5

Measurement

Measurement is the process of assigning numbers or labels to units of analysis to represent conceptual properties. This process should be familiar to readers 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 grade-point average, 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, however, 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 a 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 a 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 moves 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, the 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, whereas 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, although 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 a 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, a 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 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 the 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 similar cake.

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.

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.

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 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 [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.")

BOX 5.1 Operationalizing the Concept of "Religiosity"

The 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" (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, whereas nonmembers are nonreligious. The General Social Survey (GSS) (T. W. Smith, Marsden, and Hout, 2016) 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 although church attendance declined with age among persons older than 50, nonchurch participation measures such as Bible reading and listening to religious radio programs increased with age.
3. An 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

(continued)

(continued)

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.

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 illustrated in a study of job discrimination against openly gay men in the United States. To investigate the level of discrimination, Andras Tilcsik (2011) sent fictitious resumes to 1769 postings of white-collar, entry-level job openings. All the resumes described a male, graduating college senior having similar traits, qualifications, and experiences with one exception: an experience signaling a gay sexual orientation. The independent variable, sexual orientation, was *manipulated* by indicating in one-half of the resumes that the applicant was the treasurer of a campus gay and lesbian organization ("Gay and Lesbian Alliance") and in the other half that the applicant was the treasurer of a left-wing campus organization ("Progressive and Socialist Alliance"). To operationalize the dependent variable, employment discrimination, Tilcsik used a *measured* definition: whether the applicant was invited for an interview. Among other findings, the study showed that heterosexual applicants had a greater chance of being invited for an interview than gay applicants (11.5% versus 7.2%).

With respect to the manipulation of sexual orientation, Tilcsik points out that being a member of a left-wing campus organization is an effective control because participation in a gay organization tends to be associated with progressive or liberal political views. In addition, the position of treasurer entailed financial and managerial skills that were relevant to the job. To emphasize this point, the resume item listed four responsibilities associated with the position of treasurer; for example, "managed budget and all corresponding accounts" and "wrote grant proposals, increasing grant revenue by 25% from previous years." 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**: 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 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 amount of time students reported spending on homework each week in school and out of school. 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 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 sophisticated statistical techniques that are beyond the scope of this book.

Because of the difficulty of creating scales and because there are literally thousands of scales and indexes, we recommend that a 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.

KEY POINT

Many operational definitions (or indicators) are possible, but rarely does any one indicator 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 a 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 14 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." (945). Then they were asked to rank the cards from the most to least attractive and to indicate whether they would not want to move into any of the neighborhoods.

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

Tilcsik measured employment discrimination by direct observation of behavior: He recorded whether employers contacted applicants to set up an interview. Lesley Joy and her colleagues (Joy, Kimball, and Zabrack, 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.

In addition to 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. censuses, John Logan and Brian Stults (2011) found that black–white residential segregation in metropolitan areas peaked around 1960 or 1970 and has declined slowly but steadily since then, with segregation historically lowest and declines greatest in areas where blacks constitute less than 5 percent of the population. "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 50 metropolises with the largest black populations in 2010, the index of dissimilarity ranged from 80 for Detroit (the most segregated) to 36 for Las Vegas (the least segregated). This means that in Detroit either 80 percent of whites or 80 percent of blacks would have to move to eliminate residential segregation and thereby reduce the index to zero.

Another example of an operational definition derived from archival records comes from a study of parental role portrayals in twentieth-century children's picture books. Amy DeWitt, Cynthia Cready, and Rudy Seward (2013) analyzed parental roles that appeared in 300 "easy" storybooks selected to represent six time periods, from pre-1960 to the year 2000. Based on illustrations and narrative text, they determined whether mother and father characters were depicted in five parental roles: companion, disciplinarian, caregiver, nurturer, and provider. Each role was operationally defined in terms of specific behaviors; for example, nurturing behaviors included "physically or verbally express affection for the child," "verbally encourage the child," and "comfort the child"; caregiving behaviors included "prepare meals for and/or feed the child," "clean the child," and "pick out clothes and/or dress the child." Unexpectedly, they found little change in parental roles over time as traditional roles continued to dominate, with mothers acting as nurturers and caregivers and fathers as providers.

Selection of Operational Definitions

Given that you have a concept in mind, how do you decide on an appropriate operational definition? First, 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, in contrast, 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. Although we have said that no operational definition can capture a concept's meaning perfectly or completely, this does not license a 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.

An 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 independent 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 nature and quality of information provided by operational definitions: levels of measurement, reliability, and validity.

Levels of Measurement

Speaking in terms of variables, we can define *measurement* as "the assignment of numbers or labels to units of analysis to represent variable categories." For example, in measuring participation in interscholastic sports, Broh assigned the number 1 to "participated in both

the 10th and 12th grades" and 0 to "did not participate in both grades"; and she assigned a number between 7 and 28 to represent a student's level of self-esteem. However, 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. **Levels of measurement** indicate the various meanings of these numbers, which reflect basic empirical rules for category assignment. 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 a researcher's convenience in collecting and analyzing data. For example, political party preference might be classified as follows:

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

With nominal measurement, the empirical rule for assigning units to categories is that cases placed in the same category must be *equivalent*. Thus, we can say that all 1s share the same political preference and that 1s differ from 2s in their political preference. However, because 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).

Nominal measurement has two characteristics that apply to all levels of measurement: Variable categories 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. The 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 television. In this sense, ordinal measurement is like an 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 (231) illustrates the accuracy of comparative judgments with a familiar example:

> You 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.

On the one hand, 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.

Although social researchers often aim to create interval measures, most of what passes for this level of measurement is only a 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 courses taken in sociology, 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 on the empirical procedures used to operationalize a concept. A field researcher, for example, may obtain an ordinal measure of age 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 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]

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

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

We have seen that for any concept several operational definitions are possible and that creative insight, good judgment, and relevant theory aid in the development of operational definitions. Admittedly, these aids are 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 this 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 child has a Y chromosome and hence will be a

boy or a girl. At one time, many 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 sources, 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 underestimates 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, a person familiar with the test problems and language will always score higher than a 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 a 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. 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).

In addition to this **social desirability effect**, other response tendencies 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.

In Figure 5.1, a target is used to illustrate the relationship among 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 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.

KEY POINT

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

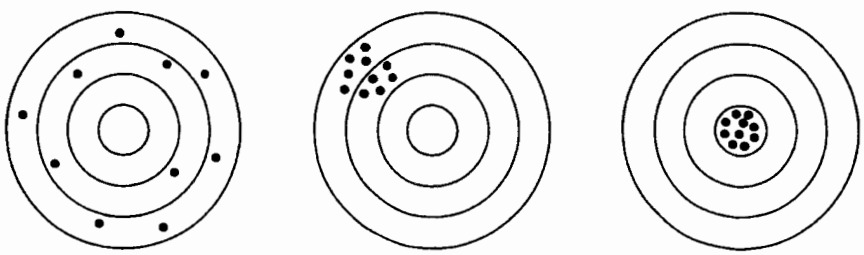

FIGURE 5.1. **Analogy of target to reliability, validity, and measurement error. Adapted from Babbie (1995).**

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 (M. 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).

Although 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 consequently 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. However, 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, a 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? 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 observers or coders using the same instrument or measure obtain 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 parental roles (DeWitt,

Cready, and Seward, 2013), several coders were trained to code role behaviors in a subset of 84 books. When their codes were compared with those of the researcher, the level of agreement, or **intercoder reliability**, was very high, with coefficients well above .90 on all behaviors.[4]

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, although 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.

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.

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 items. 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 retaining 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.

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 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. Although these all have face validity, they are not equally accurate.[5] One would not know which operationalization is most accurate without resorting to a validation method other than face validity.

A more acceptable form of subjective evaluation, **content validity**, concerns the extent to which a measure adequately represents all facets of a concept. 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 difficult for most tests of knowledge. With respect to knowledge of the current chapter, one could list all the major topics and subtopics and then develop test items for each one, making sure that the number of items per topic is 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, an 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 an individual's *current* or *future* standing on the criterion variable. For example, a mental health inventory designed 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 current 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 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? Although this sounds like face validity, the orientation is different; construct validity is based on 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.[6]

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 self-respect. 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" (18). Thus, because clinical observations indicated that depression often accompanies low self-esteem, people who score low on the scale should report more depressive feelings such as unhappiness and gloom and appear more depressed to outside observers. 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.

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-on independent criterion	No clear, single criterion
Degree of validity a matter of:	Predictive accuracy	A consistent pattern of relationships

In construct validation, however, 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 examine 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. M. Rosenberg (1965) validated his self-esteem scale in this way by showing that scores on it were correlated with symptoms of depression 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.

Abundant evidence supports the convergent validity of the Rosenberg Self-Esteem Scale. For example, based on a sample of 9th and 10th graders, David Demo (1985) showed that scores on Rosenberg's scale were positively associated with scores on another popular scale, the Coopersmith Self-Esteem Inventory, and with peer ratings of each individual's self-esteem. Other studies similarly have found positive associations between the Rosenberg Self-Esteem Scale and the Lerner Self-Esteem Scale (Savin-Williams and Jaquish, 1981) and with an indicator of general self-regard (Fleming and Courtney, 1984).

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 13-item Love Scale. A valid measure of love should differentiate love from liking because the two concepts are empirically related but conceptually distinct. Rubin therefore developed a parallel scale of liking. When both the love and the 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 7-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, nonwhites, 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.)

Although construct validation is now considered the model validation procedure for most social measurement, it is not without its problems. It is cumbersome and requires abundant evidence; more important, however, it can lead to inconsistent and equivocal outcomes. On the one hand, 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 a lack of construct validity (Zeller and Carmines, 1980).

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 20-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 20 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

(continued)

(continued)

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 and Fascinating Womanhood. The National Organization for Women (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. Members of NOW had an average score of 91, whereas members of Fascinating Womanhood had an average score of 51.

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.[7] 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. Sometimes a measure is tested or validated by others. At other times, new measures are introduced (perhaps with minimal evidence of reliability or validity) and widely used 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 at "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 die-hard 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" (15).

Summary

- *Measurement* is the process of assigning numerals or labels to units of analysis to represent conceptual properties.
- The measurement 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.
- 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.
- In addition to all the features of lower measurement levels, ratio measurement contains 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.
- 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 of 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 observers or coders who apply the same measure (intercoder reliability).
- 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).
- Criterion-related validation applies to measures (or tests) that are intended to indicate a person's current 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, including 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
systematic measurement error
reactive measurement effect
social desirability effect
random measurement error
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 on 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 because 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. DeWitt, Cready, and Seward (2013) used Cohen's kappa coefficient to calculate interrater reliability. Kappa is thought to be a better measure than simple percent agreement because it takes into account agreement that could occur by chance.
5. 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).
6. As you can see, this process is 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).
7. 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).