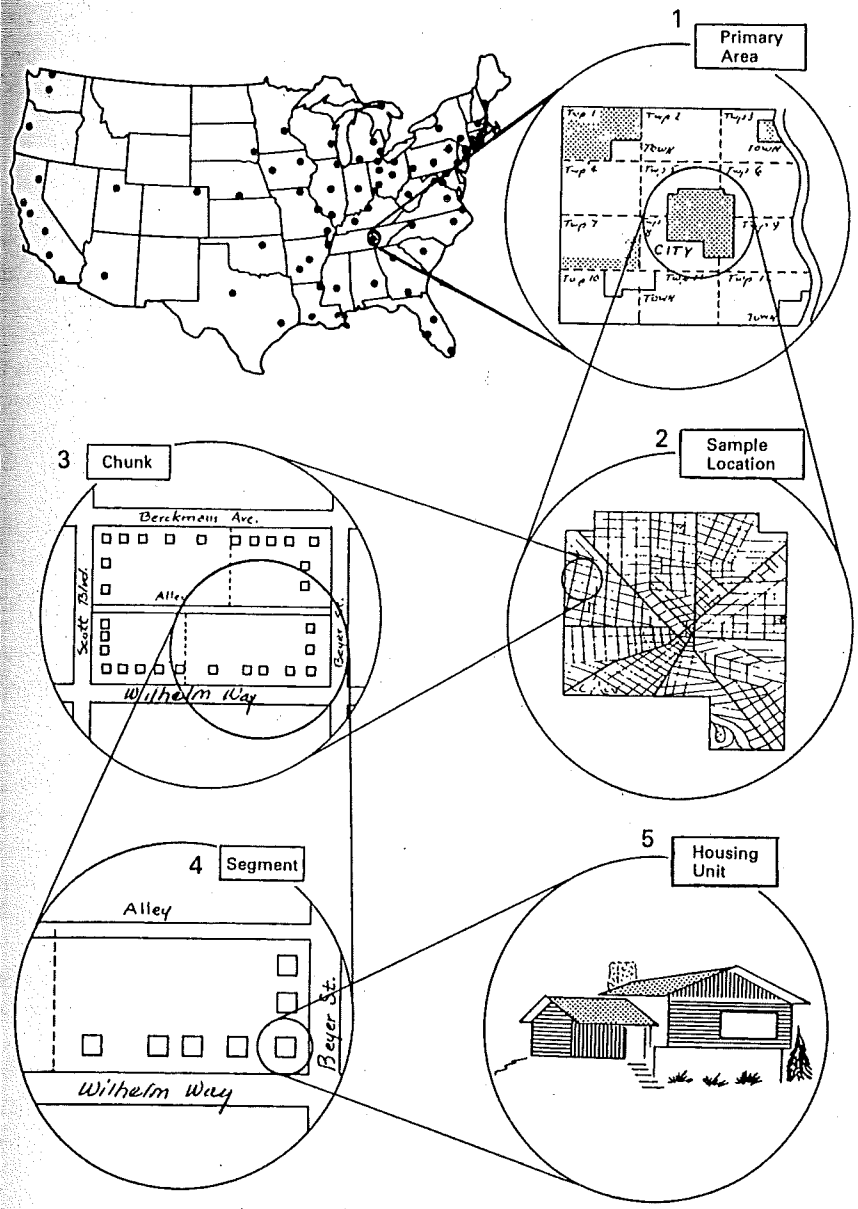
Multistage sample designs may involve either simple random sampling or stratified random sampling at each stage of the design. In other words, one can stratify clusters just as one stratifies individual cases. The NELS stratified clusters, as schools were divided into several strata according to four variables: school type (public, Catholic private, other private), region of the country, urbanicity (whether the school's location was urban, suburban, or rural), and percent minority. A number of schools were then selected within each stratum.

One problem with cluster sampling is that natural clusters often vary considerably in the numbers of cases they contain. This, in fact, was true of the schools in the NELS. There are two ways of handling this problem in order to reduce potential error. First, you can stratify clusters by population size. In this way, a relatively few clusters with extremely large populations can be sampled from a separate stratum to guarantee that they are represented in the sample. Otherwise, with simple random sampling one runs the risk of not selecting clusters that account for a major proportion of the population. Imagine, for example, drawing a cluster sample, with counties as the primary sampling units, of the population of Nevada, in which over 80 percent of the people reside in two of the state's sixteen counties. Any sample that excludes these two counties would be a biased sample of Nevada residents. Yet the chances are fairly high that neither of the state's two largest counties would be included in a simple random sample of counties (e.g., the probability that neither would be selected is 55 percent in a sample of four counties and 23 percent in a sample of eight counties). The best procedure, therefore, would be to stratify by county size, placing the two most populous counties in the topmost stratum, and then sample 100 percent from this stratum (i.e., select both counties).

When clusters vary in size, another common solution is to make the selection of clusters or cases within clusters proportionate to the size of the cluster. This method, called **probability proportionate to size sampling**, was applied to the selection of schools in the NELS. One way it works is to assign each cluster a probability of selection proportionate to the number of cases in the cluster. Thus, a school with 1000 students would be twice as likely to be selected as a school with 500 stu-

BOX 6.2 Sampling Design for a Nationwide Survey

Survey organizations such as the Survey Research Center (SRC) at the University of Michigan use multistage cluster sampling to conduct nationwide surveys. The steps in-



volved in selecting the SRC's national sample in the 1970s are roughly diagrammed in the accompanying figure. The steps are numbered and labeled according to the type of unit selected.

Step 1. The United States is divided into *primary areas* consisting of counties, groups of counties, or metropolitan areas. These areas are stratified by region, and a proportionate stratified sample of seventy-four areas is selected.

Step 2. The seventy-four areas are divided into *locations* such as towns, cities, and residual areas. After these have been identified and stratified by population size, a proportionate stratified sample of locations is drawn within each area.

Step 3. All sample locations are divided into *chunks.* A chunk is a geographic area with identifiable boundaries such as city streets, roads, streams, and county lines. After division into chunks, a random sample of chunks is drawn.

Step 4. Interviewers scout each sample chunk and record addresses and estimates of the number of housing units at each address. They then divide the chunks into smaller units called *segments*, and a random sample of segments is selected.

Step 5. Within each sample segment, either all or a sample of the housing units, usually about four, are chosen for a given study. Finally, for every housing unit in the sample, interviewers randomly choose one respondent from among those eligible, which ordinarily consists of all U.S. citizens 18 years of age or older.

Source: Survey Research Center, 1976. *Interviewer's Manual,* rev. ed., p. 36. Ann Arbor: Institute for Social Research. Used by permission.

dents. If an equal number of students is selected from each school, then each student in the population has an equal probability of being selected.[6]

Table 6.4 compares multistage cluster sampling with stratified random sampling. While cluster samples are more cost efficient, they are less precise, size for size, than either simple random or stratified random samples. The latter designs have a single source of sampling error; however, there are sampling errors associated with each stage of a multistage cluster sample. For example, in a two-stage cluster sample of college students at 4-year colleges, the sample of 4-year colleges would represent the population of colleges within a range of error, and the sample of students within a selected college represents all students in the college within a range of error. The sampling error in the total sample is thus compounded and can be quite large relative to the error produced by one simple random selection.

In stratified sampling, efficiency increases to the extent that strata are internally homogeneous; for example, the drinking behavior of women is similar and the drinking behavior of men is similar. By contrast, the ideal in cluster sampling is for clusters to be internally *heterogeneous* compared with the differences between clusters. Because you are selecting a *sample* of clusters, you essentially want each cluster to be a replica of the population. But this is rarely the case. In fact, there generally is much less variability (hence sampling error) within than across natural clusters. For example, residents in a single block are likely to be fairly homogeneous

TABLE 6.4. Comparison of Stratified and Multistage Cluster Sampling

	Stratified random sampling	*Multistage cluster sampling*
Purpose	reduce margin of error; sample small subsets of population	reduce costs
Population breakdown	by variable categories (strata)	by geographical units or natural groupings (clusters)
Sampling	within every stratum	within randomly selected clusters
Optimally efficient	when there are large differences between homogeneous strata	when there are small differences between heterogeneous clusters

with respect to income, but quite different in terms of average income from the residents in another block. For this reason, one way of increasing efficiency is to select more clusters and fewer elements within the clusters. By increasing the number of clusters, sampling error is reduced at the stage that is subject to the greatest error.[7]

Although increasing the number of clusters generally decreases sampling error, it also increases the cost of cluster sampling. A greater number of clusters means more widely scattered interviews, hence more time and expense invested in travel and more time and expense required for listing elements within the selected clusters. Reducing costs, of course, is the reason for cluster sampling. Consequently, researchers are likely to resolve this dilemma by simply selecting as many clusters as they can afford (Babbie, 1995:215).

Systematic Sampling

Before the computer age—before random number programs and computerized lists became available—social scientists often used a technique called "systematic sampling" to closely approximate simple random sampling. Simple random sampling requires not only a complete sampling frame, such as a list, but also the numbering of population elements. (In effect, one selects random numbers that correspond to case numbers.) A list in a computer file meets both of these requirements. But if you only have an unnumbered hard-copy list of the population and the list is quite long, such as a university student directory, it is much easier to draw a systematic sample than a simple random sample.

Systematic sampling consists of selecting every *K*th (e.g., fifteenth or twentieth) case from a complete list or file of the population, starting with a randomly chosen case from the first *K* cases on the list. Such a procedure has two requirements: a sampling interval (*K*) and a random start (Sudman, 1976). The **sampling interval** is merely the ratio of the number of cases in the population to the desired sample size. A *random start* refers to the process of randomly selecting the initial case between 1 and *K*. For example, suppose you wanted to draw a sample of one hundred students from the 2500 students listed in the campus directory. Dividing 2500 by 100, you obtain a sampling interval of 25. You would then select at random a number between 1 and 25 and, starting with that number, select every twenty-fifth student thereafter. Suppose the random number turns out to be 19. Your sample then would consist of cases numbered 19, 44, 69, 94, 119 . . . 2494.

Although each case in a systematic sample has an equal chance of being selected, each *combination* of cases does not. (There are only twenty-five possible combinations in this example.) Technically speaking, therefore, a systematic sample is not the equivalent of a simple random sample. It is a random sample, however, and in practice researchers regard it as a close approximation of a simple random sample.

A potential problem as well as a potential advantage with systematic sampling arises from the fact that most available lists are not random but, rather, are ordered in some fashion, such as alphabetical or by rank, age, or size of unit. The danger is that the available population listing may have a periodic or cyclical pattern that corresponds to the sampling interval. A classic example is a World War II study in which a one out of ten systematic sample of soldiers was drawn. The list from which the sample was drawn, however, was ordered by squads of ten soldiers, with each squad ordered by rank (sergeants, corporals, privates). When the sample was drawn, the interval of ten corresponded to all the sergeants on the list.

Fortunately, the danger of periodicity in a systematic sample is actually quite small. More often, an ordered list is advantageous because it provides an implicitly stratified sample; if the list is ordered in terms of a variable correlated with the study variable, this will increase precision. For example, organizational size—in terms of either assets or number of employees—is a key variable in studies of organizations. Therefore, if a list of business establishments is ordered by number of employees, a systematic sample will implicitly stratify the businesses by size.

Nonprobability Sampling

Nonprobability sampling refers to processes of case selection other than random selection. Without random selection, nonprobability samples have two basic weaknesses: (1) they do not control for investigator bias in the selection of units and (2) their pattern of variability cannot be predicted from probability sampling theory, thereby making it impossible to calculate sampling error or to estimate sample precision. However, while one should be ever mindful of these weaknesses, it would be a mistake to rule out nonprobability sampling. In many instances, this form of sampling either is more appropriate and practical than probability sampling or is the only viable means of case selection.

1. Consider the situation in which very few cases can be included in the sample. For an investigator doing an intensive study in which the unit of analysis is the city, corporation, or university, the expense of studying more than one or a few cases would be prohibitive. As we have seen, probability sampling is less reliable the smaller the sample. With extremely small samples of one or few cases, generalization from sample to population essentially becomes a matter of judgment. The selection of cases is therefore better left to expert judgment (in other words, nonrandom selection) than to the whims of chance (Wallis and Roberts, 1956:117–18).

2. When studying past events, the archaeologist or historian often finds only a fraction of relevant materials available or accessible. Similarly in contemporary societies, certain individuals or institutions may refuse to cooperate in an inquiry. Under these circumstances, the researcher must either accept a nonprobability method of case selection or abandon the study altogether.

3. In the early stages of investigating a problem, when the objective is to become more informed about the problem itself, probability sampling simply may be unnecessary. It will suffice to select a range of cases nonrandomly without concern for precise statistical generalization.

4. If the population itself contains few cases or if an adequate sampling frame cannot be obtained or constructed, there is no point in considering probability sampling. With small populations (say, fewer than one hundred), each case should be studied in its own right in comparison with all others. If the population is unknown or not readily identifiable, as in many sociological studies of deviance, sampling generally will consist of studying any and all identifiable and cooperative units.

KEY POINT

Probability sampling is preferred over *nonprobability sampling* because it controls for selection bias and allows for the calculation of sampling error; however, it is not always necessary or feasible.

Because nonprobability sampling is a residual category pertaining to any method of *non*random case selection, the various types of nonprobability samples are far-ranging and less precisely defined than the probability sampling designs examined in the previous section. With this in mind, let us now examine three primary types of nonprobability samples: convenience, purposive, and quota.

Convenience Sampling

In this form of sampling (also called "haphazard," "fortuitous," and "accidental" sampling), the researcher simply selects a requisite number from cases that are conveniently available. In terms of the study of campus alcohol consumption, you might decide to (1) ask a professor for permission to administer a questionnaire to all the students in his or her classes, (2) interview whoever comes down the dormitory hall when you happen to be there, or (3) find a convenient spot in the campus pub from which to observe behavior. Television stations and newspapers wanting to tap public opinion on specific issues may interview conveniently available commuters, shoppers, store clerks, and others (Chein, 1981). Such methods of case selection are easy, quick, and inexpensive. If the research is at an early stage and generalizability is not an issue, then they may be perfectly appropriate. However, convenience sampling is a matter of catch-as-catch-can. There is no way of determining to whom, other than the sample itself, the results apply. Consequently, such samples should be avoided whenever the goal is to make inferences about a population.

Purposive Sampling

In this form of sampling, the investigator relies on his or her expert judgment to select units that are "representative" or "typical" of the population. The general strategy is to identify important sources of variation in the population and then to se-

lect a sample that reflects this variation. One might select a single unit or subpopulation that is thought to be typical of the population in important respects or select a few units that correspond to key population differences. For example, in his study of "community ideology"—the values and beliefs associated with different kinds of communities—David Hummon (1990) selected residents of four distinct communities in northern California: a central city, an upper-class suburb, a working-class suburb, and a rural small town. Because the type of community in which people lived was likely to affect their community ideology, this was a crucial variable in Hummon's study. He therefore wanted to be sure that it was represented adequately in his sample. Furthermore, in selecting the specific communities, Hummon carefully examined census data to identify typical areas or towns and to avoid populations that contained relatively large proportions of aged, minority, or institutionalized persons. Such purposive sampling fit the aims of the study extremely well because Hummon's interest was not in estimating the distribution of various beliefs about communities within a particular population but in identifying the key elements in various community ideologies.

Another purposive sampling technique, applied to election forecasting, is the selection of "barometer" states or precincts that have served as reliable predictors of overall election returns in previous years. Still another is the selection of deviant or extreme cases in order to discover why they deviate from the norm. Finally, one may select both typical and extreme instances within the same study. For example, if one were testing an innovation in the schools, it would be wise to investigate at least one instance of the "typical school" in a particular city (or state or nation) and at least one instance of the best and worst problem schools (Cook and Campbell, 1979:76). If comparable results were found in all three instances, one would have a reasonably firm basis for inferring that the effect applies to the full range of schools.

As with any nonprobability method of case selection, purposive sampling for heterogeneity and/or typicality is an unacceptable substitute for probability sampling when precise and accurate generalizations are required. However, with studies of more limited scope or in situations that preclude random selection, purposive sampling is an acceptable alternative. It generally offers much stronger, less tenuous inferences than convenience sampling, although such inferences are very much dependent on the researcher's expert judgment. The major weakness of purposive sampling is that making an informed selection of cases requires considerable knowledge of the population before the sample is drawn.

Quota Sampling

Quota sampling is a form of purposive sampling that bears a superficial resemblance to stratified random sampling. Like the latter, quota sampling divides the population into relevant strata such as age, gender, and race. Once the strata are established, each interviewer is given quotas for each category of people that he or she is supposed to interview within a specified geographic area. For example, an interviewer covering an area of Boston might be assigned quotas of one man under 30, three men 30 and over, three unemployed women, and two employed women. To fill the quotas, the interviewer is free to choose anyone who meets the quota requirements.

The difference between quota sampling and stratified random sampling lies precisely in how cases are selected once quotas have been set. In stratified random sampling, the requisite number of cases within each stratum must be drawn by simple random sampling. But in quota sampling, the quota of cases within a stratum may be filled in whatever way the investigator chooses. Each stratum in a quota sample thus constitutes a convenience sample of the corresponding stratum of the population (Chein, 1981:426); moreover, the total sample is also a convenience sample, albeit one that resembles the population with respect to some key characteristics.

Quota categories ordinarily are chosen so that the sample will be representative of the population on important characteristics. By agreeing with the population on these characteristics, the method assumes that the sample will be like the population in other ways as well. Experience has shown, however, that this is an unsafe assumption. Interviewers allowed to self-select respondents are subject to several kinds of bias. If left to their own devices, they will interview their friends in excessive proportions; they also have been known to fill quotas by concentrating on areas where there are large numbers of potential respondents, such as transportation or entertainment centers, college campuses, and business districts (Chein, 1981:426). When quotas are filled by home visits, interviewers tend to select houses that are perceived as more attractive or as housing persons with relatively high income (Carter, Troldahl, and Schuneman, 1963). The potential biases in such procedures are obvious: Friends are likely to resemble closely the interviewers themselves; people who gravitate to population centers may differ sharply from those who do not; and a preference for nicer neighborhoods and homes will almost certainly create a strong socioeconomic bias. Thus, representative quotas on some characteristics do not ensure representation on others.

For some years before the election forecast failure in 1948 (see Box 6.1), quota sampling was the predominant sampling design in opinion surveys. Probability sampling designs are now preferred. Yet, in spite of its weaknesses, quota sampling continues to be used by some market research and opinion survey organizations because of two main advantages (Kalton, 1983). First, unlike probability sampling, there is no need to construct a sampling frame for selecting respondents. Second, whereas probability sampling requires that interviewers make several call-backs, if necessary, to contact predesignated respondents who are not at home, call-backs are unnecessary in a quota sample. The interviewer must simply find any eligible person according to the relevant quota criteria. For these two reasons, interviewing can be completed more quickly, more easily, and at much lower cost in quota sampling studies. To reduce the hazards of the quota technique, investigators have introduced tight geographic controls, specifying the section of the city, block, and even household where respondents may be selected. This will not eliminate sample biases entirely, but it does refine quota sampling to the point that it "is clearly superior to loose judgment or convenience samples" (Sudman, 1976:200).

Other Sampling Designs

In addition to the foregoing "pure" forms, investigators have developed a variety of other sampling techniques and designs. Each of these methods addresses a recur-

ring problem in social research: what to do when research purposes require the rigorous, systematic selection of cases but available resources or other practical problems preclude the use of conventional probability sampling designs.

Consider, for example, the Chicago Homeless Study (Rossi et al., 1987; Rossi, 1989), which attempted to estimate the composition and size of Chicago's homeless population. Because of the difficulties in conducting empirical research on the homeless, estimates of the extent of homelessness vary widely and knowledge of the conditions that produce it is minimal. Conventional probability sampling in surveys assumes "that persons can be enumerated and sampled within their customary dwelling unit"; however, this assumption fails by any definition of the homeless. For the Chicago study, therefore, the investigators devised a strategy that sampled persons from nondwelling units at times of maximum separation between the homeless and those with homes. The sampling design consisted of two complementary probability samples: (1) a systematic sample of persons spending the night in over twenty shelters provided for homeless persons and (2) a one-stage, stratified cluster sample of the 19,409 census blocks within the Chicago city limits. Blocks were stratified by the expected number of homeless to be found on each, according to police estimates. Within each selected block, between the hours of 1:00 AM and 6:00 AM, interviewers conducted thorough searches of every conceivable place, including alleys, hallways, roofs, basements, and parked cars. All persons encountered were first given a short screening interview to determine if they were homeless. If so, the person was given a 15- to 20-minute interview.[8]

John Watters and Patrick Biernacki (1989) devised a very different strategy for sampling another "hidden" population—injecting drug users. This is a group at high risk for the transmission of the human immunodeficiency virus that causes AIDS. Due to the clandestine nature of illicit drug use, members of this target population are very difficult to locate. One option rejected by Watters and Biernacki was use of a convenience sample of injecting drug users enrolled in methadone maintenance treatment programs. While this option would have provided easy access to some of the target population, there was reason to believe that persons in methadone clinics were a biased sample of drug users who differed in many ways from those outside the clinics.[9] The sampling strategy they decided on, which they called "targeted sampling," contained several steps: (1) through observation and the use of several sources of information, they determined which neighborhoods had the highest concentrations of drug users and drug-related activity; (2) on the basis of intensive observation and interaction with people in these areas, they identified potential respondents; and (3) they established field sites in each area and developed specific plans for recruitment. During the course of the study, Watters and Biernacki were careful to select adequate numbers of particular groups of respondents and were sensitive to the effects of time and field site location on recruitment.

Combined Probability and Nonprobability Sampling

To reduce the costs associated with probability sampling, probability and nonprobability sampling may be combined in one design. This is possible whenever sampling is carried out in stages (Chein, 1981:438–39). Cases can be selected according to nonprobability procedures at one stage and by probability procedures at another. Two combinations are common in social research. The first involves multistage probability sampling of clusters, with a quota sampling of individual respondents at the final stage. Thus, one might "select a probability sample of counties in a state; within each of these counties, a probability sample of neighborhoods; and within each of the selected neighborhoods, a quota sample controlled for, say, age and sex" (Chein, 1981:438). While sometimes called "probability sampling with quotas," such a design produces, strictly speaking, a *non*probability (quota) sample—with geographic restrictions that sharply reduce interviewer selection bias.

The second combination reverses the first strategy by drawing a probability sample of cases within a nonprobability (purposive) sample of areas. For example, in their study of maternal leave policies, Kelly and Dobbin (1999) collected information on the history of personnel policies from 279 organizations. All the establishments were located in three states: California, New Jersey, and Virginia. Within each state, they randomly selected organizations from thirteen sectors or types of establishments with fifty or more workers; however, the states themselves formed a purposive national sample, carefully "selected for their varying legal environments." With this design, mathematical sampling theory applies to inferences from each probability sample to its respective sector and state population, while more risky investigator judgment must be used in making generalizations to all organizations nationwide.

Referral Sampling

Another set of techniques has been developed for sampling target populations that comprise small subgroups of the larger population (see Sudman and Kalton, 1986; Sudman, Sirken, and Cowan, 1988): for example, users of a particular service or product, members of a minority ethnic group that is widely dispersed in a larger population, laid-off workers, or persons with AIDS. If resources are plentiful and members of the target population are readily identifiable, probability sampling can be carried out simply by choosing a large enough fraction of the larger population and screening for members of the target population. However, extensive screening is very costly and time-consuming, and it is possible to miss persons because they are incorrectly identified.

Often, the sampling technique devised to handle these problems utilizes some form of referral, wherein respondents who are initially contacted are asked to supply the names and addresses of members of the target population. In **network sampling**, respondents initially contacted in screening a probability sample of the larger population are asked to identify members of the target population who are socially linked to the respondent (e.g., relative, neighbor, co-worker, friend). In a study of male Vietnam veterans, for instance, George Rothbart, Michelle Fine, and Seymour Sudman (1982) first selected respondents by screening the larger population for eligible persons. Each respondent to a screening call was then asked if he or she had a son, brother, or nephew who was a veteran. The sample of 1011 consisted of 535 veterans who were directly contacted and 476 veteran kin who were nominated in the initial screening process. Since veterans living within initially contacted house-

holds comprised only 6.2 percent of the larger population, a sample of 1011 would have required 16,439 interviews without network sampling. With network sampling, however, the number of screening interviews was slightly more than half this total—8698. Thus, the major advantage of this technique is a reduction in screening costs. On the other hand, the final probability sample creates unequal selection probabilities that increase sampling error and that must be taken into account in the analysis, and failures by informants to identify members of the target population may result in bias.

Another referral technique, called **snowball sampling**, uses a process of chain referral: When members of the target population are located, they are asked to provide names and addresses of other members of the target population, who are then contacted and asked to name others, and so on. A basic assumption of snowball sampling is that members of the target population often know each other. This technique has been used to create sampling frames (Sudman and Kalton, 1986) and is sometimes associated with probability sampling (Goodman, 1961), but most applications involve nonprobability methods of selection. For example, D. Michael Lindsay (2008) used snowball sampling to examine the role of public leaders and elite actors in the American evangelicalism movement. Initially, he interviewed 157 religious and nonprofit leaders from a diverse group of organizations within the movement; then, at the end of each interview, he asked these informants to identify other leaders who shared their religious commitments. Thus, with the help of the early informants, he contacted and interviewed a total sample of 360 elites.

Patrick Biernacki and Dan Waldorf (1981:144) note that snowball sampling is particularly applicable to studies of deviant behavior, where "moral, legal, or social sensitivities surrounding the behavior in question . . . pose some serious problems for locating and contacting potential respondents." In these studies, members of the target population are often socially invisible by virtue of their illicit, clandestine activities. Their characteristics, therefore, are unknown, and drawing a probability sample is virtually impossible. Often the best that one can do is to use all available means to find eligible respondents and start referral chains. The quality of the sample ultimately depends on the researcher's ability to develop initial contacts and referral chains that represent a range of characteristics in the target population.

Factors Affecting Choice of Sampling Design

Now that we have examined the basic sampling designs, let us look more closely at some of the factors that enter into the researcher's selection of an appropriate design. Which sampling design is adopted depends on several preliminary considerations:

1. What is the stage of research?
2. How will the data be used?
3. What are the available resources for drawing the sample?
4. How will the data be collected?

As you will see, these considerations are interrelated and vary in their relative importance in affecting the choice of a sampling design.

Stage of Research and Data Use

The first two considerations—stage of research and data use—dictate how accurate the sample must be as a description of the population (Sudman, 1976). Accuracy is least important in the exploratory phases of research, when the goal is to discover interesting patterns and generate hypotheses for later study. In fact, under these circumstances, generalizing to a specified population and estimating sample precision are usually unimportant or irrelevant. The researcher often must rely on his or her judgment in drawing a sample and may end up selecting cases for such reasons as availability or willingness to participate in the research; therefore, convenience or purposive sampling is appropriate.

Accuracy would appear to be most important in large-scale fact-finding studies that provide input for major policy decisions. Perhaps the best example of this is the aforementioned Current Population Survey (CPS). The CPS is the only up-to-date source of information on a number of important characteristics of the total population—providing, for example, monthly estimates of unemployment and employment. As Seymour Sudman (1976:3) notes, "major government economic and welfare programs are influenced by changes of a few tenths of a percent in CPS data from month to month." Consequently, a very large (approximately 50,000) and carefully controlled probability sample is necessary to guarantee a high degree of precision.

Obviously, the desired level of precision in most studies lies somewhere between exploratory data gathering and the CPS. The important point here is that the researcher must have some sense of how good the sample must be to meet the objectives of the research. For example, if you need to assess opinions about student government among students at your school for the purpose of documenting support for government reforms, a haphazard, poor-quality sample would be inappropriate. You would need to make sure that the sample is adequate for the task of making reasonably accurate generalizations about the student population. On the other hand, if you merely wanted to get some sense of the variability in opinions about student government for a class project, a small-scale convenience sample would suffice.

Available Resources

If sample accuracy were the sole criterion, selecting an appropriate sampling design would be a simple matter of using the procedure that yields the smallest margin of error. However, accuracy ultimately must be balanced against cost. Available resources such as time, money, materials, and personnel place limitations on how cases can be selected. The primary reason for multistage cluster sampling, as we have seen, is that it reduces time and expense due to interviewer travel and the compilation of necessary case listings. Therefore, some form of this design is recommended for all large-scale surveys. Because it requires less time, systematic sampling sometimes is

used in lieu of simple random sampling. And the only viable rationale for quota sampling similarly rests on critical savings in time and expense in comparison with probability sampling.

For the student doing unfunded research, such comparisons probably seem irrelevant. The underlying message, however, is that one should make efficient use of available resources in designing one's sample. The sampling design should be appropriate for available resources, but lack of funds should not be used as an excuse for drawing a haphazard, low-quality sample. As a general rule, one should define the target population to fit the scope of the study; if the study is small enough, a well-designed and executed sample is possible with very limited resources.

Method of Data Collection

A common impression of sampling is that it is only a tool of survey researchers. Social scientists who employ the other major forms of data collection (experimentation, field research, and documentary research) have been seen as having a peripheral interest in sampling. The truth is, however, that all researchers must deal with case selection. The four data-collection strategies simply handle the issue of sampling in different ways. As you will see in subsequent chapters, convenience sampling is the rule in experiments which tend to use small samples drawn from readily available populations. Some form of probability sampling is found in most surveys, purposive or judgmental sampling typifies field research, and—to the extent that the population is definable—probability sampling often is used in the analysis of available data.

Factors Determining Sample Size

Having carefully defined the target population, obtained a good sampling frame, and come up with an appropriate sampling design, the researcher finally must decide on an adequate sample size. As with the choice of a sampling design, several interrelated factors affect the decision about sample size: (1) heterogeneity of the population, (2) desired precision, (3) type of sampling design, (4) available resources, and (5) number of breakdowns planned in data analysis. The most sophisticated research applications enter these factors into mathematical equations for determining sample size. Such technical matters are, of course, largely beyond the scope of this book. Still, we think that understanding some of the statistical sampling theory underlying sample size considerations is important. The few mathematical principles that we introduce in the following discussion apply directly to simple random sampling. However, they also apply, in a more complex fashion than we discuss, to other forms of probability sampling and, in both a logical and an intuitive sense, to nonprobability sampling designs.

Population Heterogeneity

Heterogeneity refers to the degree of dissimilarity (and, conversely, homogeneity to the degree of similarity) among cases with respect to a particular characteristic. In general, the more heterogeneous the population with respect to the characteristic being studied, the more cases required to yield a reliable sample estimate. The logic of this principle can be seen by considering the extremes: If all cases were exactly alike, a sample of one case would suffice; and if no two cases were alike on a given characteristic, only a complete census could satisfactorily represent the population.

The best statistical measure of population heterogeneity for a quantitative variable is the **standard deviation**, conventionally symbolized by the Greek letter sigma (σ). There is a direct link between this measure and the concept of standard error, introduced earlier in our discussion of probability sampling theory. The standard error, recall, indicates the degree of error in a sample estimate (the "average" amount by which a sample estimate deviates from the population value it estimates). The formula used to calculate the standard error is $\sigma/\sqrt{N}$, that is, the standard deviation divided by the square root of the sample size N. From this formula one can see that the standard error is directly related to the heterogeneity of the population as measured by σ and is inversely related to sample size. Thus, the greater the heterogeneity of the population, the larger the sample necessary to achieve a given level of precision.

KEY POINT

As heterogeneity, or variation, of a variable increases, the standard error increases.

Desired Precision

Technically speaking, *precision* refers to the degree of variability or error in a sample estimate, hence to the standard error. Intuitively, however, the concept of "precision" is perhaps best conveyed by relating it to the size of the confidence interval used to estimate a population value. Thus, it is more precise to say that the average number of drinks consumed is likely to fall between 2.0 and 4.0 than to say that the average is likely to fall between 1.0 and 5.0. For a given confidence level (say, 95 percent), the size of the confidence interval is directly related to the standard error; that is, the smaller the standard error, the smaller the confidence interval and the *more* precise the sample estimate. Of course, the larger the sample, the smaller the standard error. Therefore, it follows that the larger the sample, the greater the precision of the sample estimate.

Two facts about this relationship are especially noteworthy because they defy intuition to a certain extent. First, ordinarily it is the absolute size of the sample rather than the proportion of the population sampled that determines precision. As long as the population is relatively large, the proportion of the population sampled has a negligible effect on precision. For example, in 2000 the population of Vermont was a little over 600,000 and the population of Massachusetts was over 6 million. Now, if one were to take a simple random sample of 2000 in each of these states, the sample proportion of the total population would be about 1 of every 300 persons in Vermont and about 1 of every 3000 persons in Massachusetts. Yet, based on these samples, an estimate (say, of average income) would be just as precise for Massachusetts as for Vermont.

We can get a mathematical understanding of this relationship by re-examining the formula for the standard error. Although the formula given above shows that the standard error is determined only by the standard deviation and sample size, this formula actually applies to populations of theoretically infinite size. For finite populations, the formula should be multiplied by a correction factor equal to $\sqrt{1-f}$, where f is the **sampling fraction**, or proportion of the population included in the sample (Kish, 1965:43–44). Notice, however, that if the sampling fraction is very near zero, then the correction factor becomes $\sqrt{1}$, or 1, which has no effect on the standard error. In most practical examples, the population is so much larger than the sample that f is extremely small—near zero—and the correction factor can be ignored. Only with small populations would we expect f to be much larger than zero; however, for a small population a small f implies a small sample size, and it is the latter that has the greater impact on the standard error.

Although precision is governed primarily by the absolute numerical size of the sample rather than by the proportion of the population sampled, the sample need not be enormous in size to yield very precise results. This is the second crucial fact about precision and sample size. The sampling error tends to be quite small for a sample of size 2000 to 3000, and increasing the sample size beyond this number decreases the error by so little that it usually is not worth the additional cost.

The mathematical explanation for this once again can be found in the standard error formula. Notice that the standard error goes down as the *square root* of the sample size goes up. Because of the square root function, each time we wish to decrease the standard error by one-half we must increase the sample size fourfold. At this rate, the precision gained with increased sample size reaches a point of minute, diminishing returns after a few thousand cases. Consider, for example, how sample size affects the standard error of a percentage, such as the estimated percentage of the vote a candidate will receive in an election. With two candidates and an evenly split vote, the standard error will be 5.0 percent for a sample size of 100. This error decreases to 2.5 percent when the sample reaches 400 and to 1.0 percent when the sample reaches 2500, as Table 6.5 shows. To get the error down to 0.5 percent would require a sample size of 10,000.

The reader should now begin to understand how election forecasters can make accurate predictions using samples of a few thousand out of millions of voters. The

TABLE 6.5. Standard Error of a Percentage of 50 Percent, Broken Down by Sample Size*

Sample size	*Standard error (percent)*
100	5.0
400	2.5
2500	1.0
10,000	0.5

*Standard errors are smaller for percentages greater than or less than 50 percent.

KEY POINT

The margin of sampling error in a sample estimate nearly always depends on the absolute size of the sample and *not* on the size of the sample relative to the population.

huge size of the population has no effect on the precision of sample estimates. And a sample size of 2000 to 3000 is large enough to predict accurately all but the closest elections.

We hasten to add that 2000 to 3000 should not be regarded as *the* standard size for reliable sample results. Not only does sample size depend on factors other than precision, but necessary levels of precision also vary widely from one study to the next. At one extreme, the CPS requires a sample of some 50,000 housing units, a sample "large enough so that the sampling errors of the total estimates of unemployment are only about 0.1%" (Sudman, 1976:3). At the other extreme, thirty cases generally is regarded as minimally adequate for statistical data analysis, although most social researchers would probably recommend at least one hundred.

Sampling Design

The type of sampling design also affects decisions about sample size. Recall that one way of increasing precision other than selecting a larger sample is to use stratified, rather than simple random sampling. In other words, for the same level of precision, a stratified random sample requires fewer cases than a simple random sample. A cluster sample, on the other hand, requires a somewhat larger number of cases for precision equal to that of a simple random sample. A nonmathematical explanation for these comparisons rests on the concept of "heterogeneity."

Remember that stratified random sampling provides greater sampling efficiency (or precision) when the stratifying variable is related to the variable one is estimating. If it is related, much of the variation or heterogeneity in the estimated variable resides in differences between the strata, and each stratum tends to be relatively homogeneous. With respect to alcohol consumption, for example, there is much less variability among men or among women than between men and women. By sampling every stratum, the effect of stratified random sampling is to eliminate this *between* source of heterogeneity, leaving only the variation within the strata. And because the strata are relatively homogeneous, the standard error of each stratum is small in comparison with the population as a whole, and stratified sampling produces a smaller margin of error than a simple random sample of the same size.

Although cluster sampling has the advantage of low cost, it lacks efficiency in precision. In two-stage cluster sampling, for example, there is variability both between and within the clusters. By taking a sample of clusters, the "between source" of heterogeneity that was eliminated in stratified sampling is still present. If the variability between clusters tends to be large compared with the variability within the clusters, then the sampling error could be considerable, depending on how many clusters are selected. (For this reason, we want clusters to be heterogeneous compared with differences between the clusters—just the opposite of the strategy in stratified sampling.) Furthermore, each stage in a multistage cluster design contributes a source of vari-

ability or error to the total sample; consequently, the more stages, the larger the total sampling error tends to be and the larger the sample required for a given level of precision.

Available Resources

Each individual case requires an expenditure of available resources. Consequently, at some point, cost must enter into the equation for determining sample size. When a fixed amount of money and/or time has been allocated for a project, Sudman (1976:88–89) offers the following rule of thumb for survey research: Allocate one-half of both money and time to data collection (and the other half to data analysis). Once the data-collection procedure is specified, the sample size can be determined on a time- and cost-per-case basis. That is, the number of cases will be equal to the total time for data collection divided by the time per case or the total funds for data collection divided by the cost per case.

KEY POINT

The two overriding factors in choosing a sampling design and an appropriate sample size are how precise sample estimates need to be and the cost of sampling.

Number of Breakdowns Planned

The number of variables and variable categories into which the data are to be grouped and analyzed also must be taken into account in determining sample size. In general, the more breakdowns planned in the analysis—the more complex the relationships under investigation or the more distinct subcategories of separate interest—the larger the sample must be. Consider, for example, what happens when a sample of 1000 is divided into males/females, then into blacks/whites, then into people over/under 18 years of age, then into urban/rural residence. If we were interested in describing rural black young men, we might find that a breakdown of the 1000 cases looks something like this:

Sample	1000
Males	489
Black males	60
Young black males	20
Rural young black males	4

With just four cases available, the sample clearly would be too small for a reliable analysis of this particular subgroup. To avoid this kind of problem, it is important to estimate the number of breakdowns to be made during data analysis and to make sure that the total sample size will provide enough cases in each subcategory. If the requisite sample size turns out to be rather large, it might be more efficient to sample the relevant subcategories separately, as in stratified and quota sampling.

Final Notes on Sampling Errors and Generalizability

In sections on probability sampling and sample size considerations, we spoke repeatedly of sampling error—the deviation of a sample estimate from the true population value. The sampling error referred to in our discussions is random error, produced by the random selection of elements. Though unavoidable in sampling, this error can be estimated and reduced by increasing the size or efficiency of one's sample. There is, however, another type of sampling error, called **sample bias**,[10] which is nonrandom, difficult to detect, and often much more damaging to sample accuracy.

In probability sampling designs, the two most common sources of sample bias are **coverage error** due to incomplete sampling frames and **nonresponse bias** due to incomplete data collection. As an example of coverage error, we mentioned earlier that telephone directories provide inadequate sampling frames to the extent that they exclude the poor, who cannot afford telephones, and the more wealthy, who tend to have unlisted numbers. The problem of nonresponse bias arises when, through refusals to cooperate, unreturned questionnaires, missing records, or some other means, the sample turns out to be a fraction of the number of cases originally selected for observation. The crux of this problem is that nonobservations tend to differ in systematic ways from observations. Thus, in surveys, mail surveys in particular, highly educated respondents are more likely to cooperate than poorly educated ones. Also, those who feel most strongly about the topics or issues of a study are more likely to respond than those in the middle (see Box 6.1).

Obviously, the researcher should try to reduce the likelihood of these biases. For coverage error, researchers should assess how the sampling frame may differ from the target population and consider how such differences may affect survey estimates. To reduce nonresponse bias, the conventional wisdom has been to increase response rates because it is assumed that nonresponse bias increases as the response rate declines. Robert Groves (2006) has shown, however, that the correlation between response rates and nonresponse bias is relatively weak and that this bias depends primarily on "how strongly correlated the survey variable of interest is with . . . the likelihood of responding" (p. 670). This makes it incumbent on the survey researcher to consider variables that might be related to responding and to seek data that can be used to estimate and reduce nonresponse effects. In a study of voluntarism, for example, knowing that women are more likely to volunteer than men, you should examine the response rates of men and women and adjust survey estimates based on these rates. Otherwise, if men are overrepresented among nonrespondents, estimated rates of voluntarism will be too high.

Although Groves cautions against "the blind pursuit of high response rates in probability samples" (p. 668), high response rates continue to be valued (Dillman, 2007). Some researchers and sponsors have more confidence in the representativeness of a survey with a high response rate than one with a very low rate. Two trends indicate, however, that survey researchers may have to shift the emphasis from obtaining a very high response rate to assessing nonresponse bias and adjusting survey estimates based on variables that are likely to be related to the tendency to respond. First, as we note in chapter 9, response rates in most surveys are declining. Second, some procedures used to increase response rates, such as multiple contacts

and recontacting people who initially refuse to participate, have raised ethical concerns among institutional review boards (Dillman, 2007:28).

The underlying issue addressed in this chapter is generalizability. Advances in knowledge depend on our ability to generalize. With this in mind, the reader should be aware of two limitations to sampling in social research. First, even though social scientists would like to make the broadest possible generalizations, for reasons of convenience and economy they often study local or special populations. Investigators in the Worcester aging and religion study, for example, were not interested specifically in religion among older Worcester residents, which turned out to be their target population, but rather in religion and aging in the United States and beyond. They chose Worcester because that was the city in which the investigators were teaching. Drawing a probability sample made their results generalizable (within an estimable margin of error) to all Worcester residents 65 years and older. But inferences beyond this population are based on little more than intuition.

The second limitation concerns generalizing from samples to populations. As we have seen, the manner of selecting cases determines to whom and with what level of precision research results may be applied. Sample generalizability is the primary generalization issue addressed in social science research, especially survey studies, which are often called "sample surveys" because the sampling of units is such an integral part of the research design. Survey researchers are also largely responsible for the development of many of the sampling techniques discussed in this chapter. So it is no wonder that most of our examples came from survey studies and involved people as units.

However, the issue of generalizability is not limited to units of analysis, nor is the sampling of units solely the province of survey research. Generalizability pertains to all features of research, including the time of the study, the research setting, and the operational definitions. We might ask, for example, whether the same results would be obtained if a different set of questions were asked or if the study were conducted in a different place or at a different time of day or year. The social scientific ideal is to establish theoretical propositions that hold under all conditions—that do not pertain to any particular population, time, or setting. With this goal in mind, social researchers have no particular target population as the focus of their research. To establish the broadest possible generalizations, researchers use a variety of techniques in addition to random sampling, which we discuss in later chapters. These include replication (chapter 7), theoretical sampling (chapter 11), and triangulation (chapter 13).

Summary

Sampling is the process of selecting a subset of cases in order to draw conclusions about the entire set. Sampling is unavoidable given the scientific goal of generalization, and it requires special attention in social research given the inherent variability of social units of analysis. Sometimes sampling can yield more accurate information than "complete" enumerations of the population. Yet, even if this were never true, we would still sample because it is usually impossible for practical reasons to examine all cases, and observing a sample of cases is simply more efficient

anyway: It saves time and money and can be as accurate as research purposes demand.

Prior to sampling, one must select the unit of analysis. Sampling then begins with a description of the target population, the collection of units about which one wishes to generalize. For sampling to be feasible, the target population should be defined by objective criteria that clearly indicate its limits of inclusion. Even then, there is often an imperfect fit between the target population and the sampling frame, which consists of either (1) a list of cases from which the sample is actually selected or (2) a rule defining membership that provides a basis for case selection.

The procedure for selecting a sample is called the "sampling design." The major distinction among designs is between probability and nonprobability sampling. Probability sampling is based on a process of random selection, which gives each case in the population an equal chance of being included in the sample. This process, which is absent from nonprobability sampling, eliminates investigator bias in selecting cases and permits the application of mathematical probability theory for estimating sample accuracy.

Probability sampling designs include simple random sampling, stratified random sampling, cluster sampling, and systematic sampling. In simple random sampling, random selection from the entire population makes it equally possible to draw any combination of cases. The distribution of a statistical property for all possible combinations of random samples of a fixed size is called the "sampling distribution." A given sample statistic differs from the estimated population value by an amount known as the "sampling error"; an index of the size of such errors in a given sampling distribution is called the "standard error." In general, the larger the sample, the smaller the standard error. To take into account random sampling error, researchers use interval estimates, called "confidence intervals," in making inferences from a sample statistic to its population parameter.

In stratified random sampling, the population is divided into strata and independent random samples are drawn from each stratum. The number of cases in each stratum of the sample may be either proportionate or disproportionate to the number of cases in the population. Stratifying improves sampling efficiency, provided that the stratifying variable is related to the variable under study and that the cost of stratifying is low compared with the cost of increasing sample size. It also may be used to increase the number of cases in certain variable categories that would prove too small if a simple random sample were drawn.

In cluster sampling, the population is divided into natural groupings or areas, called "clusters," and a random sample of clusters is drawn. When this is done in stages, as it usually is, moving from larger to smaller clusters, it is referred to as "multistage sampling." Clustering reduces the costs of interviewer travel and sampling frame construction, but it does so at a loss of sample precision. Systematic sampling, which often provides a reasonable approximation to simple random sampling, consists of selecting cases from an available list at a fixed interval after a random start.

Nonprobability sampling designs include convenience, purposive, and quota samples. Convenience sampling is a rubric for various nonrandom and unsystematic processes of case selection that offer no basis for generalizing. Purposive sam-

pling involves the careful selection of typical cases or of cases that represent relevant dimensions of the population. Quota sampling allocates quotas of cases for various strata and then allows for the nonrandom selection of cases to fill the quotas. Although once used in national surveys to reduce costs, it has been replaced by probability sampling because it tends to introduce biases in case selection.

Other sampling strategies have been devised when research goals require rigorous, systematic methods of case selection but conventional probability sampling methods are not feasible. If selection can be done in stages, using probability sampling at one stage and nonprobability sampling at another can reduce costs. For rare target populations, sampling may entail a process of referral. Network sampling asks respondents to identify members of the target population who are socially linked to the respondent in a specified way. Snowball sampling uses a chain of referrals, with initial contacts asked to name other members of the target population with particular characteristics, who are asked in turn to name others, and so on.

Factors affecting the choice of a sampling design include (1) the stage of research and data use, with research in later stages intended to provide accurate population descriptions requiring the most sophisticated probability sampling designs; (2) available resources such as time, money, and personnel; and (3) the method of data collection. In general, the more heterogeneous the population, the greater the desired precision and available resources; and the larger the number of breakdowns planned during the data analysis, the larger the sample should be. Holding sample size constant, one will generally get the greatest precision with a stratified sample, followed by a simple random sample, and then a cluster sample. Finally, the quality of a sample ultimately depends not only on sample variability due to random selection but also on sample bias in the form of coverage error and nonresponse error.

Key Terms

- population
- sample
- target population
- sampling frame
- sampling design
- probability sampling
  - simple random sampling
  - stratified random sampling
  - cluster sampling
  - systematic sampling
- random selection
  - parameter
  - statistic
  - sampling distribution
  - sampling error
  - standard error
- confidence interval
- stratum
- disproportionate stratified sampling
- multistage sampling
- primary sampling unit
- probability proportionate to size sampling
- sampling interval
- nonprobability sampling
  - convenience sampling
  - purposive sampling
  - quota sampling
- network sampling
- snowball sampling
- heterogeneity
- standard deviation
- sampling fraction
- sample bias
- coverage error
- nonresponse bias

Exercises

1. This exercise will help familiarize you with some types of probability sampling. First, using the A section of the latest issue of your campus telephone directory as a starting point, draw up a sampling frame composed of the first fifty names listed.
   a. *Simple random sample.* Select a random sample of ten names using the Research Randomizer (http://www.randomizer.org/). List all the random numbers you select, and then list the ten names in your sample. Repeat this procedure in drawing a random sample of five names.
   b. *Stratified systematic sample.* Divide the names in your sampling frame into strata on the basis of gender. Now select a one out of five systematic sample within each stratum. Begin with a random start, and indicate the random number in each stratum with which you began. List the names that you obtain.

2. Each election year, the press relies on exit polls to forecast and analyze election outcomes. Exit polls are surveys of voters immediately after they have cast their votes at their polling places. The polling places (or precincts) are randomly selected to represent each state and, for national polls, the nation as a whole. Interviewers give selected voters a questionnaire that takes only a couple minutes to complete. It asks for whom they voted, about important issues, and about demographics such as gender, age, and race. Participation is voluntary and anonymous. Interviewing starts when the polls open and continues throughout the day until about an hour before they close at night.

   In one national election exit poll, voters were selected in the following way. Within each state, a complete list of precincts was obtained. The precincts were divided up into counties, and within each county precincts were selected randomly so that the odds of being selected were proportionate to the number of people who typically voted in that precinct. At each selected precinct, one or sometimes two interviewers stood outside and attempted to randomly select roughly one hundred voters during the day as they exited from voting. The interviewers accomplished this task by counting voters as they left the polling place and selecting every voter at a specific interval (say, every third or fifth voter). The interval was chosen so that the approximately one hundred interviews were spread evenly over the course of the day.
   a. What is the overall sampling design? (Be specific. This is a complex design, involving more than one sampling method.)
   b. What sampling frame was used in the design?
   c. Is stratification incorporated into the design? If so, what was the stratifying variable?
   d. What sampling method was used to select voters within each selected precinct?
   e. Some journalists and commentators would like to have had mid-day numbers or estimates from the exit polls. Why would the mid-day numbers have a much larger margin of error than when the polls were completed at the end of the day? Give *two* possible reasons, one involving random sampling error and the other involving sampling bias.
   f. Evaluate the sampling strategy with respect to coverage error; assume that the target population is all voters in the national election.

3. The General Social Survey (GSS), which we describe throughout chapter 9, consists of a personal interview survey of adults (18 and over) living in households in the United States. Read about the sampling design of the GSS in Appendix A of the GSS Codebook, which can be found online (http://publicdata.norc.org:41000/gss/Documents/Codebook/A.pdf). Briefly describe the overall design of the 2006 GSS. How has the design changed from 1972 to the present? When did the GSS switch from a "block quota" sample to a full probability sample?

4. Suppose the governor hires you to conduct a survey of state firefighters. The state plans to launch an advertising campaign to hire more firefighters. The goals of the survey are to describe the state population of firefighters in terms of basic demographic characteristics such as age, race, religion, and social class and to determine what leads people to become firefighters. You want to be able to estimate margins of error, and you also would like to have reliable information on female firefighters, who are estimated to make up only 5 percent of firefighters statewide.

   a. What is your target population for this survey?
   b. Assume you decide to do a personal interview survey. No complete list of all firefighters in the state is available; however, you have found a Web site with a complete directory of state fire departments. How would you go about selecting a probability sample of *firefighters*? In other words, what sampling design would you use? Keep in mind the goals of the survey. Be specific and be sure to fully describe how you would draw the sample.
   c. Now assume that you are going to do a mail questionnaire survey and that you were able to obtain a complete list of firefighters working in the state. However, the list contains no information other than each firefighter's name, address, and length of service. What sampling design would you use? Be specific and be sure to fully describe how you would draw the sample.

5. Because of your developing methodological expertise, someone who is about to conduct a random sample survey asks you, "What proportion of the population should I sample to give me adequate precision?" How would you respond to this question?

Notes

1. The U.S. Constitution mandates a complete enumeration of the population every 10 years. It therefore would require a constitutional amendment to replace the Census with a sample survey of the U.S. population. The primary purpose of the Census is to reapportion seats in the U.S. House of Representatives. Therefore, it would require a very large sample to provide detailed information on small geographic areas, as the Census does.

2. The lottery method also has proven inadequate on occasion due to insufficient mixing. Two dramatic examples of this problem are the military draft lotteries of 1940 and 1970. Statistical analyses convincingly have shown that both lotteries were biased, apparently due to the failure of physical mixing to achieve randomness (Fienberg, 1971). The 1970 lottery, based on birthdays, was set up in the following way: 366 cylindrical capsules with rounded ends were used; slips of paper with the January dates were inserted in thirty-one capsules and placed in a large, square wooden box; the February dates were inserted and then placed in the box and mixed with the January dates, and this procedure was repeated for each of the remaining months, in turn; the box then was shut and shaken several times; finally, the 366 capsules were poured from the box into a large bowl, from which they were drawn. When the lottery drawing sequence was analyzed, it was shown that the sequence was not random but, rather, a reflection of "the order in which capsules were placed in the wooden box during the initial mixing procedure" (Fienberg, 1971:260). As a result, those with birthdays in the later months of the year had lower lottery numbers, hence were more likely to be drafted, than those with birthdays in the early months of the year.

3. This example is modeled after presentations by Morris James Slonim (1957, 1960) and Isidor Chein (1981).

4. If each one of a set of events is equally likely to occur, then the probability of any one event occurring is defined as 1 divided by the number of events. Assuming a perfect coin, for example, we have two events—a head and a tail—with the probability of one event (say, a head) in a single flip equal to 1/2.

5. Basically, this is accomplished by a weighting procedure that compensates for oversampling in some strata. The necessary mathematical formulas can be found in Kish (1965).

6. The following example illustrates how the probabilities are equal. Suppose that the total number of students in the population is 500,000 and that 100 schools are selected. The probability of selecting a school A, with 1000 students, is $100 \times 1000/500{,}000$, or .20. The probability of selecting a school B, with 500 students, is $100 \times 500/500{,}000$, or .10. If fifty students are selected from each school, the probability of selecting a student from school A is $.20 \times 50/1000$, or .01, which is the same as the probability of selecting a student from school B ($.10 \times 50/500 = .01$).

7. Another problem with cluster sampling is that, except for rare cases in which clusters are equal in size, the probabilities of case selection may vary widely from cluster to cluster. Consequently, estimates of population characteristics require some rather complex statistical solutions, which are beyond the scope of this book.

8. Richard Appelbaum (1990), who contends that counting the homeless is an intractable problem and that "every effort is likely to produce an undercount," criticized the Rossi street survey on two counts. First, since the homeless were self-identified, there is the strong possibility of response bias—either through avoidance of interviewers, refusals to be interviewed, or unwillingness to admit to being homeless. Second, the survey reached only the visible homeless, which excludes the sizeable number of people in rooming houses, jails, and halfway houses, as well as people riding trains and buses.

9. In spite of the potential bias, studies of noninstitutional hidden populations are rare. Indeed, "classic studies of deviance have depended largely or wholly on populations captured in the nets of institutions of social control" (Watters and Biernacki, 1989:417).

10. These two types of error—random sampling error and sample bias—are analogous to random and systematic measurement error.