Substance abuse is a social problem of remarkable proportions and is related to many forms of crime. For example, more than half of violent offenders incarcerated in State and Federal prisons were under the influence of alcohol when they committed their crimes (Bureau of Justice Statistics 1998). Alcohol also is involved in about half of all fatal traffic crashes, and more than 1 million arrests are made annually for driving under the influence. As many as two-thirds of all persons arrested in urban areas test positive for drugs (Gruenewald et al.}
1997). College presidents rate alcohol abuse as the number-one campus problem (Wechsler et al. 1994), and it is a factor in as many as two-thirds of on-campus sexual assaults (National Institute of Alcohol Abuse and Alcoholism [NIAAA], 1995). All told, the annual costs of prevention and treatment for alcohol and drug abuse exceed $4 billion (Gruenewald et al. 1997).

Whether your goal is to examine the factors related to criminal offending, to deliver useful services, or to design effective social policies, at some point you will probably need to read the research literature on substance abuse and perhaps even design your own study of it. Every time you begin to review or design relevant research, you will have to answer two questions: “What is meant by ‘substance abuse’ in this research?” (which concerns conceptualization) and “How was substance abuse measured?” (which concerns measurement). Both questions must be answered to evaluate the validity of substance abuse research. You cannot make sense of the results of a study until you know how the concepts were defined and measured. Nor are you ready to begin a research project until you have defined your concepts and constructed valid measures of them. Measurement validity is essential to successful research; in fact, without valid measures it is fruitless to attempt to achieve the other two aspects of validity: causal validity (see Chapter 5) and generalizability (see Chapter 4).

In this chapter, we first address the issue of conceptualization, using substance abuse and related concepts as examples. We then focus on measurement, reviewing first how measures of substance abuse have been constructed, using such operations as available data, questions, observations, and less direct and obtrusive measures. Then we explain how to assess the validity and reliability of these measures. The final topic is the level of measurement reflected in different measures. By chapter’s end, you should have a good understanding of measurement, the first of the three legs on which a research project’s validity rests.

CONCEPTS

A May 1997 New York Times article (Johnson 1997) reported that five U.S. colleges were participating in a pilot program to ban alcohol in their fraternities. Moreover, the article claimed that substance-free housing would soon become the norm on U.S. campuses. Do you know what the article means? Some of these concepts—alcohol, colleges, campuses, and pilot program—are widely understood and commonly used. However, do we all have the same thing in mind when we hear these terms? For example, are junior colleges subsumed within the term college? Does the concept of on campus extend to fraternity houses that are not physically on college property? Does substance-free housing mean banning tobacco products as well as alcohol?

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

Concepts such as substance-free housing require an explicit definition before they are used in research because we cannot be certain that all readers will share the same definition. It is even more important to define concepts that are somewhat abstract or unfamiliar. When we refer to concepts such as poverty or social control or strain, we cannot be certain that others know exactly what we mean.
Clarifying the meaning of such concepts does not just benefit those unfamiliar with them; even experts often disagree about their meaning. But we need not avoid using these concepts; we just have to specify clearly what we mean when we use them, and we must expect others to do the same.

**Conceptualization in Practice**

If we are to do an adequate job of conceptualization, we must do more than just think up some definition, any definition, for our concepts. We may need to distinguish subconcepts, or dimensions, of the concept. We also should ask how the concept’s definition fits within the theoretical framework guiding the research and what assumptions underlie this framework.

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

**Defining Youth Gangs**

Do you have a clear image in mind when you hear the term *youth gangs*? Although this is a very ordinary term, social scientists’ attempts to define precisely the concept, youth gang, have not yet succeeded: “Neither gang researchers nor law enforcement agencies can agree on a common definition . . . and a concerted national effort . . . failed to reach a consensus” (Howell 2003:75). Exhibit 3.1 lists a few of the many alternative definitions of youth gangs.

What is the basis of this conceptual difficulty? Howell (2003:27–28) suggests that defining the term *youth gangs* has been difficult for four reasons:

- Youth gangs are not particularly cohesive.
- Individual gangs change their focus over time.
- Many have a “hodgepodge of features,” with diverse members and unclear rules.
- There are many incorrect but popular “myths” about youth gangs.

In addition, youth gangs are only one type of social group, and it is important to define youth gangs in a way that distinguishes them from these other types of groups, for example, childhood play groups, youth subculture groups, delinquent groups, and adult criminal organizations. You can think of “social group” as a broader concept that has multiple dimensions, one of which is youth gangs. In the same way, you can think of substance abuse as a concept with three dimensions: alcohol abuse, drug abuse, and polysubstance abuse. Whenever you define a concept, you need to consider whether the concept is unidimensional or multidimensional. If it is multidimensional, your job of conceptualization is not complete until you have specified the related subconcepts that belong under the umbrella of the larger concept.
Defining Substance Abuse

What observations or images should we associate with the concept substance abuse? Someone leaning against a building with a liquor bottle, barely able to speak coherently? College students drinking heavily at a party? Someone in an Alcoholics Anonymous group drinking one beer? A 10-year-old boy drinking a small glass of wine in an alley? A 10-year-old boy drinking a small glass of wine at the dinner table in France? Do all these images share something in common that we should define as substance abuse for the purposes of a particular research study? Do some of them. Should we take into account cultural differences? Social situations? Physical tolerance for alcohol? Individual standards?

Many researchers now use the definition of substance abuse contained in the Diagnostic and Statistical Manual of Mental Disorders (1994) of the American Psychiatric Association (Mueser et al. 1990): “repeated use of a substance to the extent that it interferes with adequate social, vocational, or self-care functioning” (p. 33). In contrast, substance dependence is defined as “development of tolerance to a substance such that the person requires larger dosages to achieve the same psychoactive effect, and the experience of withdrawal symptoms and craving after a period of abstinence from the substance.” Note that these definitions rely on behavioral and biological criteria rather than social expectations or cultural norms.
We cannot judge the *DSM-IV* definition of substance abuse as correct or incorrect. Each researcher has the right to conceptualize as her or she sees fit. However, we can say that the *DSM-IV* definition of substance abuse is useful, even good, in part because it has been very widely adopted. If we conceptualize substance abuse in the same way that the *DSM-IV* does, many others will share our definition and understand what we are talking about. This definition of substance abuse has two other attractive features: It is stated in clear and precise language that should minimize differences in interpretation and maximize understanding; and it can clearly be distinguished from the more specific concept of substance dependence.

One caution is in order. The definition of any one concept rests on a shared understanding of the terms used in the definition. So if our audience does not already have a shared understanding of terms such as *adequate social functioning*, *self-care functioning*, and *repeated use*, we must also define these terms before we are finished with the process of defining substance abuse.

**Defining Community Policing**

Most of us believe that the police are supposed to prevent crime, not just respond to it after it occurs. However, early research demonstrated that there was little difference in rates of crime between jurisdictions where there was a strong police presence in the form of visible patrols and those jurisdictions where there were virtually no patrols (Bayley 1994). In the 1980s, police began to acknowledge the limitations of basic patrols and seek more effective prevention strategies. David Bayley (1994) describes three insights that helped shape the new crime-prevention strategies that emerged: (1) that police could not prevent crime without community help, (2) that police must do more than react to crime, and (3) that simply patrolling an area is too passive. Most people label the crime-prevention efforts that emerged from these insights as “community policing.” Unfortunately, there is still a large amount of variation and disagreement among both practitioners and researchers alike on what exactly community policing really is. As Bayley describes,

> I have heard police describe community policing as being foot patrol, aggressive enforcement of minor ordinances, electronic surveillance of shopping malls, enhanced traffic enforcement, and any police action that instills public confidence. . . . One police chief thought that any contact between the police and the public was community policing. (P. 104)

After examining the evolution of police practices from the common practice of basic patrolling to the new community approaches, Bayley (1994) observed that four elements were common: consultation, adaptation, mobilization, and problem solving. He conceptualized each of these elements as follows:

*Consultation*: Establishing new mechanisms for discussing police priorities and strategies with their communities.

*Adaptation*: Because crime and order needs vary from place to place, adaptation involves reshaping command structures to that local police commanders can use resources more flexibly.
**Mobilization:** Rather than relying on just their own efforts to prevent crime, mobilization involves police developing programs that enlist the active assistance of the public (e.g., Neighborhood Watches).

**Problem Solving:** Instead of simply responding to crimes after they have occurred, problem solving involves police studying conditions that lead to calls for their services, drawing up plans to correct these conditions, and taking the lead in evaluating and implementing remedial actions.

As you can see, community policing involves several different dimensions. There is wide variability in the extent to which police departments have adopted these different dimensions and the future of community policing, at least in the United States is unclear. However, in other industrialized countries, including Japan and Singapore, community policing has become entrenched (Bayley 1994).

**Defining Poverty**

Decisions about how to define a concept reflect the theoretical framework that guides the researchers. For example, the concept *poverty* has always been somewhat controversial, because different notions of what poverty is shape estimates of how prevalent it is and what can be done about it.

Most of the statistics that you see in the newspaper about the poverty rate reflect a conception of poverty that was formalized by Mollie Orshansky of the Social Security Administration in 1965 and subsequently adopted by the federal government and many researchers (Putnam 1977). She defined poverty as an *absolute* standard, based on the amount of money required to purchase an emergency diet that is estimated to be nutritionally adequate for about 2 months. The idea is that people are truly poor if they can barely purchase the food they need and other essential goods. This poverty standard is adjusted for household size and composition (number of children and adults), and the minimal amount needed for food is multiplied by 3 because a 1955 survey indicated that poor families spend about one-third of their incomes on food (Orshansky 1977).

Some social scientists disagree with the absolute standard and have instead urged adoption of a *relative* poverty standard or income inequality (Blau & Blau 1982). The idea behind this relative conception is that poverty should be defined in terms of what is normal in a given society at a particular time.

Some social scientists prefer yet another conception of poverty. With the *subjective* approach, poverty is defined as what people think would be the minimal income they need to make ends meet. Of course, many have argued that this approach is influenced too much by the different standards that people use to estimate what they need (Ruggles 1990:20–23).

Which do you think this is a more reasonable approach to defining poverty: an absolute standard, a relative standard, or a subjective standard? Our understanding of the concept of poverty is sharpened when we consider the theoretical ramifications of these alternative definitions.

**Defining Strain**

Some concepts have multiple dimensions, bringing together several related concepts under a larger conceptual umbrella. One such concept is *strain*, first defined by Merton (1958) and
later by Cloward and Ohlin (1960). In its simplest form, strain theory contends that
delinquency and other forms of criminality and/or deviance are the result of frustrated
needs or wants; this frustration is generally believed to result from a breakdown in the rela-
tionship between socially induced aspirations or goals and socially approved ways of
achieving these goals. According to strain theory, delinquency is a response to actual or
anticipated failure to achieve socially induced needs or goals (e.g., status wealth, power,
social acceptance). To determine the relationship between perceived strain and delinquency
in a sample of 13-through 19-year-olds from the National Youth Survey (NYS), Elliott,
Huizinga, and Ageton (1985) conceptualized strain as goal expectation discrepancies in
both the home and school contexts. For example, strain would be increased at the individ-
ual level if being a good student was important to someone, but he or she did not perceive
that teachers thought of him or her as a good student.

Relying on Merton's (1938) earlier conceptualization, Messner and Rosenfeld (1994)
utilized the macro-level analogue of strain called anomie. Anomie, in its most general for-
mulation, is defined as a weakening in the normative regulation of behavior. Anomie theory
posits that diminished normative regulation results from an overemphasis on cultural goals
(e.g., wealth, prestige) relative to the legitimate means to achieve them (e.g., availability of
a college education, high status jobs), or from the differential distribution of opportunities
to achieve highly valued goals (some groups in society may not have the same opportuni-
ties to achieve these goals as others). According to Messner and Rosenfeld, there are four
cultural values within American society that contribute to anomie and therefore to crime
in general: achievement, individualism, universalism, and materialism. To fully define
anomie, Messner and Rosenfeld thus had to define these concepts. For example, to describe
what they meant by the value of materialism, they state,

In American culture, success is signified in a profoundly significant way: by
the accumulation of monetary rewards. Money is awarded special priority in
American culture. The point to emphasize here is not that Americans are uniquely
materialistic, for a strong interest in material well-being can be found in most
societies. Rather, the distinctive feature of American culture is the preeminent role
of money as the “metric” of success. Monetary success is inherently open-ended.
It is always possible in principle to have more money. Hence, the American Dream
offers no final stopping point. It requires never-ending achievement. The pressure
to accumulate money is therefore relentless, which entices people to pursue their
monetary goals by any means necessary. (P. 71)

Concepts and Variables

After we define the concepts in a theory, we can identify variables corresponding to the
concepts and develop procedures to measure them. This is an important step. Consider the
concept of social control, which Donald Black (1984) defines as “all of the processes by
which people define and respond to deviant behavior.” What variables can represent this
conceptualization of social control? Proportion of persons arrested in a community?
Average length of sentences for crimes? Types of bystander reactions to public intoxication?
Some combination of these?
Although we must proceed carefully to specify what we mean by a concept like social control, some concepts are represented well by the specific variables in the study and need not be defined so carefully. We may define binge drinking as heavy episodic drinking and measure it, is a variable, by asking people how many drinks they consumed in succession during some period (see Wechsler et al. 1994). That is pretty straightforward.

Be aware that not every concept in a study is represented by a variable. For example, if the term *tolerance of drinking* is defined as the absence of rules against drinking in a fraternity, it brings to mind a phenomenon that varies across different fraternities at different colleges. But if we study social life at only those fraternities that prohibit drinking, tolerance of drinking would not be a variable: All the fraternities studied have the same level of tolerance, and thus tolerance of drinking is a constant and not a variable. Of course, the concept of tolerance of drinking would still be important for understanding social life in the “dry” fraternities.

**MEASUREMENT OPERATIONS**

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

*Operation* The procedure for actually measuring the concepts we intend to measure, identifying the value of a variable for each case.

*Operationalization* The process of specifying the operations that will indicate the value of a variable for each case.

Exhibit 3.2 represents the operationalization process in three studies. The first researcher defines her concept (binge drinking) and chooses one variable (frequency of heavy episodic drinking) to represent it. This variable is then measured with responses to a single question, or indicator: “How often within the past two weeks did you consume five or more drinks containing alcohol in a row?” The second researcher defines her concept, poverty, as having two aspects or dimensions, subjective poverty and absolute poverty. Subjective poverty is measured with responses to a survey question: “Do you consider yourself poor?” Absolute poverty is measured by comparing family income to the poverty threshold. The third researcher decides that her concept, social class, can be indicated with three measured variables: income, education, and occupational prestige. The values of these three variables for each case studied are then combined into a single indicator.

Good conceptualization and operationalization can prevent confusion later in the research process. For example, a researcher may find that substance abusers who join a self-help group are less likely to drink again than those who receive hospital-based substance abuse treatment. But what is it about these treatment alternatives that is associated with successful abstinence? Level of peer support? Beliefs about the causes of alcoholism? Financial
investment in the treatment? If the researcher had considered such aspects of the concept of substance abuse treatment before collecting her data, she might have been able to measure different aspects of treatment and so figure out which, if any, were associated with differences in abstinence rates. Because she did not measure these variables, she will not contribute as much as she might have to our understanding of substance abuse treatment.

Social researchers have many options for operationalizing their concepts. Measures can be based on activities as diverse as asking people questions, reading judicial opinions, observing social interactions, coding words in books, checking census data, enumerating the contents of trash receptacles, or drawing urine and blood samples. We focus here on the operations of using published data, asking questions, observing behavior, and using unobtrusive means of measuring people’s behavior and attitudes.

**Using Available Data**

Government reports are rich and readily accessible sources of criminal justice data, as are datasets available from nonprofit advocacy groups, university researchers, and some private businesses. For example, law-enforcement and health statistics provide several community-level indicators of substance abuse (Gruenewald et al. 1997). Statistics on arrests for the sale and possession of drugs, drunk driving arrests, and liquor law violations (such as sales to minors) can usually be obtained on an annual basis, and often quarterly, from local police departments or state crime information centers.

Indicators like these cannot be compared across communities or over time without reviewing carefully how they were constructed. The level of alcohol in the blood that is legally required to establish intoxication can vary among communities, creating the

---

### EXHIBIT 3.2 Concepts, Variables, and Indicators

<table>
<thead>
<tr>
<th>Concept</th>
<th>Variable</th>
<th>Indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binge drinking</td>
<td>Frequency of heavy episodic drinking</td>
<td>“How often within the past two weeks did you consume five or more drinks containing alcohol in a row?”</td>
</tr>
<tr>
<td>Subjective poverty</td>
<td></td>
<td>“Would you say you are poor?”</td>
</tr>
<tr>
<td>Poverty</td>
<td>Absolute poverty</td>
<td>Family income ÷ poverty</td>
</tr>
<tr>
<td></td>
<td>Income</td>
<td></td>
</tr>
<tr>
<td>Social class</td>
<td>Education</td>
<td>Income + education + prestige</td>
</tr>
<tr>
<td></td>
<td>Occupational prestige</td>
<td></td>
</tr>
</tbody>
</table>
appearance of different rates of substance abuse even though drinking and driving practices may be identical. Enforcement practices can vary among police jurisdictions and over time (Gruenewald et al. 1997:14). We also cannot assume that available data are accurate, even when they appear to measure the concept in which we are interested in a way that is consistent across communities. “Official” counts of homeless persons have been notoriously unreliable because of the difficulty of locating homeless persons on the streets, and government agencies have, at times, resorted to “guesstimates” by service providers (Rossi 1989). Even available data for such seemingly straightforward measures as cause of death can contain a surprising amount of error. For example, between 30% and 40% of death certificates incorrectly identify the cause of death (Altman 1998).

Government statistics that are generated through a central agency like the U.S. Bureau of the Census are often of high quality, but caution is warranted when using official data collected by local levels of government. For example, the Uniform Crime Reports (UCR) program administered by the Federal Bureau of Investigation imposes standard classification criteria, with explicit guidelines and regular training at the local level, but data are still inconsistent for many crimes. Consider only a few of the many sources of inconsistency between jurisdictions: variation in the classification of forcible rape cases due to differences in what is considered to be “carnal knowledge of a female;” different decisions about what is considered “more than necessary force” in the definition of “strong-arm” robberies; whether offenses in which threats were made but no physical injury occurred are classified as aggravated or simple assaults. (Mosher, Miethe, & Phillips 2002:66). A new National Incident-Based Reporting System (NIBRS) corrects some of the problems with the UCR, but it requires much more training and documentation and has not yet been widely used (Mosher, Miethe, & Phillips 2002:70).

In some cases, problems with an available indicator can be lessened by selecting a more precise indicator. For example, the number of single-vehicle nighttime crashes, whether fatal or not, is a more specific indicator of the frequency of drinking and driving than just the number of single-vehicle fatal accidents (Gruenewald et al. 1997:40–41). Focusing on a different level of aggregation may also improve data quality, because procedures for data collection may differ between cities, counties, states, and so on (Gruenewald et al. 1997:40–41). It is only after such factors as legal standards, enforcement practices, and measurement procedures have been taken into account that comparisons among communities become credible.

Constructing Questions

Asking people questions is the most common and probably the most versatile operation for measuring social variables. Most concepts about individuals can be defined in such a way that measurement with one or more questions becomes an option. In this section, we introduce some options for writing single questions; in Chapter 8, we explain why single questions can be inadequate measures of some concepts, and then examine approaches that rely on multiple questions to measure a concept.

Measuring variables with single questions is very popular. Public opinion polls based on answers to single questions are reported frequently in newspaper articles and TV newscasts: “Do you favor or oppose U.S. policy in . . . ?” “If you had to vote today, for which candidate would you vote?” Criminal justice surveys also rely on single questions to measure many
variables: “Overall, how satisfied are you with the police in your community?” “How would you rate your current level of safety?”

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

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

_____ Greater than other campuses
_____ Less than other campuses
_____ About the same as other campuses

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

Most surveys of a large number of people contain primarily fixed-choice questions, which are easy to process with computers and analyze with statistics. With fixed-choice questions, respondents are also more likely to answer the question that the researcher really wants them to answer. Including response choices reduces ambiguity and makes it easier for respondents to answer. However, fixed-response choices can obscure what people really think if the choices do not match the range of possible responses to the question; many studies show that some respondents will choose response choices that do not apply to them simply to give some sort of answer (Peterson 2000:39). We will discuss question wording and response options in greater detail in Chapter 7.

Open-ended questions—questions without explicit response choices, to which respondents write in their answers—are preferable when the range of responses cannot adequately be anticipated, namely, questions that have not previously been used in surveys and questions that are asked of new groups. Open-ended questions can also lessen confusion about the meaning of responses involving complex concepts. The next question is an open-ended version of the earlier fixed-choice question:

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

Observations can be used to measure characteristics of individuals, events, and places. The observations may be the primary form of measurement in a study, or they may supplement measures obtained through questioning. Reiss (1971a) developed a careful method of observing phenomena that he termed systematic social observation (SSO). In his study of police interaction with the public, Reiss’s SSO method involved riding in police squad cars, observing police-citizen interactions and recording features of these interactions characteristics on a form.

Sampson and Raudenbush (1999) refined the SSO technique in their study of neighborhood disorder and crime. Teams drove in “a sport utility vehicle at a rate of five miles per hour down every street” in a sample of Chicago neighborhoods (Raudenbush and Sampson 1999). Two video cameras recorded people and activities on both sides of the vehicle, while a trained observer completed a log for each block. The resulting 23,816 observer logs contained information about building conditions and land use, while the videotapes were coded to measure features of streets, buildings, businesses, and social interaction on 15,141 blocks. Direct observation is often the method of choice for measuring behavior in natural settings, as long as it is possible to make the requisite observations.

Collecting Unobtrusive Measures

Unobtrusive measures allow us to collect data about individuals or groups without their direct knowledge or participation. In their classic book (now revised), Webb et al. ([1966] 2000) identified four types of unobtrusive measures: physical trace evidence, archives (available data), simple observation, and contrived observation (using hidden recording hardware or manipulation to elicit a response). Let us consider the first two types in more detail: physical trace evidence and archives.

The physical traces of past behavior are one type of unobtrusive measure that is most useful when the behavior of interest cannot be directly observed (perhaps because it is hidden or occurred in the past) and has not been recorded in a source of available data. To measure the prevalence of drinking in college dorms or fraternity houses, we might count the number of empty bottles of alcoholic beverages in the surrounding dumpsters. However, you can probably see that care must be taken to develop trace measures that are useful for comparative purposes. For instance, comparison of the number of empty bottles in dumpsters outside different dorms can be misleading; at the very least, you would need to take into account the number of residents in the dorms, the time since the last trash collection, and the accessibility of each dumpster to passersby.

Unobtrusive measures can also be created from such diverse forms of media as newspaper archives or magazine articles, TV or radio talk shows, legal opinions, historical documents, personal letters, or e-mail messages. An investigation of the drinking climate on campuses might include a count of the amount of space devoted to ads for alcoholic beverages in a sample of issues of the student newspaper. Campus publications also might be coded to indicate the number of times that statements discouraging substance abuse appear. With this tool, you could measure the frequency of articles reporting substance abuse--related crimes, the degree of approval of drinking expressed in TV shows or songs, or the relationship between region of the country and amount of space devoted in the print media to drug usage.
Combining Measurement Operations

Using available data, asking questions, making observations, and using unobtrusive indicators are interrelated measurement tools, each of which may include or be supplemented by the others. From people’s answers to survey questions, the U.S. Bureau of the Census develops widely consulted census reports containing available data on people, firms, and geographic units in the United States. Data from employee surveys may be supplemented by information available in company records. Interviewers may record observations about those whom they question. Researchers may use insights gleaned from questioning participants to make sense of the social interaction they have observed. Unobtrusive indicators can be used to evaluate the honesty of survey responses.

Questioning can be a particularly poor approach for measuring behaviors that are very socially desirable, such as voting or attending church, or that are socially stigmatized or illegal, such as abusing alcohol or drugs. Triangulation, the use of two or more different measures of the same variable, can strengthen measurement considerably (Brewer & Hunter 1989:17). When we achieve similar results with different measures of the same variable, particularly when they are based on such different methods as survey questions and field-based observations, we can be more confident in the validity of each measure. If results diverge with different measures, it may indicate that one or more of these measures are influenced by more measurement error than we can tolerate. Divergence between measures could also indicate that they actually operationalize different concepts. An interesting example of this interpretation of divergent results comes from research on crime. Official crime statistics only indicate those crimes that are reported to and recorded by the police; when surveys are used to measure crimes with self-reports of victims, many “personal annoyances” are included as if they were crimes (Levine 1976). We will talk more about triangulation in Chapter 10.

EVALUATION OF MEASURES

The issue of measurement error is very important. Do the operations developed to measure our concepts actually do so; are they valid? If we have weighed our measurement options, carefully constructed our questions and observational procedures, and carefully selected indicators from the available data, we should be on the right track. But we cannot have much confidence in a measure until we have empirically evaluated its validity.

Measurement Validity

We can consider measurement validity the first concern in establishing the validity of research results, because without having measured what we think we measured, we really do not know what we are talking about.

Measurement validity The extent to which measures indicate what they are intended to measure.
As an example of measurement validity, consider the following question: “How prevalent is youth violence and delinquency in the United States; how many juveniles are involved in delinquency?” Data on the extent of juvenile delinquency come from two primary sources: official statistics and unofficial statistics. Official statistics are based on the aggregate records of juvenile offenders and offenses processed by agencies of the criminal justice system: police, courts, and corrections. One primary source of official statistics on juvenile delinquency is the Uniform Crime Reports (UCR), produced by the Federal Bureau of Investigation (FBI). Unofficial statistics are data produced by people or agencies outside the criminal justice system such as victimization surveys and self-report studies. The validity of official statistics for measuring the extent of juvenile delinquency is a heated debate among criminologists. Although some researchers believe official reports are a valid measure of serious delinquency, others contend that UCR data say more about the behavior of the police than about delinquency. These criminologists think the police are predisposed against certain groups of people or certain types of crimes.

Unquestionably, official reports underestimate the actual amount of delinquency because a great deal of delinquent behavior never comes to the attention of police (Mosher, Miethe, & Phillips 2002). Sometimes delinquent acts are committed and not observed; or they are observed and not reported. There is also evidence that the UCRs often reflect the political climate and police policies as much as they do criminal activity. Take the United States’ “War on Drugs,” which heated up in the 1980s. During this time, arrest rates for drug offenses soared, giving the illusion that drugs used was increasing at an epidemic pace. However, self-report surveys that asked citizens directly about their drug use behavior during this time period found that use of most illicit drugs was actually declining (Regoli & Hewitt 1994). In your opinion, then, which measure of drug use, the UCR or self-report surveys, was more valid? Before you answer this question, let us continue with the delinquency example.

Despite the limitations of official statistics for measuring delinquency, these data were relied on by criminologists and used as a valid measure of the prevalence of delinquency for many decades. As such, delinquency and other violent offenses were thought to primarily involve minority populations and/or disadvantaged youth. In 1947, however, Wallerstein and Wyle (1947) surveyed a sample of 700 juveniles and found that 91% admitted to having committed at least one offense that was punishable by one or more years in prison, and 99% admitted to at least one offense for which they could have been arrested had they been caught. In 1958, Short and Nye (1957–1958) reported the results from the first large-scale self-report study involving juveniles from a variety of locations. In their research, Short and Nye concluded that delinquency was widespread throughout the adolescent population and that youth from high-income families were just as likely to engage in delinquency as youth from low-income families. Contemporary studies using self-report data from the National Youth Survey (NYS) indicate that the actual amount of delinquency is much greater than that reported by the UCR and that unlike this official data, where nonwhites are overrepresented, self-report data indicate that white juveniles report nearly the identical number of delinquencies as nonwhites, but fewer of them are arrested (Elliott & Ageton 1980). This is just one example of measurement validity and should convince you that we must be very careful in designing our measures and in subsequently evaluating how well they have performed.

The extent to which measures indicate what they are intended to measure can be assessed with one or more of four basic approaches: face validation, content validation,
criterion validation, and construct validation. Whatever the approach to validation, no one measure will be valid for all times and places. For example, the validity of self-report measures of substance abuse varies with such factors as whether the respondents are sober or intoxicated at the time of the interview, whether the measure refers to recent or lifetime abuse, and whether the respondents see their responses as affecting their chances at receiving housing, treatment, or some other desired outcome (Babor, Stephens, & Marlatt 1987). In addition, persons with severe mental illness are, in general, less likely to respond accurately (Corse, Hirschinger, & Zanis 1995). These types of possibilities should always be considered when evaluating measurement validity.

**Face Validity**

Researchers apply the term *face validity* to the confidence gained from careful inspection of a concept to see if it is appropriate “on its face,” simply whether it appears to measure what it intends. For example, measuring people’s favorite color seems unlikely on its face to tell us much about their alcohol consumption patterns. A measure with greater face validity would be a count of how many drinks they had consumed in the past week.

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

**Content Validity**

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

An example of a measure that covers a wide range of meaning is the Michigan Alcoholism Screening Test (MAST). The MAST includes 24 questions representing the following subscales: recognition of alcohol problems by self and others; legal, social, and work problems; help seeking; marital and family difficulties; and liver pathology (Skinner & Sheu 1982). Many experts familiar with the direct consequences of substance abuse agree that these dimensions capture the full range of possibilities. Thus, the MAST is believed to be valid from the standpoint of content validity.

**Criterion Validity**

When people drink an alcoholic beverage, the alcohol is absorbed into their blood and then gradually metabolized (broken down into other chemicals) in their liver (NIAAA 1997). The alcohol that remains in their blood at any point, unmetabolized, impairs both thinking and behavior (NIAAA 1994). As more alcohol is ingested, cognitive and behavioral consequences multiply. These biological processes can be identified with direct measures of alcohol concentration in the blood, urine, or breath. Questions about drinking behavior, on the other hand, can be viewed as attempts to measure indirectly what biochemical tests measure directly.

*Criterion validity* is established when the scores obtained on one measure can be accurately compared to those obtained with a more direct or already validated measure of the
same phenomenon (the criterion). A measure of blood-alcohol concentration or a urine test could serve as the criterion for validating a self-report measure of drinking, as long as the questions we ask about drinking refer to the same period. Observations of substance use by friends or relatives could also, in some circumstances, serve as a criterion for validating self-report substance use measures.

Criterion validation studies of substance abuse measures have yielded inconsistent results. Self-reports of drug use agreed with urinalysis results for about 85% of the drug users who volunteered for a health study in several cities (Weatherby et al. 1994). On the other hand, the post-treatment drinking behavior self-reported by 100 male alcoholics was substantially less than the drinking behavior observed by the alcoholics’ friends or relatives (Watson et al. 1984). Such inconsistent findings can occur because of differences in the adequacy of a measures across settings and populations. This underscores our point that you cannot assume that a measure that was validated in one study is also valid in another setting or with a different population.

An attempt at criterion validation is well worth the effort because it greatly increases confidence that the measure is measuring what was intended. However, often no other variable might reasonably be considered a criterion for feelings or beliefs or other subjective states. Even with variables for which a reasonable criterion exists, the researcher may not be able to gain access to the criterion, as would be the case with a tax return or employer document as criterion for self-reported income.

Construct Validity

Measurement validity also can be established by showing that a measure is related to a variety of other measures as specified in a theory. This validation approach, known as construct validity, is commonly used in social research when no clear criterion exists for validation purposes. For example, in one study of the validity of the Addiction Severity Index (ASI), A. Thomas McLellan et al. (1985) compared subject scores on the ASI to a number of indicators that they felt from prior research should be related to substance abuse: medical problems, employment problems, legal problems, family problems, and psychiatric problems. They could not use a criterion validation approach because they did not have a more direct measure of abuse, such as laboratory test scores or observer reports. However, their extensive research on the subject had given them confidence that these sorts of problems were all related to substance abuse, and thus their measures seemed to valid from the standpoint of construct validity. Indeed, the researchers found that individuals with higher ASI ratings tended to have more problems in each of these areas, giving us more confidence in the ASI’s validity as a measure.

A somewhat different approach to construct validation is termed discriminant validity. In this approach, scores on the measure to be validated are compared to scores on another measure of the same variable and to scores on variables that measure different but related concepts. Discriminant validity is achieved if the measure to be validated is related most strongly to its comparison measure and less to the measures of other concepts. McLellan et al. (1985) found that the ASI passed this test, too: The ASI’s measures of alcohol and drug problems were related more strongly to other measures of alcohol and drug problems than they were to measures of legal problems, family problems, medical problems, and the like.

The distinction between criterion and construct validation is not always clear. Opinions can differ about whether a particular indicator is indeed a criterion for the concept that is
to be measured. For example, if you need to validate a question-based measure of sales ability for applicants to a sales position, few would object to using actual sales performance as a criterion. But what if you want to validate a question-based measure of the amount of social support that people receive from their friends? Should you just ask people about the social support they have received? Could friends’ reports of the amount of support they provided serve as a criterion? Are verbal accounts of the amount of support provided adequate? What about observations of social support that people receive? Even if you could observe people in the act of counseling or otherwise supporting their friends, can an observer be sure that the interaction is indeed supportive? There isn’t really a criterion here, just related concepts that could be used in a construct validation strategy. Even biochemical measures of substance abuse are questionable as criteria for validating self-reported substance use. Urine test results can be altered by ingesting certain substances, and blood tests vary in their sensitivity to the presence of drugs over a particular period of time.

What construct and criterion validation have in common is the comparison of scores on one measures to scores on other measures that are predicted to be related. It is not so important that researchers agree that a particular comparison measure is a criterion rather than a related construct. But it is very important to think critically about the quality of the comparison measure and whether it actually represents a different measure of the same phenomenon. For example, it is only a weak indication of measurement validity to find that scores on a new self-report measure of alcohol use are associated with scores on a previously used self-report measure of alcohol use.

Reliability

Reliability means that a measurement procedure yields consistent scores when the phenomenon being measured is not changing (or that the measured scores change in direct correspondence to actual changes in the phenomenon). If a measure is reliable, it is affected less by random error, or chance variation, that if it is unreliable. Reliability is a prerequisite for measurement validity; we cannot really measure a phenomenon if the measure we are using gives inconsistent results.

Reliability A measure is reliable when it yields consistent scores or observations of a given phenomenon on different occasions. Reliability is a prerequisite for measurement validity.

There are four possible methods for measuring the reliability of a measure: test-retest reliability, interitem reliability, alternate-forms reliability, and interobserver reliability.

Test-Retest Reliability

When researchers measure a phenomenon that does not change between two points separated by an interval of time, the degree to which the two measurements yield comparable, if not identical, values is the test-retest reliability of the measure. If you take a test of your math ability and then retake the test 2 months later, the test is performing reliably if you receive a similar score both times, presuming that nothing happened during the 2 months to change your math ability. Of course, if events between the test and the retest
have changed the variable being measured, then the difference between the test and retest scores should reflect that change.

When ratings by an observer, rather than ratings by the subjects themselves, are being assessed at two or more points in time, test-retest reliability is termed intraobserver reliability or intrarater reliability.

One example of how test-retest reliability may be assessed is a study by Sobell et al. (1988) of alcohol abusers’ past drinking behavior (using the Lifetime Drinking History questionnaire) and life changes (using the Recent Life Changes questionnaire). All 69 subjects in the study were patients in an addiction treatment program. They had not been drinking prior to the interview (determined by a breath test). The two questionnaires were administered by different interviewers about 2 or 3 weeks apart, both times asking the subjects to recall events 8 years prior to the interviews. Reliability was high: 92% of the subjects reported the same life events both times, and at least 81% of the subjects were classified consistently at both interviews as having had an alcohol problem or not. When asked about their inconsistent answers, subjects reported that in the earlier interview they had simply dated an event incorrectly, misunderstood the question, evaluated the importance of an event differently, or forgotten an event. Answers to past drinking questions were less reliable when they were very specific, apparently because the questions exceeded subjects’ capacities to remember accurately.

Interitem Reliability (Internal Consistency)

When researchers use multiple items to measure a single concept, they are concerned with interitem reliability (or internal consistency). For example, if we are to have confidence that a set of questions (such as those in Exhibit 3.3) reliably measures attitudes toward violence, the answers to the questions should be highly associated with one another. The stronger the association among the individual items, and the more items that are included, the higher reliability of the index.

Alternate-Forms Reliability

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

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

Interobserver Reliability

When researchers use more than one observer to rate the same persons, events, or places, interobserver reliability is their goal. If observers are using the same instrument to rate the same thing, their ratings should be very similar. If they are similar, we can have much more confidence that the ratings reflect the phenomenon being assessed rather than the orientations of the observers.
Ways to Improve Reliability and Validity

We must always assess the reliability of a measure if we hope to able to establish its validity. Remember that a reliable measure is not necessarily a valid measure, as Exhibit 3.4 illustrates. This discrepancy is a common flaw of self-report measures of substance abuse. The multiple questions in self-report indexes of substance abuse or answered by most respondents in a consistent way, so the indexes are reliable. However, a number of respondents will not admit to drinking, even though they drink a lot. Their answers to the questions are consistent, but they are consistently misleading. So the indexes based on self-report are reliable but invalid. Such indexes are not useful and should be improved or discarded. Unfortunately, many measures are judged to be worthwhile on the basis only of a reliability test.

The reliability and validity of measures in any study must be tested after the fact to assess the quality of the information obtained. But then if it turns out that a measure cannot be considered reliable and valid, little can be done to save the study. Hence it is supremely important to select in the first place measures that are likely to be reliable and valid. In studies that use interviewers or observers, careful training is often essential to achieving a consistent approach. In most cases, however, the best strategy is to use measures that have been used before and whose reliability and validity have been established in other contexts. But the
selection of “tried and true” measures still does not absolve researchers from the responsibility of testing the reliability and validity of the measure in their own studies.

The process of evaluating the reliability and validity of measures about individuals is termed psychometrics. Measures of individuals that range from tests you take in school to personality assessments you complete on the job are advertised as “psychometrically valid” after multiple studies have demonstrated their reliability and validity. The process of evaluating the reliability and validity of measures about organizations, neighborhoods, or other collective units is termed ecometrics, a term coined by Raudenbush and Sampson (1999). For example, Raudenbush and Sampson’s ecometric evaluation of their observational measures of Chicago neighborhoods included a test of the consistency of ratings by multiple observers of the same neighborhoods (p. 7).

It may be possible to improve the reliability and validity of measures in a study that already has been conducted if multiple measures were used. For example, in our study of housing for homeless mentally ill persons, funded by the National Institute of Mental Health, we assessed substance abuse with several different sets of direct questions as well as with
reports from subjects’ case managers and others (Goldfinger et al. 1996). We found that the observational reports were often inconsistent with self-reports and that different self-report measures were not always in agreement and were thus unreliable. A more reliable measure of substance abuse was initial reports of lifetime substance abuse problems. This measure was extremely accurate in identifying all those who subsequently abused substances during the project. We concluded that the lifetime measure was a valid way to identify persons at risk for substance abuse problems.

**LEVELS OF MEASUREMENT**

When we know a variable’s level of measurement, we can better understand how cases vary on that variable and so understand more fully what we have measured. Level of measurement also has important implications for the types of statistics that can be used with the variable, as you will learn in Chapter 11. There are four levels of measurement: nominal, ordinal, interval, and ratio. Exhibit 3.5 depicts the differences among these four levels.

**Level of measurement** The complexity of the mathematical means that can be used to express the relationship between a variable’s values. The nominal level of measurement, which is qualitative, has no mathematical interpretation; the quantitative levels of measurement (ordinal, interval, and ratio) are progressively more complex mathematically.

**Nominal Level of Measurement**

The nominal level of measurement (also called the categorical or qualitative level) identifies variables whose values have no mathematical interpretation; they vary in kind or quality but not in amount. In fact, it is conventional to refer to the values of nominal variables as attributes instead of values. Gender is one example. The variable “gender” has two attributes (or categories or qualities): male and female. We might indicate male with the value 1 and female with the value 2, but these numbers do not tell us anything about the difference between male and female except that they are different. Female is not one unit more of “gender” than male, nor is it twice as much “gender.” Ethnicity, occupation, religious affiliation, and region of the country are also measured at the nominal level. A person may be Spanish or Portuguese, but one ethnic group does not represent more ethnicity than another, just a different ethnicity. A person may be a doctor or a truck driver, but one does not represent three units more occupation than the other.

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

- A variable’s attributes or values are mutually exclusive attributes if every case can have only one attribute.
- A variable’s attributes or values are exhaustive attributes when every case can be classified into one of the categories.
When a variable’s attributes are mutually exclusive and exhaustive, every case corresponds to one, and only one, attribute. Imagine the challenge of coming up with an exhaustive set of attributes when a large sample is being studied.

**Ordinal Level of Measurement**

The first of the three quantitative levels is the *ordinal level of measurement*. At this level, the numbers assigned to cases specify only the order of the cases, permitting greater-than and less-than distinctions; absolute mathematical distinctions cannot be made between categories. The Core Alcohol and Drug Survey (Core Institute 1994) measures substance abuse with a series of questions that permit ordinal distinctions (see Exhibit 3.6). Although these
**EXHIBIT 3.6 Example of Ordinal Measures: Core Alcohol and Drug Survey**

<table>
<thead>
<tr>
<th>Within the last year, about how often have you used . . . (mark one for each line)</th>
<th>Did Not Use</th>
<th>Once a Year</th>
<th>6 Times a Year</th>
<th>Once a Month</th>
<th>Twice a Month</th>
<th>Once a Week</th>
<th>3 Times a Week</th>
<th>5 Times a Week</th>
<th>Every day</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Tobacco (some, chew, snuff)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Alcohol (beer, wine, liquor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Marijuana (pot, hash, hash oil)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. Cocaine (crack, rock, freebase)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Amphetamines (diet pills, speed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. Sedatives (downers, ludes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g. Hallucinogens (LSD, PCP)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h. Opiates (heroin, smack, horse)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Inhalants (glue, solvents, gas)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j. Designer drugs (ecstasy, MDMA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k. Steroids</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l. Other illegal drugs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Core Institute 1994.
categories may seem at first to be mathematically distinct, you can really only determine the extent to which a respondent who checks one category consumes more or less than another respondent.

The properties of variables measured at the ordinal level are illustrated in Exhibit 3.5 by the contrast between the level of conflict in two groups. The first group, symbolized by two people shaking hands, has a low level of conflict. The second group, symbolized by two persons using fists against each other, has a higher level of conflict. The third group, symbolized by two people pointing guns at each other, has an even higher level of conflict. To measure conflict, we would put the groups “in order” by assigning the number 1 to the low-conflict group, the number 2 to the group using fists, and the number 3 to the high-conflict group using guns. The numbers thus indicate only the relative position or order of the cases. Although low level of conflict is represented by the number 1, it is not mathematically two less units of conflict than the high level of conflict, which is represented by the number 3. These numbers really have no mathematical qualities; they are just used to represent relative rank in the measurement of conflict.

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

**Interval Level of Measurement**

The numbers indicating the values of a variable at the interval level of measurement represent fixed measurement units but have no absolute, or fixed, zero point. This level of measurement is represented in Exhibit 3.5 by the difference between two Fahrenheit temperatures. Although 60 degrees is 30 degrees hotter than 30 degrees, 60 in this case is not twice as hot as 30. Why not? Because heat does not begin at 0 degrees on the Fahrenheit scale.

An interval-level measure is created by a scale that has fixed measurement units but no absolute, or fixed, zero point. The numbers can therefore be added and subtracted, but ratios are not meaningful. Again, the values must be mutually exclusive and exhaustive.

Social scientists often treat indexes that were created by combining responses to a series of variables measured at the ordinal level as interval-level measures. An index of this sort could be created with responses to the Core Institute’s (1994) questions about friends’ disapproval of substance use (see Exhibit 3.7). The survey has 13 questions on the topic, each of which has the same three response choices. If Don’t disapprove is valued at 1, Disapprove is valued at 2, and Strongly disapprove is valued at 3, the summed index of disapproval would range from 12 to 36. The average could then be treated as a fixed unit of measurement. So a score of 20 could be treated as if it were 4 more units than a score of 16.

**Ratio Level of Measurement**

The numbers indicating the values of a variable at the ratio level of measurement represent fixed measuring units and an absolute zero point (zero means absolutely no amount of whatever the variable indicates). On a ratio scale, 10 is two points higher than 8 and is also two times greater than 5. Ratio numbers can be added and subtracted, and because the
numbers begin at an absolute zero point, they can be multiplied and divided (so ratios can be formed between the numbers). For example, people’s ages can be represented by values ranging from 0 years (or some fraction of a year) to 120 or more. A person who is 30 years old is 15 years older than someone who is 15 years old (30 – 15 = 15) and is twice as old as that person (30/15 = 2). Of course, the numbers also are mutually exclusive and exhaustive, so that every case can be assigned one and only one value.

Exhibit 3.5 displays an example of a variable measured at the ratio level. The number of people in the first group is 5, and the number in the second group is 7. The ratio of the two groups’ sizes is then 1.4, a number that mirrors the relationship between the sizes of the groups. Note that there does not actually have to be any group with a size of 0; what

EXHIBIT 3.7 Ordinal Level Variables Can Be Added to Create an Index With Interval-Level Properties: Core Alcohol and Drug Survey

<table>
<thead>
<tr>
<th>How do you think your close friends feel (or would feel) about you . . . (mark one for each line)</th>
<th>Do Not Disapprove</th>
<th>Disapprove</th>
<th>Strongly Disapprove</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Trying marijuana once or twice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Smoking marijuana occasionally</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Smoking marijuana regularly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. Trying cocaine once or twice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Taking cocaine regularly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. Trying LSD once or twice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g. Taking LSD regularly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h. Trying amphetamines once or twice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Taking amphetamines regularly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j. Taking one or two drinks of an alcoholic beverage (beer, wine, liquor) nearly every day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k. Taking four or five drinks nearly every day</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l. Having five or more drinks in one sitting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m. Taking steroids for bodybuilding or improved athletic performance</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Core Institute 1994.
is important is that the numbering scheme begins at an absolute zero, in this case, the absence of any people. The number of days a convicted felon was sentenced to prison would represent a ratio level of measurement because sentence length begins with an absolute 0 point. The number of days an addict stays clear after treatment, too, has a ratio level of measurement.

The Case of Dichotomies

Dichotomies, variables having only two values, are a special case from the standpoint of levels of a measurement. Although variables with only two categories are generally thought of as nominally measured, we can also think of a dichotomy as indicating the presence or absence of an attribute. Suppose, for example, we were interested in differences between individuals who had never used illegal drugs in the last year compared to those who had used at least one illegal drug in the last year. We could create a variable that indicated this dichotomous distinction by coding those individuals who said they did not use any of the substances listed as 0, and all others as 1. Viewed in this way, there is an inherent order to the two values: In one group the attribute of consuming illegal substances is absent (those coded 0), and in another it is present (those coded 1).

Comparison of Levels of Measurement

Exhibit 3.8 summarizes the types of comparisons that can be made with different levels of measurement, as well as the mathematical operations that are legitimate. All four levels of measurement allow researchers to assign different values to different cases. All three quantitative measures allow researchers to rank cases in order.

An important thing to remember is that researchers choose levels of measurement in the process of operationalizing the variables; the level of measurement is not inherent in the variable itself. Many variables can be measured at different levels, with different procedures.

<table>
<thead>
<tr>
<th>Examples of Comparison Statements</th>
<th>Relevant Level of Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Appropriate Math Operations</td>
</tr>
<tr>
<td>$A$ is equal to (not equal to) $B$</td>
<td>$=$ ($\neq$)</td>
</tr>
<tr>
<td>$A$ is greater than (less than) $B$</td>
<td>$&gt;$ ($&lt;$)</td>
</tr>
<tr>
<td>$A$ is three more than (less than) $B$</td>
<td>$+$ ($-$)</td>
</tr>
<tr>
<td>$A$ is twice (half) as large as $B$</td>
<td></td>
</tr>
</tbody>
</table>
For example, The “Core Alcohol and Drug Survey” (Core Institute 1994) identifies binge drinking by asking students, “Think back over the last two weeks. How many times have you had five or more drinks at a sitting?” You might be ready to classify this as a ratio-level measure, but you must first examine the fixed response options given to respondents. This is a **closed-ended question**, and students are asked to indicate their answer by checking None, Once, Twice, 3 to 5 times, 6 to 9 times, or 10 or more times. Use of these categories makes the level of measurement ordinal. The distance between any two cases cannot be clearly determined. A student with a response in the “6 to 9 times” category could have binged just one more time than a student who responded 3 to 5 times, or they could have binged 4 more times. With these response categories, you cannot mathematically distinguish the number of times a student binged, only the relative amount of binging behavior.

The more information available, the more ways you have to compare cases. You also have more possibilities for statistical analysis with quantitative than with qualitative variables. Thus, it often is a good idea to try to measure variables at the highest level of measurement possible, if doing so does not distort the meaning of the concept that is to be measured. For example, measure age in years, if possible, rather than in categories. You can always recategorize the responses into categories that compare teenagers to young adults, but it is impossible to obtain the actual age in years when the question is asked using an ordinal response format.

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

**CONCLUSION**

Remember always that measurement validity is a necessary for social research. Gathering data without careful conceptualization or conscientious efforts to operationalize key concepts often is a wasted effort.

The difficulties of achieving valid measurement vary with the concept being operationalized and the circumstances of the particular study. The examples in this chapter of difficulties in achieving valid measures of substance abuse should sensitize you to the need for caution, particularly when the concepts you wish to measure are socially stigmatized and/or illegal.

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

Alternate-forms reliability  Level of measurement
Closed-ended question  Measurement validity
Concept  Mutually exclusive attributes
Conceptualization  Nominal level of measurement
Construct validity  Open-ended questions
Content validity  Operation
Criterion validity  Operationalization
Dichotomy  Ordinal level of measurement
Discriminant validity  Psychometrics
Ecometrics  Ratio level of measurement
Exhaustive attributes  Reliability
Face validity  Split-halves reliability
Indicator  Systematic social observation
Interitem reliability  Test-retest reliability
Interobserver reliability  Triangulation
Interval level of measurement  Unobtrusive measures
Intraobserver reliability (intrarater reliability)  

HIGHLIGHTS

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

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

- The validity of measures should always be tested. There are four basic approaches: face validation, content validation, criterion validation, and construct validation. Criterion validation provides the strongest evidence of measurement validity, but there often is no criterion to use in validating social science measures.

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

- Level of measurement indicates the type of information obtained about a variable and the type of statistics that can be used to describe its variation. The four levels of measurement can be ordered by complexity of the mathematical operations they permit: nominal (least complex), ordinal, interval, ratio (most complex). The measurement level of a variable is determined by how the variable is operationalized. Dichotomies are a special case of measurement.

EXERCISES

1. Are important concepts in criminological research always defined clearly? Are they defined consistently? Search the literature for six articles that focus on "violent crime," "domestic violence," or some other concept suggested by your instructor. Is the concept
defined clearly in each article? How similar are the definitions? Write what you have found in a short report.

2. What are some of the research questions you could attempt to answer with available statistical data? Visit your library and ask for an introduction to the government documents collection. Inspect the volumes from the Federal Bureau of Investigation (FBI) Uniform Crime Report (UCR) or the Sourcebook for Criminal Justice Statistics, both of which report statistics on crimes committed by offender characteristics. List 10 questions you could explore with such data.

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

DEVELOPING A RESEARCH PROPOSAL

At this point you can begin the processes of conceptualization and operationalization. You will need to assume that your primary research method will be conducting a survey.

1. List at least 10 variables that will be measured in your research. No more than two of these should be sociodemographic indicators like race or age. The inclusion of each variable should be justified in terms of theory or prior research that suggests it would be an appropriate independent or dependent variable, or will have some relation to either of these.

2. Write a conceptual definition for each variable. Whenever possible, this definition should come from the existing literature, either a book you have read for a course or the research literature that you have been searching. Ask two class members for feedback on your definitions.

3. Develop measurement procedures for each variable. Several measures should be single questions and indexes that were used in prior research (search the Web and the journal literature in Soc Abstracts or Psych Abstracts). Make up a few questions and one index yourself. Ask classmates to answer these questions and give you feedback on their clarity.

4. Propose tests of reliability and validity for four of the measures.

---

Student Study Site

The Companion Website for *The Practice of Research in Criminology and Criminal Justice*, Third Edition

http://www.sagepub.com/prccj3

Visit the Web-based student study site to enhance your understanding of the chapter content and to discover additional resources that will take your learning one step further. You can enhance your understanding of the chapters by using the comprehensive study material, which includes e-flashcards, web exercises, practice self-tests, and more. You will also find special features, such as Learning from Journal Articles, which incorporates SAGE’s online journal collection, and Russ Schutt’s popular Interactive Exercises.
WEB EXERCISES

1. Use the Web to find information regarding alcohol consumption and crime. Write a short report on your findings. How is “alcohol consumption” conceptualized and measured in the various sources you find?

2. How would you define “rape”? Write a brief definition. Based on this conceptualization, what circumstances constitute rape? Describe a method of measurement that would be valid for a study of rape (as you define it). Now go to the Rape Victim Advocates’ Web site at www.RapeVictimAdvocates.org. Go to “Myths & Facts.” Discuss some facts about rape that you were previously unaware of or some myths you believed. Rewrite your definition of rape based on your new knowledge. What additional circumstances constitute rape based on your new conceptualization?

ETHICS EXERCISES

1. In order to measure disorder in Chicago neighborhoods, Sampson and Raudenbush (1999) recorded the street scene with video cameras in a van with darkened windows. Do you judge this measurement procedure as ethical? Refer to each of the guidelines in chapter 2. How could the guideline about “informed consent” be interpreted to permit this type of observational procedure?

2. Both some Homeland Security practices and inadvertent releases of Web searching records have raised new concerns about the use of unobtrusive measures of behavior and attitudes. If all identifying information is removed, do you think criminologists should be able to study who is stopped by police for traffic violations? What types of books are checked out in libraries in different communities? The extent of use of pornography in different cities by analyzing store purchases? How much alcohol different types of people use by linking credit card records to store purchases?

SPSS EXERCISES

HOMICIDE.POR contains a sample of homicide defendants from a sample of 33 U.S. counties for the year 1988.

1. Obtain a frequency distribution for the variables INTIMATE, NUMVICT, and PRIMTIME. At what levels (nominal or categorical, ordinal, interval, ratio) are each of these variables measured?

2. What conclusions do you make about the victim-offender relationship, number of victims, and length of sentence received based on these frequency distributions?