Qualitative and Quantitative Measurement

The Need for Measurement
Quantitative and Qualitative Measurement
The Measurement Process
Reliability and Validity
A Guide to Quantitative Measurement
Scales and Indexes
Conclusion

Measurement, in short, is not an end in itself. Its scientific worth can be appreciated only in an instrumentalist perspective, in which we ask what ends measurement is intended to serve, what role it is called upon to play in the scientific situation, what functions it performs in inquiry.
—Abraham Kaplan, *The Conduct of Inquiry,* p. 171

Who is poor and how much poverty exists? U.S. government officials in the 1960s answered these questions using the poverty line to measure poverty. New programs were to provide aid to poor people (for schooling, health care, housing assistance, and so forth). They began with the idea of being so impoverished that a family was unable to buy enough food to prevent malnourishment. Studies at the time showed that low-income people were spending one-third of their income on food. Officials visited grocery stores and calculated how much low-cost nutritional food for a family would cost and multiplied the amount by 3 to create a poverty line. Since then, the number has been adjusted for inflation. When Brady (2003:730) reviewed publications from 1990–2001, he found that 69.8 percent of poverty studies in the United States used the official government rate. However, numerous studies found that the official U.S. measure of poverty has major deficiencies. When the National Research Council examined the measure in 1995, members declared it outdated and said it should not be retained. The poverty measure sets an arbitrary income level and "it obscures differences in the extent of poverty among population groups and across geographic contexts and provides an inaccurate picture of trends over time" (Brady, 2003:718). It fails to capture the complex nature of poverty and does not take into account new family situations, new aid programs, changes in taxes, and new living expenses. Adding to the confusion, we cannot compare U. S. poverty reduction over time to those in other countries because each country uses different poverty measures. All of the methodological improvements as to how we measure poverty would result in counting far more people as being poor, so few government officials want to change the measure.

THE NEED FOR MEASUREMENT

As researchers, we encounter measures everyday such as the Stanford Binet IQ test to measure intelligence, the index of dissimilarity to measure racial segregation, or uniform crime reports to measure the amount of crime. We need measures to test a hypothesis, evaluate an explanation, provide empirical support for a theory, or study an applied issue. The way we measure a range of social life—aspects such as self-esteem, political power, alienation, or racial prejudice—is the focus of this chapter. We measure in both quantitative and qualitative studies, but quantitative researchers are most concerned with measurement. In *quantitative studies*, measurement is a distinct step in the research process that occurs prior to data collection. Quantitative measurement has a special terminology and set of techniques because the goal is to precisely capture details of the empirical social world and express what we find in numbers.

In *qualitative studies*, we measure with alternatives to numbers, and measurement is less a separate research step. Because the process is more inductive, we are measuring and creating new concepts simultaneously with the process of gathering data.

Measuring is not some arcane, technical issue (like pulling out a tape measure to determine an object's length or putting an object on a scale to check its weight) that we can skip over quickly. Measurement intimately connects how we perceive and think about the social world with what we find in it. Poor-quality measures can quickly destroy an otherwise good study. Measurement also has consequences in everyday life. For example, psychologists and others debate the meaning and measures of intelligence. We use IQ "tests" to measure a person's intelligence in schools, on job applications, and in statements about racial or other inherited superiority. But what is intelligence? Most such IQ "tests" measure only analytic reasoning (i.e., one's capacity to think abstractly and to infer logically). However, we recognize other types of intelligence: artistic, practical, mechanical, and creative. Some people suggest even more types, such as social-interpersonal, emotional, body-kinesthetic, musical, or spatial. If there are many forms of intelligence but we narrowly measure only one type, we limit the way schools identify and nurture learning; the way we select, evaluate, and promote employees; and the way society as a whole values diverse human capabilities.

As the chapter opening indicated, the way we measure poverty determines whether people receive assistance from numerous social programs (e.g., subsidized housing, food aid, health care, childcare). Some say that people are poor if they cannot afford to buy food required to prevent malnutrition. Others say that *poor* means having an annual income that is less than one-half of the average (median) income. Still others say that *poor* means someone who earns less than a "living wage" based on a judgment about an income needed to meet minimal community standards of health, safety and decency in hygiene, housing, clothing, diet, transportation, and so forth. Decisions about measuring poverty can greatly influence the daily living conditions of millions of people.

We use many measures in daily life. For example, this morning I woke up and hopped onto a bathroom scale to see how well my diet is working. I glanced at a thermometer to find out whether to wear a coat. Next, I got into my car and checked the gas gauge to be sure I could make it to campus. As I drove, I watched the speedometer so I would not get a speeding ticket. By 8:00 A.M., I had measured weight, temperature, gasoline volume, and speed—all measures about the physical world. Such precise, well-developed measures of daily life are fundamental in the natural sciences.

Our everyday measures of the nonphysical world are usually less exact. We are measuring when we say that a restaurant has excellent food, that Pablo is really smart, that Karen has a negative attitude toward life, that Johnson is really prejudiced, or that last night's movie contained lots of violence. Such everyday judgments as "really prejudiced" or "lots of violence" are sloppy and imprecise.

Measurement instruments also extend our senses. The astronomer or biologist uses the telescope or the microscope to extend natural vision.

Measuring helps us see what is otherwise invisible, and it lets us observe things that were once unseen and unknown but predicted by theory. For example, we may not see or feel magnetism with our natural senses. Magnetism comes from a theory about the physical world. We see its effects indirectly; for instance, metal flecks move near a magnet. The magnet allows us to "see" or measure the magnetic fields. In contrast to our natural senses, scientific measurement is more sensitive and varies less with the specific observer and yields more exact information. We recognize that a thermometer gives more specific, precise information about temperature than touch can. Likewise, a good bathroom scale gives us more specific, constant, and precise information about the weight of a 5-year-old girl than we can get by lifting her and then calling her "heavy" or "light."

Before we can measure, we need to have a very clear idea about what we are interested in. This is a key principle; measurement connects ideas we carry in our heads with specific things we do in the empirical world to make those ideas visible. Natural scientists use many theories, and they created measures to "see" very tiny things (molecules or insect organs) or very large things (huge geological land masses or planets) that are not observable through ordinary senses. All researchers are constantly creating new measures.[1]

We might easily see age, sex, and race that are measured in social research (e.g., physical wrinkles of age, body parts of each sex, skin tones, and eye shape), but many aspects of the social world (e.g., attitudes, ideology, divorce rates, deviance, social roles) are difficult to observe directly. Just as natural scientists created indirect measures of the "invisible" molecules and the force of gravity, social scientists created measures for difficult-to-observe parts of the social world.

QUANTITATIVE AND QUALITATIVE MEASUREMENT

In all social research—both qualitative and quantitative studies—we connect data to ideas or concepts. We can think of the data in a study as the empirical representation of a concept. Measurement links the data to the concepts, yet the measurement process differs depending on whether our data and research approach are primarily quantitative or qualitative. Three features separate quantitative from qualitative approaches to measurement.

The first difference is timing. In quantitative research, we think about variables and convert them into specific actions during a planning stage that is before and separate from gathering or analyzing data. In qualitative research, we measure while in the data collection phase.

A second difference involves the data itself. In a quantitative study, we use techniques that will produce data in the form of numbers. Usually this happens by moving deductively from abstract ideas to specific data collection techniques, and to precise numerical information that the techniques yield. Numerical data represent a uniform, standardized, and compact way to empirically represent abstract ideas. In a qualitative study, data sometimes come in the form of numbers; more often, the data are written or spoken words, actions, sounds, symbols, physical objects, or visual images (e.g., maps, photographs, videos). Unlike a quantitative study, a qualitative study does not convert all observations into a single, common medium such as numbers but leaves the data in a variety of nonstandard shapes, sizes, and forms. While numerical data convert information into a standard and condensed format, qualitative data are voluminous, diverse, and nonstandard.

A third difference involves how we connect concepts with data. In quantitative research, we contemplate and reflect on concepts before we gather data. We select measurement techniques to bridge the abstract concepts with the empirical data. Of course, after we collect and examine the data, we do not shut off our minds and continue to develop new ideas, but we begin with clearly thought-out concepts and consider how we might measure them.

In qualitative research, we also reflect on concepts before gathering data. However, many of the concepts we use are developed and refined during or after the process of data collection. We reexamine and reflect on the data and concepts simultaneously and interactively. As we gather data, we are simultaneously reflecting on it and generating new

ideas. The new ideas provide direction and suggest new ways to measure. In turn, the new ways to measure shape how we will collect additional data. In short, we bridge ideas with data in an ongoing, interactive process.

To summarize, we think about and make decisions regarding measurement in quantitative studies before we gather data. The data are in a standardized, uniform format: numbers. In contrast, in a qualitative study, most of our thinking and measurement decisions occur in the midst of gathering data, and the data are in a diffuse forms.

THE MEASUREMENT PROCESS

When we measure, we connect an invisible concept, idea, or construct in our minds with a technique, process, or procedure with which we observe the idea in the empirical world.[2] In quantitative studies, we tend to start with abstract ideas and end with empirical data. In qualitative studies, we mix data and ideas while gathering data. However, in a specific study, things are messy and tend to be more interactive than this general statement suggests.

We use two major processes in measurement: conceptualization and operationalization. **Conceptualization** refers to taking an abstract construct and refining it by giving it a conceptual or theoretical definition. A **conceptual definition** is a statement of the idea in your head in specific words or theoretical terms that are linked to other ideas or constructs. There is no magical way to turn a construct into a precise conceptual definition; doing so involves thinking carefully, observing directly, consulting with others, reading what others have said, and trying possible definitions.

A good definition has one clear, explicit, and specific meaning. There is no ambiguity or vagueness. Sometimes conceptualization is highly creative and produces new insights. Some scholarly articles have been devoted to conceptualizing key concepts. Melbin (1978) conceptualized *night* as a frontier, Gibbs (1989) analyzed the meaning of the concept of *terrorism*, and Ball and Curry (1995) discussed what *street gang* means. The key point is this: We need clear, unambiguous definitions of concepts to develop sound explanations.

A single construct can have several definitions, and people may disagree over definitions. Conceptual definitions are linked to theoretical frameworks. For example, a conflict theorist may define *social class* as the power and property that a group of people in society has or lacks. A structural functionalist defines *social class* in terms of individuals who share a social status, lifestyle, or subjective identification. Although people disagree over definitions, we as researchers should always state explicitly which definition we are using.

Some constructs (e.g., alienation) are highly abstract and complex. They contain lower level concepts within them (e.g., powerlessness), which can be made even more specific (e.g., a feeling of little power concerning where one can live). Other constructs are concrete and simple (e.g., age). We need to be aware of how complex and abstract a construct is. For example, it is easier to define a concrete construct such as *age* (e.g., number of years that have passed since birth) than a complex, abstract concept such as *morale.*

Before we can measure, we must distinguish exactly what we are interested in from other nearby things. This is common sense. How can we measure something unless we know what we are looking for? For example, a biologist cannot observe a cancer cell unless he or she first knows what a cancer cell is, has a microscope, and can distinguish the cell from noncell "stuff" under the microscope. The process of measurement involves more than simply having a measurement instrument (e.g., a microscope). We need three things in the measurement process: a construct, a measure, and the ability to recognize what we are looking for.[3]

For example, let us say that I want to measure teacher morale. I must first define *teacher morale.* What does the construct of *morale* mean? As a variable construct, morale takes on different values: high versus low or good versus bad. Next I must

Conceptualization The process of developing clear, rigorous, systematic conceptual definitions for abstract ideas/concepts.

Conceptual definition A careful, systematic definition of a construct that is explicitly written down.

create a measure of my construct. This could take the form of survey questions, an examination of school records, or observations of teachers. Finally, I must distinguish morale from other things in the answers to survey questions, school records, or observations.

The social researcher's job is more difficult than that of the natural scientist because social measurement involves talking with people or observing their behavior. Unlike the planets, cells, or chemicals, the answers people give and their actions can be ambiguous. People can react to the very fact that they are being asked questions or observed. Thus, the social researcher has a double burden: first, to have a clear construct, a good measure, and an ability to recognize what is being looked for, and second, to try to measure fluid and confusing social life that may change just because of an awareness that a researcher is trying to measure.

How can I develop a conceptual definition of *teacher morale*, or at least a tentative working definition to get started? I begin with my everyday understanding of morale: something vague such as "how people feel about things." I ask some of my friends how they define it. I also look at an unabridged dictionary and a thesaurus. They give definitions or synonyms such as "confidence, spirit, zeal, cheerfulness, esprit de corps, mental condition toward something." I go to the library and search the research literature on morale or teacher morale to see how others have defined it. If someone else has already given an excellent definition, I might borrow it (citing the source, of course). If I do not find a definition that fits my purposes, I turn to theories of group behavior, individual mental states, and the like for ideas. As I collect various definitions, parts of definitions, and related ideas, I begin to see the boundaries of the core idea.

By now, I have many definitions and need to sort them out. Most of them say that morale is a spirit, feeling, or mental condition toward something, or a group feeling. I separate the two extremes of my construct. This helps me turn the concept into a variable. High morale involves confidence, optimism, cheerfulness, feelings of togetherness, and willingness to endure hardship for the common good. Low morale is the opposite; it is a lack of confidence, pessimism, depression, isolation, selfishness, and an unwillingness to put forth effort for others.

Because I am interested in *teacher* morale, I learn about teachers to specify the construct to them. One strategy is to make a list of examples of high or low teacher morale. High teacher morale includes saying positive things about the school, not complaining about extra work, or enjoying being with students. Low morale includes complaining a lot, not attending school events unless required to, or looking for other jobs.

Morale involves a feeling toward something else; a person has morale with regard to something. I list the various "somethings" toward which teachers have feelings (e.g., students, parents, pay, the school administration, other teachers, the profession of teaching). This raises an issue that frequently occurs when developing a definition. Are there several types of teacher morale, or are all of these "somethings" aspects of one construct? There is no perfect answer. I have to decide whether morale means a single, general feeling with different parts or dimensions or several distinct feelings.

What unit of analysis does my construct apply to: a group or an individual? Is morale a characteristic of an individual, of a group (e.g., a school), or of both? I decide that for my purposes, morale applies to groups of people. This tells me that my unit of analysis will be a group: all teachers in a school.

I must distinguish the construct of interest from related ideas. How is my construct of teacher morale similar to or different from related concepts? For example, does *morale* differ from *mood?* I decide that mood is more individual and temporary than morale. Likewise, morale differs from optimism and pessimism. Those are outlooks about the future that individuals hold. Morale is a group feeling. It may include positive or negative feelings about the future as well as related beliefs and feelings.

Conceptualization is the process of thinking through the various possible meanings of a construct. By now, I know that teacher morale is a mental state or feeling that ranges from high (optimistic, cheerful) to low (pessimistic, depressed); morale has several dimensions (regarding students, regarding other teachers); it is a characteristic of a group;

and it persists for a period of months. I have a much more specific mental picture of what I want to measure than when I began. If I had not conceptualized, I would have tried to measure what I started with: "how people feel about things."

Even with all of the conceptualization, some ambiguity remains. To complete the conceptualization process, boundaries are necessary. I must decide exactly what I intend to include and exclude. For example, what is a teacher? Does a teacher include guidance counselors, principals, athletic coaches, and librarians? What about student teachers or part-time or substitute teachers? Does the word *teachers* include everyone who teaches for a living, even if someone is not employed by a school (e.g., a corporate trainer, an on-the-job supervisor who instructs an apprentice, a hospital physician who trains residents)? Even if I restrict my definition to people in schools, what is a school? It could include a nursery school, a training hospital, a university's Ph.D. program, a for-profit business that prepares people to take standardized tests, a dog obedience school, a summer camp that teaches students to play basketball, and a vocational school that teaches how to drive semitrailer trucks.

Some people assume *teacher* means a full-time, professionally trained employee of a school teaching grades 1 through 12 who spends most of the day in a classroom with students. Others use a legal or official government definition that could include people certified to teach, even if they are not in classrooms. It excludes people who are uncertified, even if they are working in classrooms with students. The central point is that conceptualization requires me to be very clear in my own thinking. I must know exactly what I mean by *teachers* and *morale* before I can begin to measure. I must state what I think in very clear and explicit terms that other people can understand.

Operationalization links a conceptual definition to a set of measurement techniques or procedures, the construct's **operational definition** (i.e., a definition in terms of the specific operations or actions). An operational definition could be a survey questionnaire, a method of observing events in a field setting, a way to measure symbolic content in the mass media, or any process that reflects, documents, or represents the abstract construct as it is expressed in the conceptual definition.

We often can measure a construct in several ways; some are better and more practical than other ways. The key point is that we must fit the measure to the specific conceptual definition by working with all practical constraints within which we must operate (e.g., time, money, available participants). We can develop a new measure from scratch or use one that other researchers are using (see Expansion Box 1, Five Suggestions for Coming Up with a Measure).

EXPANSION BOX 1

Five Suggestions for Coming Up with a Measure

1. *Remember the conceptual definition.* The underlying principle for any measure is to match it to the specific conceptual definition of the construct that will be used in the study.
2. *Keep an open mind.* Do not get locked into a single measure or type of measure. Be creative and constantly look for better measures. Avoid what Kaplan (1964:28) called the "law of the instrument," which means being locked into using one measurement instrument for all problems.
3. *Borrow from others.* Do not be afraid to borrow from other researchers, as long as credit is given. Good ideas for measures can be found in other studies or modified from other measures.
4. *Anticipate difficulties.* Logical and practical problems often arise when trying to measure variables of interest. Sometimes a problem can be anticipated and avoided with careful forethought and planning.
5. *Do not forget your units of analysis.* Your measure should fit with the units of analysis of the study and permit you to generalize to the universe of interest.

Operationalization The process of moving from a construct's conceptual definition to specific activities or measures that allow a researcher to observe it empirically.

Operational definition A variable in terms of the specific actions to measure or indicate it in the empirical world.

Operationalization connects the language of theory with the language of empirical measures. Theory has many abstract concepts, assumptions, definitions, and cause-and-effect relations. By contrast, empirical measures are very concrete actions in specific, real situations with actual people and events. Measures are specific to the operations or actions we engage in to indicate the presence or absence of a construct as it exists in concrete, observable reality.

Quantitative Conceptualization and Operationalization

Quantitative measurement proceeds in a straightforward sequence: first conceptualization, next operationalization, and then application of the operational definition or the collection of data. We must rigorously link abstract ideas to measurement procedures that can produce precise information in the form of numbers. One way to do this is with rules of correspondence or an auxiliary theory. The purpose of the rules is to link the conceptual definitions of constructs to concrete operations for measuring the constructs.[4]

Rules of correspondence are logical statements of the way an indicator corresponds to an abstract construct. For example, a rule of correspondence says that we will accept a person's verbal agreement with a set of ten specific statements as evidence that the person strongly holds an antifeminist attitude. This auxiliary theory may explain how and why indicators and constructs connect. Carmines and Zeller (1979:11) noted, "The auxiliary theory specifying the relationship between concepts and indicators is equally important to social research as the substantive theory linking concepts to one another." Perhaps we want to measure alienation. Our definition of the alienation has four parts, each in a different sphere of life: family relations, work relations, relations with community, and relations with friends. An auxiliary theory may specify that certain behaviors or feelings in each sphere of life are solid evidence of alienation. In the sphere of work, the theory says that if a person feels a total lack of control over when, where, and with whom he or she works, what he or she does when working, or how fast he or she must work, that person is alienated.

Figure 1 illustrates the measurement process linking two variables in a theory and a hypothesis. We must consider three levels: conceptual, operational, and empirical.[5] At the most abstract level, we may be interested in the causal relationship between two constructs, or a **conceptual hypothesis**. At the level of operational definitions, we are interested in testing an **empirical hypothesis** to determine the degree of association between indicators. This is the level at which we consider correlations, statistics, questionnaires, and the like. The third level is the empirical reality of the lived social world. As we link the operational indicators (e.g., questionnaire items) to a construct (e.g., alienation), we capture what is taking place in the lived social world and relate it back to the conceptual level.

As we measure, we link the three levels together and move deductively from the abstract to the concrete. First, we conceptualize a variable, giving it a clear conceptual definition; next we operationalize it by developing an operational definition or set of indicators for it; and lastly, we apply indicators to collect data and test empirical hypotheses.

Let us return to the example mentioned earlier. How do I give my teacher morale construct an operational definition? First, I read the research reports of others and see whether a good indicator already exists. If there are no existing indicators, I must invent one from scratch. Morale is a mental state or feeling, so I measure it indirectly through people's words and actions. I might develop a questionnaire

Rules of correspondence Standards that researchers use to connect abstract constructs with measurement operations in empirical social reality.

Conceptual hypothesis A type of hypothesis that expresses variables and the relationships among them in abstract, conceptual terms.

Empirical hypothesis A type of hypothesis in which the researcher expresses variables in specific empirical terms and expresses the association among the measured indicators in observable, empirical terms.

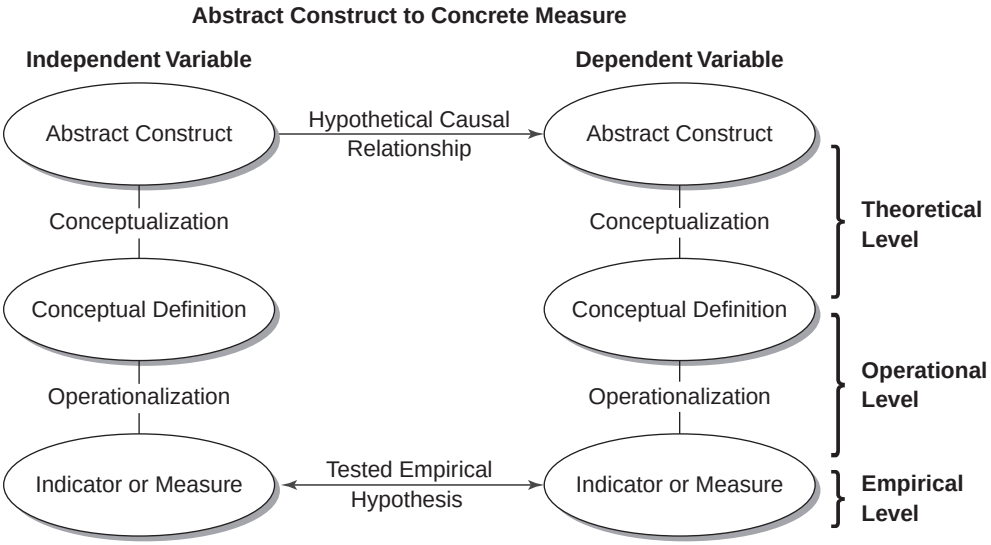


FIGURE 1 Conceptualization and Operationalization

for teachers and ask them about their feelings toward the dimensions of morale in my definition. I might go to the school and observe the teachers in the teachers lounge, interacting with students, and attending school activities. I might use school personnel records on teacher behaviors for statements that indicate morale (e.g., absences, requests for letters of recommendation for other jobs, performance reports). I might survey students, school administrators, and others to find out what they think about teacher morale. Whichever indicator I choose, I further refine my conceptual definition as I develop it (e.g., write specific questionnaire questions).

Conceptualization and operationalization are necessary for each variable. In the preceding example, morale is one variable, not a hypothesis. It could be a dependent variable caused by something else, or it could be an independent variable causing something else. It depends on my theoretical explanation.

Qualitative Conceptualization and Operationalization

Conceptualization. In qualitative research, instead of refining abstract ideas into theoretical definitions early in the research process, we refine rudimentary "working ideas" during the data collection and analysis process. *Conceptualization* is a process of forming coherent theoretical definitions as we struggle to "make sense" or organize the data and our preliminary ideas about it.

As we gather and analyze qualitative data, we develop new concepts, formulate definitions for major constructs, and consider relationships among them. Eventually, we link concepts and constructs to create theoretical relationships. We form and refine constructs while examining data (e.g., field notes, photos and maps, historical documents), and we ask theoretical questions about the data (e.g., Is this a case of class conflict? What is the sequence of events and could it be different? Why did this happen here but not somewhere else?).

We need clear, explicit definitions expressed in words and descriptions of specific actions that link to other ideas and are tied to the data. In qualitative research, conceptualization flows largely from the data.

Operationalization. In qualitative studies, operationalization often precedes conceptualization

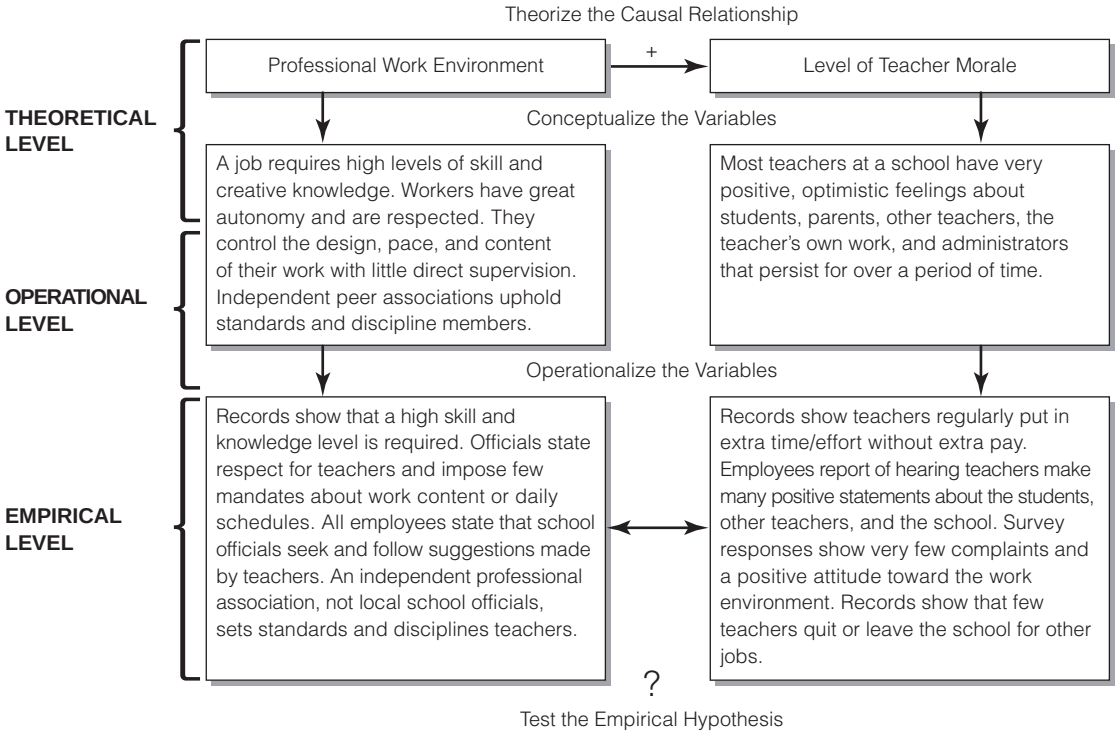


FIGURE 2 Example of the Deductive Measurement Process for the Hypothesis: A Professional Work Environment Increases the Level of Teacher Morale

(see Figure 2) and gives deductive measurement (see Figure 3 for inductive measurement). We may create conceptual definitions out of rudimentary "working ideas" while we are making observations or gathering data. Instead of turning refined conceptual definitions into measurement operations, we operationalize by describing how specific observations and thoughts about the data contribute to working ideas that are the basis of conceptual definitions.

Thus, qualitative research operationalization largely involves developing a description of how we use working ideas while making observations. Oerationalization describes how we gathered specific observations or data and we struggled to understand the data as the data evolved into abstract constructs. In this way, qualitative operationalization is more an after-the-fact description than a preplanned technique.

Just as quantitative operationalization deviates from a rigid deductive process, qualitative researchers may draw on ideas from beyond the data of a specific research setting. Qualitative operationalization includes using preexisting techniques and concepts that we blend with those that emerged during the data collection process.

Fantasia's (1988) field research on contested labor actions illustrates qualitative operationalization. Fantasia used *cultures of solidarity* as a central construct. He related this construct to ideas of conflict-filled workplace relations and growing class consciousness among nonmanagerial workers. He defined a culture of solidarity as a type of cultural expression created by workers that evolves in particular places over time. The workers over time develop shared feelings and a sense of unity that is in opposition to management and business owners. It is an interactive process. Slowly over

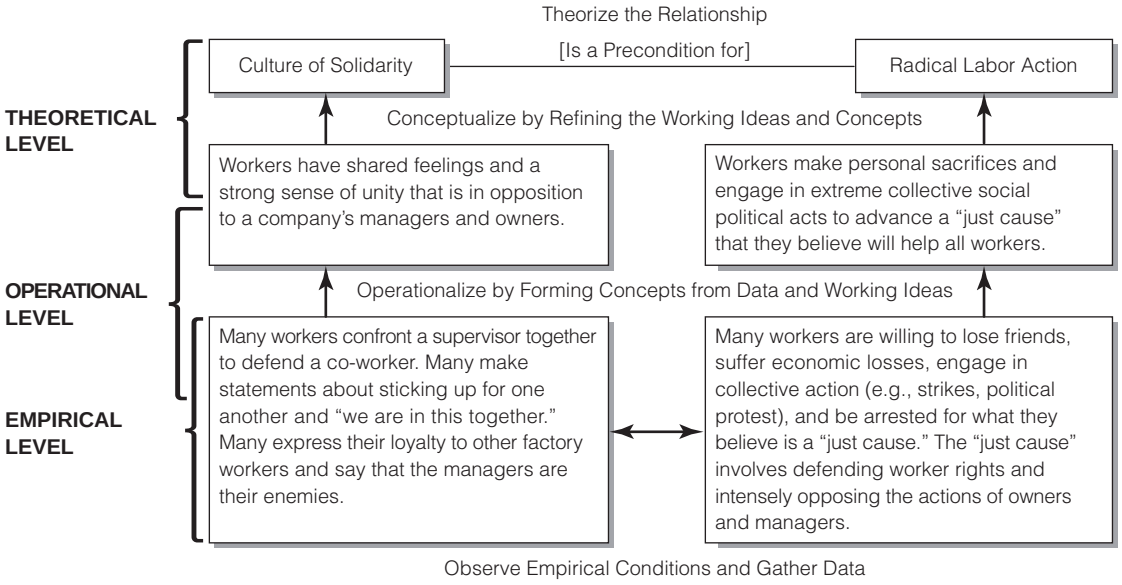


FIGURE 3 Example of the Inductive Measurement Process for the Proposition: Radical Labor Action Is Likely to Occur Where a Culture of Solidarity Has Been Created

time, the workers arrive at common ideas, understandings, and actions. It is "less a matter of disembodied mental attitude than a broader set of practices and repertoires available for empirical investigation" (Fantasia:14).

To operationalize the construct, Fantasia describes how he gathered data. He presents them to illustrate the construct, and explains his thinking about the data. He describes his specific actions to collect the data (e.g., he worked in a particular factory, attended a press conference, and interviewed people). He also shows us the data in detail (e.g., he describes specific events that document the construct by showing several maps indicating where people stood during a confrontation with a foreperson, retelling the sequence of events at a factory, recounting actions by management officials, and repeating statements that individual workers made). He gives us a look into his thinking process as he reflected and tried to understand his experiences and developed new ideas drawing on older ideas.

Casing. In qualitative research, ideas and evidence are mutually interdependent. This applies particularly to case study analysis. Cases are not given preestablished empirical units or theoretical categories apart from data; they are defined by data and theory. By analyzing a situation, the researcher organizes data and applies ideas simultaneously to create or specify a case. Making or creating a case, called **casing**, brings the data and theory together. Determining what to treat as a case resolves a tension or strain between what the researcher observes and his or her ideas about it. "Casing, viewed as a methodological step, can occur at any phase of the research process, but occurs especially at the beginning of the project and at the end" (Ragin, 1992b:218).

RELIABILITY AND VALIDITY

All of us as researchers want reliability and validity, which are central concerns in all measurement. Both connect measures to constructs. It is not

Casing Developing cases in qualitative research.

possible to have perfect reliability and validity, but they are ideals toward which we strive. Reliability and validity are salient because our constructs are usually ambiguous, diffuse, and not observable. Reliability and validity are ideas that help to establish the truthfulness, credibility, or believability of findings. Both terms also have multiple meanings. As used here, they refer to related, desirable aspects of measurement.

Reliability means dependability or consistency. It suggests that the same thing is repeated or recurs under the identical or very similar conditions. The opposite of reliability is an erratic, unstable, or inconsistent result that happens because of the measurement itself. *Validity* suggests truthfulness. It refers to how well an idea "fits" with actual reality. The absence of validity means that the fit between the ideas we use to analyze the social world and what actually occurs in the lived social world is poor. In simple terms, validity addresses the question of how well we measure social reality using our constructs about it.

All researchers want reliable and valid measurement, but beyond an agreement on the basic ideas at a general level, qualitative and quantitative researchers see reliability and validity differently.

Reliability and Validity in Quantitative Research

Reliability. **Measurement reliability** means that the numerical results an indicator produces do not vary because of characteristics of the measurement process or measurement instrument itself. For example, I get on my bathroom scale and read my weight. I get off and get on again and again. I have a reliable scale if it gives me the same weight each time, assuming, of course, that I am not eating, drinking, changing clothing, and so forth. An unreliable scale registers different weights each time, even though my "true" weight does not change. Another example is my car speedometer. If I am driving at a constant slow speed on a level surface but the speedometer needle jumps from one end to the other, the speedometer is not a reliable indicator of how fast I am traveling. Actually, there are three types of reliability.[6]

Measurement reliability The dependability or consistency of the measure of a variable.

Stability reliability Measurement reliability across time; a measure that yields consistent results at different time points assuming what is being measured does not itself change.

Representative reliability Measurement reliability across groups; a measure that yields consistent results for various social groups.

Three Types of Reliability

1. Stability reliability is reliability across time. It addresses the question: Does the measure deliver the same answer when applied in different time periods? The weight-scale example just given is of this type of reliability. Using the test-retest method can verify an indicator's degree of stability reliability. Verification requires retesting or readministering the indicator to the same group of people. If what is being measured is stable and the indicator has stability reliability, then I will have the same results each time. A variation of the test-retest method is to give an alternative form of the test, which must be very similar to the original. For example, I have a hypothesis about gender and seating patterns in a college cafeteria. I measure my dependent variable (seating patterns) by observing and recording the number of male and female students at tables, and noting who sits down first, second, third, and so on for a 3-hour period. If, as I am observing, I become tired or distracted or I forget to record and miss more people toward the end of the 3 hours, my indicator does not have a high degree of stability reliability.

2. Representative reliability is reliability across subpopulations or different types of cases. It addresses the question: Does the indicator deliver the same answer when applied to different groups? An indicator has high representative reliability if it yields the same result for a construct when applied to different subpopulations (e.g., different classes, races, sexes, age groups). For example, I ask a question about a person's age. If people in their twenties answered my question by overstating their true age

whereas people in their fifties understated their true age, the indicator has a low degree of representative reliability. To have representative reliability, the measure needs to give accurate information for every age group.

A *subpopulation analysis* verifies whether an indicator has this type of reliability. The analysis compares the indicator across different subpopulations or subgroups and uses independent knowledge about them. For example, I want to test the representative reliability of a questionnaire item that asks about a person's education. I conduct a subpopulation analysis to see whether the question works equally well for men and women. I ask men and women the question and then obtain independent information (e.g., check school records) and check to see whether the errors in answering the question are equal for men and women. The item has representative reliability if men and women have the same error rate.

3. Equivalence reliability applies when researchers use **multiple indicators**—that is, when a construct is measured with multiple specific measures (e.g., several items in a questionnaire all measure the same construct). Equivalence reliability addresses the question: Does the measure yield consistent results across different indicators? If several different indicators measure the same construct, then a reliable measure gives the same result with all indicators.

We verify equivalence reliability with the *split-half method.* This involves dividing the indicators of the same construct into two groups, usually by a random process, and determining whether both halves give the same results. For example, I have fourteen items on a questionnaire. All measure political conservatism among college students. If my indicators (i.e., questionnaire items) have equivalence reliability, then I can randomly divide them into two groups of seven and get the same results. For example, I use the first seven questions and find that a class of fifty business majors is twice as conservative as a class of fifty education majors. I get the same results using the second seven questions. Special statistical measures (e.g., Cronbach's alpha) also can determine this type of reliability. A special type of equivalence reliability, intercoder reliability, can be used when there are several observers, raters, or coders of information. In a sense, each observer is an indicator. A measure is reliable if the observers, raters, or coders agree with each other. This measure is a common type of reliability reported in content analysis studies. For example, I hire six students to observe student seating patterns in a cafeteria. If all six are equally skilled at observing and recording, I can combine the information from all six into a single reliable measure. But if one or two students are lazy, inattentive, or sloppy, my measure will have lower reliability. Intercoder reliability is tested by having several coders measure the exact same thing and then comparing the measures. For instance, I have three coders independently code the seating patterns during the same hour on three different days. I compare the recorded observations. If they agree, I can be confident of my measure's intercoder reliability. Special statistical techniques measure the degree of intercoder reliability.

How to Improve Reliability. It is rare to have perfect reliability. We can do four things to improve reliability: (1) clearly conceptualize constructs, (2) use a precise level of measurement, (3) use multiple indicators, and (4) use pilot tests.

1. *Clearly conceptualize all constructs.* Reliability increases when each measure indicates one and only one concept. This means we must develop unambiguous, clear theoretical definitions. Constructs should be specified to eliminate "noise" (i.e., distracting or interfering information) from other constructs. For example, the indicator of a pure chemical compound is more reliable than the indicator in which the chemical is mixed with other material or dirt. In the latter case, separating the

Equivalence reliability Measurement reliability across indicators; a measure that yields consistent results using different specific indicators, assuming that all measure the same construct.

Multiple indicators The use of multiple procedures or several specific measures to provide empirical evidence of the levels of a variable.

"noise" of other material from the pure chemical is difficult.

Let us return to the example of teacher morale. I should separate morale from related ideas (e.g., mood, personality, spirit, job attitude). If I did not do this, I could not be sure what I was really measuring. I might develop an indicator for morale that also indicates personality; that is, the construct of personality contaminates that of morale and produces a less reliable indicator. Bad measurement occurs by using one indicator to operationalize different constructs (e.g., using the same questionnaire item to indicate morale and personality).

2. *Increase the level of measurement.* Levels of measurement are discussed later in this chapter. Indicators at higher or more precise levels of measurement are more likely to be reliable than less precise measures because the latter pick up less detailed information. If more specific information is measured, it is less likely that anything other than the construct will be captured. The general principle is: Try to measure at the most precise level possible. However, quantifying at higher levels of measurement is more difficult. For example, if I have a choice of measuring morale as either high or low, or in ten categories from extremely low to extremely high, it would be better to measure it in ten refined categories.

3. *Use multiple indicators of a variable.* A third way to increase reliability is to use multiple indicators because two (or more) indicators of the same construct are better than one.[7] Figure 4 illustrates the use of multiple indicators in hypothesis testing. Three indicators of the one independent variable construct are combined into an overall measure, A, and two indicators of a dependent variable are combined into a single measure, B. For example, I have three specific measures of A, which is teacher morale: (a1) the answers to a survey question on attitudes about school, (a2) the number of absences for reasons other than illness and (a3) the number of complaints others heard made by a teacher. I also have two measures of my dependent variable B, giving students extra attention: (b1) number of hours a teacher spends staying after school hours to meet individually with students and (b2) whether the teacher inquires frequently about a student's progress in other classes.

With multiple indicators, we can build on triangulation and take measurements from a wider range of the content of a conceptual definition (i.e., sample from the conceptual domain). We can measure different aspects of the construct with its own indicator. Also, one indicator may be imperfect, but several measures are less likely to have the same error. James (1991) provides a good example of this principle applied to counting persons who are homeless. If we consider only where people sleep (e.g., using sweeps of streets and parks and counting people in official shelters), we miss some because many people who are homeless have temporary shared housing (e.g., sleep on the floor of a friend or family member). We also miss some by using records of official service agencies because

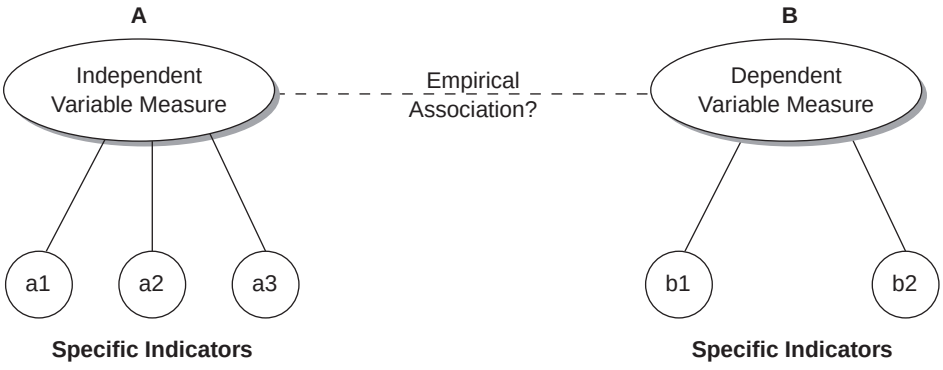


FIGURE 4 Measurement Using Multiple Indicators

many people who are homeless avoid involvement with government and official agencies. However, if we combine the official records with counts of people sleeping in various places and conduct surveys of people who use a range of services (e.g., street clinics, food lines, temporary shelters), we can get a more accurate picture of the number of people who are homeless. In addition to capturing the entire picture, multiple indicator measures tend to be more stable than single item measures.

4. *Use pilot studies and replication*. You can improve reliability by first using a pilot version of a measure. Develop one or more draft or preliminary versions of a measure and try them before applying the final version in a hypothesis-testing situation. This takes more time and effort. Returning to the example discussed earlier, in my survey of teacher morale, I go through many drafts of a question before the final version. I test early versions by asking people the question and checking to see whether it is clear.

The principle of using pilot tests extends to replicating the measures from researchers. For example, I search the literature and find measures of morale from past research. I may want to build on and use a previous measure if it is a good one, citing the source, of course. In addition, I may want to add new indicators and compare them to the previous measure (see Example Box 1, Improving the Measure of U.S. Religious Affiliation). In this way, the quality of the measure can improve over time as long as the same definition is used (see Table 1 for a summary of reliability and validity types).

Validity. Validity is an overused term. Sometimes, it is used to mean "true" or "correct." There are several general types of validity. Here we are concerned with **measurement validity**, which also has several types. Nonmeasurement types of validity are discussed later.

When we say that an indicator is valid, it is valid for a particular purpose and definition. The same indicator may be less valid or invalid for other purposes. For example, the measure of morale discussed above (e.g., questions about feelings toward school) might be valid for measuring morale among teachers but invalid for measuring morale among police officers.[8]

EXAMPLE BOX 1

Improving the Measure of U.S. Religious Affiliation

Quantitative researchers measure individual religious beliefs (e.g., Do you believe in God? in a devil? in life after death? What is God like to you?), religious practices (e.g., How often do you pray? How frequently do you attend services?), and religious affiliation (e.g., If you belong to a church or religious group, which one?). They have categorized the hundreds of U.S. religious denominations into either a three-part grouping (Protestant, Catholic, Jewish) or a three-part classification of fundamentalist, moderate, or liberal that was introduced in 1990.

Steensland and colleagues (2000) reconceptualized affiliation, and, after examining trends in religious theology and social practices, argued for classifying all American denominations into six major categories: Mainline Protestant, Evangelical Protestant, Black Protestant, Roman Catholic, Jewish, and Other (including Mormon, Jehovah's Witnesses, Muslim, Hindu, and Unitarian). The authors evaluated their new six-category classification by examining people's religious views and practices as well as their views about contemporary social issues. Among national samples of Americans, they found that the new classification better distinguished among religious denominations than did previous measures.

At its core, measurement validity tells us how well the conceptual and operational definitions mesh with one other: The better the fit, the higher is the measurement validity. Validity is more difficult to achieve than reliability. We cannot have absolute confidence about validity, but some measures are *more valid* than others. The reason is that constructs are abstract ideas, whereas indicators refer to concrete observation. This is the gap between our mental pictures about the world and the specific

Measurement validity How well an empirical indicator and the conceptual definition of the construct that the indicator is supposed to measure "fit" together.

TABLE 1 Summary of Measurement Reliability and Validity Types

RELIABILITY (DEPENDABLE MEASURE)	VALIDITY (TRUE MEASURE)
Stability—over time (verify using test-retest method)	Face—makes sense in the judgment of others
Representative—across subgroups (verify using split-half method)	Content—captures the entire meaning
Equivalence—across indicators (verify using subpopulation analysis)	Criterion—agrees with an external source ■ Concurrent—agrees with a preexisting measure ■ Predictive—agrees with future behavior Construct—has consistent multiple indicators ■ Convergent—alike ones are similar ■ Discriminant—different ones differ

things we do at particular times and places. Validity is part of a dynamic process that grows by accumulating evidence over time, and without it, all measurement becomes meaningless.

Some researchers use rules of correspondence (discussed earlier) to reduce the gap between abstract ideas and specific indicators. For example, a rule of correspondence is: A teacher who agrees with statements that "things have gotten worse at this school in the past 5 years" and that "there is little hope for improvement" is indicating low morale. Some researchers talk about the *epistemic correlation*, a hypothetical correlation between an indicator and the construct that the indicator measures. We cannot empirically measure such correlations, but they can be estimated.[9]

Four Types of Measurement Validity.

1. Face validity is the most basic and easiest type of validity to achieve. It is a judgment by the scientific community that the indicator really measures the construct. It addresses the question: On the face of it, do people believe that the definition and method of measurement fit? For example, few people would accept a measure of college student math ability by asking students what 2 + 2 equals. This is not a valid measure of college-level math ability on the face of it. Recall that the principle of organized skepticism in the scientific community means that others scrutinize aspects of research.[10]

2. Content validity addresses this question: Is the full content of a definition represented in a measure? A conceptual definition holds ideas; it is a "space" containing ideas and concepts. Measures should sample or represent all ideas or areas in the conceptual space. Content validity involves three steps. First, specify the content in a construct's definition. Next, sample from all areas of the definition. Finally, develop one or more indicators that tap all of the parts of the definition.

Let us consider an example of content validity. I define *feminism* as a person's commitment to a set of beliefs creating full equality between men and women in areas of the arts, intellectual pursuits, family, work, politics, and authority relations. I create a measure of feminism in which I ask two survey questions: (1) Should men and women get equal pay for equal work? and (2) Should men and women share household tasks? My measure has low content validity because the two questions ask only

Face validity A type of measurement validity in which an indicator "makes sense" as a measure of a construct in the judgment of others, especially in the scientific community.

Content validity A type of measurement validity that requires that a measure represent all aspects of the conceptual definition of a construct.

about pay and household tasks. They ignore the other areas (intellectual pursuits, politics, authority relations, and other aspects of work and family). For a content-valid measure, I must either expand the measure or narrow the definition.[11]

3. Criterion validity uses some standard or criterion to indicate a construct accurately. The validity of an indicator is verified by comparing it with another measure of the same construct in which a researcher has confidence. The two subtypes of this type of validity are concurrent and predictive.[12]

To have **concurrent validity**, we need to associate an indicator with a preexisting indicator that we already judge to be valid (i.e., it has face validity). For example, we create a new test to measure intelligence. For it to be concurrently valid, it should be highly associated with existing IQ tests (assuming the same definition of intelligence is used). This means that most people who score high on the old measure should also score high on the new one, and vice versa. The two measures may not be perfectly associated, but if they measure the same or a similar construct, it is logical for them to yield similar results.

Criterion validity by which an indicator predicts future events that are logically related to a construct is called **predictive validity**. It cannot be used for all measures. The measure and the action predicted must be distinct from but indicate the same construct. Predictive measurement validity should not be confused with prediction in hypothesis testing in which one variable predicts a different variable in the future. For example, the Scholastic Assessment Test (SAT) that many U.S. high school students take measures scholastic aptitude: the ability of a student to perform in college. If the SAT has high predictive validity, students who achieve high SAT scores will subsequently do well in college. If students with high scores perform at the same level as students with average or low scores, the SAT has low predictive validity.

Another way to test predictive validity is to select a group of people who have specific characteristics and predict how they will score (very high or very low) vis-à-vis the construct. For example, I create a measure of political conservatism. I predict that members of conservative groups (e.g., John Birch Society, Conservative Caucus, Daughters of the American Revolution, Moral Majority) will score high on it whereas members of liberal groups (e.g., Democratic Socialists, People for the American Way, Americans for Democratic Action) will score low. I "validate" it by pilot-testing it on members of the groups. It can then be used as a measure of political conservatism for the public.

4. Construct validity is for measures with multiple indicators. It addresses this question: If the measure is valid, do the various indicators operate in a consistent manner? It requires a definition with clearly specified conceptual boundaries. The two types of construct validity are convergent and discriminant.

Convergent validity applies when multiple indicators converge or are associated with one another. It means that multiple measures of the same construct hang together or operate in similar ways. For example, I measure the construct "education" by asking people how much education they have completed, looking up school records, and asking the people to complete a test of school knowledge. If the measures do not converge (i.e., people who claim to have a college degree but have no records of attending college or those with college degrees perform no better than high school dropouts on my tests), my measure has weak convergent validity, and I should not combine all three indicators into one measure.

Criterion validity Measurement validity that relies on some independent, outside verification.

Concurrent validity Measurement validity that relies on a preexisting and already accepted measure to verify the indicator of a construct.

Predictive validity Measurement validity that relies on the occurrence of a future event or behavior that is logically consistent to verify the indicator of a construct.

Construct validity A type of measurement validity that uses multiple indicators and has two subtypes: how well the indicators of one construct converge or how well the indicators of different constructs diverge.

Convergent validity A type of measurement validity for multiple indicators based on the idea that indicators of one construct will act alike or converge.

Discriminant validity is the opposite of convergent validity and means that the indicators of one construct "hang together," or converge, but also are negatively associated with opposing constructs. Discriminant validity says that if two constructs *A* and *B* are very different, measures of *A* and *B* should not be associated. For example, I have ten items that measure political conservatism. People answer all ten in similar ways. But I also put five questions that measure political liberalism on the same questionnaire. My measure of conservatism has discriminant validity if the ten conservatism items converge and are negatively associated with the five liberalism ones. (See Figure 5 for a review of measurement validity.)

Reliability and Validity in Qualitative Research

Qualitative research embraces the core principles of reliability and validity, but we rarely see the terms in this approach because they are so closely associated with quantitative measurement. In addition, in qualitative studies, we apply the principles differently.

Reliability. Recall that *reliability* means dependability or consistency. We use a wide variety of techniques (e.g., interviews, participation, photographs, document studies) to record observations consistently in qualitive studies. We want to be consistent (i.e., not vacillating or being erratic) in how we make observations, similar to the idea of stability reliability. One difficulty with reliability is that we often study processes that are unstable over time. Moreover, we emphasize the value of a changing or developing interaction between us as researchers and the people we study. We believe that the subject matter and our relationship to it is an evolving process. A metaphor for the relationship is one of an evolving relationship or living organism (e.g., a plant) that naturally matures over time. Many qualitative researchers see the quantitative approach to reliability as a cold, fixed mechanical instrument that one applies repeatedly to static, lifeless material.

In qualitative studies, we consider a range of data sources and employ multiple measurement methods. We do not become locked into the quantitative-positivist ideas of replication, equivalence, and subpopulation reliability. We accept that different researchers or researchers who use alternative measures may find distinctive results. This happens because data collection is an interactive process in which particular researchers operate in an evolving setting whose context dictates using a unique mix of measures that cannot be repeated. The diverse measures and interactions with different researchers are beneficial because they can illuminate different facets or dimensions of a subject matter. Many qualitative researchers question the quantitative researcher's quest for standard, fixed measures and fear that such measures ignore the benefits of having a variety of researchers with many approaches and may neglect key aspects of diversity that exist in the social world.

Validity. *Validity* means truthfulness. In qualitative studies, we are more interested in achieving authenticity than realizing a single version of "Truth." *Authenticity* means offering a fair, honest, and balanced account of social life from the viewpoint of the people who live it every day. We are less concerned with matching an abstract construct to empirical data than with giving a candid portrayal of social life that is true to the lived experiences of the people we study. In most qualitative studies, we emphasize capturing an inside view and providing a detailed account of how the people we study understand events (see Expansion Box 2, Meanings of Validity in Qualitative Research).

There are qualitative research substitutes for the quantitative approach to validity: ecological validity or natural history methods. Both emphasize conveying an insider's view to others. Historical researchers use internal and external criticisms to determine whether the evidence is real. Qualitative researchers adhere to the core principle of validity, to be truthful (i.e., avoid false or distorted accounts) and try to create a tight fit between understandings, ideas, and statements about the social world and what is actually occurring in it.

Discriminant validity A type of measurement validity for multiple indicators based on the idea that indicators of different constructs diverge.

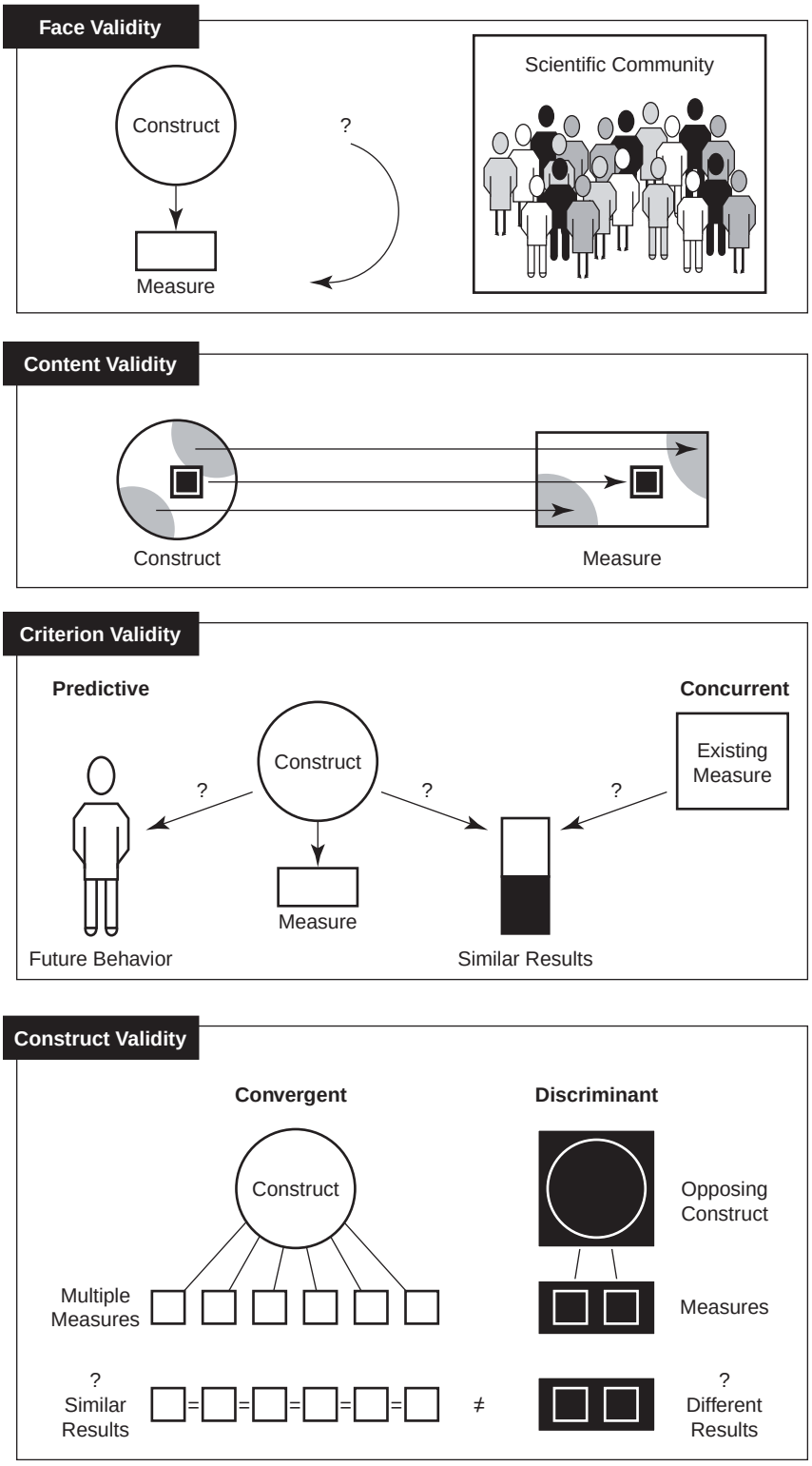


FIGURE 5 Types of Validity

EXPANSION BOX 2

Meanings of Validity in Qualitative Research

Measurement validity in qualitative research does not require demonstrating a fixed correspondence between a carefully defined abstract concept and a precisely calibrated measure of its empirical appearance. Other features of the research measurement process are important for establishing validity.

First, to be considered valid, a researcher's truth claims need to be plausible and, as Fine (1999) argued, intersubjectively "good enough" (i.e., understandable by many other people). *Plausible* means that the data and statements about it are not exclusive; they are not the only possible claims, nor are they exact accounts of the one truth in the world. This does not make them mere inventions or arbitrary. Instead, they are powerful, persuasive descriptions that reveal a researcher's genuine experiences with the empirical data.

Second, a researcher's empirical claims gain validity when supported by numerous pieces of diverse empirical data. Any one specific empirical detail alone may be mundane, ordinary, or "trivial." Validity arises out of the cumulative impact of hundreds of small, diverse details that only together create a heavy weight of evidence.

Third, validity increases as researchers search continuously in diverse data and consider the connections among them. Raw data in the natural social world are not in neatly prepackaged systematic scientific concepts; rather, they are numerous disparate elements that "form a dynamic and coherent ensemble" (Molotch et al., 2000:816). Validity grows as a researcher recognizes a dense connectivity in disparate details. It grows with the creation of a web of dynamic connections across diverse realms, not only with the number of specifics that are connected.

Relationship between Reliability and Validity

Reliability is necessary for validity and is easier to achieve than validity. Although reliability is necessary to have a valid measure of a concept, it does not guarantee that the measure will be valid. It is not a sufficient condition for validity. A measure can yield a result over and over (i.e., has reliability), but what it truly measures may not match a construct's definition (i.e., validity).

For example, I get on a scale to check my weight. The scale registers the same weight each time I get on and off during a 2-hour period. I next go to another scale—an "official" one at a medical clinic—and it reports my weight to be twice as much. The first scale yielded reliable (i.e., dependable and consistent) results, but it was not a valid measure of my weight. A diagram might help you see the relationship between reliability and validity. Figure 6 illustrates the relationship between the concepts by using the analogy of a target. The bull's-eye represents a fit between a measure and the definition of the construct.

Validity and reliability are usually complementary concepts, but in some situations, they conflict with each other. Sometimes, as validity increases, reliability becomes more difficult to attain and vice versa. This situation occurs when the construct is highly abstract and not easily observable but captures the "true essence" of an idea. Reliability is easiest to achieve when a measure is precise, concrete, and observable. For example, *alienation* is a very abstract, subjective construct. We may define it as a deep inner sense of loss of one's core humanity; it is a feeling of detachment and being without purpose that diffuses across all aspects of life (e.g., the sense of self, relations with other people, work, society, and even nature). While it is not easy, most of us can grasp the idea of alienation, a directionless disconnection that pervades a person's existence. As we get more deeply into the true meaning of the concept, measuring it precisely becomes more difficult. Specific questions on a questionnaire may produce reliable measures more than other methods, yet the questions cannot capture the idea's essence.

Other Uses of the Words *Reliable* and *Valid*

Many words have multiple definitions, creating confusion among various uses of the same word. This happens with reliability and validity. We use *reliability* in everyday language. A *reliable* person

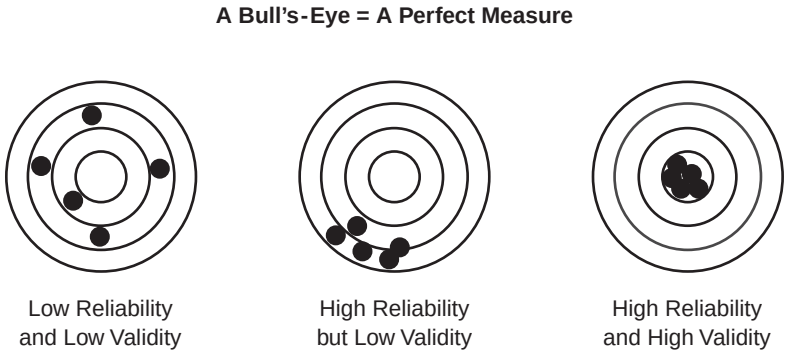


FIGURE 6 Illustration of Relationship between Reliability and Validity

Source: Adapted version of Figure 5-2 An Analogy to Validity and Reliability, page 155 from Babbie, E. R. 1986. *The Practice of Social Research,* Fourth Edition. Belmont, CA: Wadsworth Publishing Company.

is a dependable, stable, and responsible person who responds in similar, predictable ways in different times and conditions. A *reliable* car is dependable and trustworthy; it starts and peforms in a predicable way. Sometimes, we say that a study or its results are *reliable*. This means that other researchers can reproduce the study and will get similar results.

Internal validity means we have not made errors internal to the design of a research project that might produce false conclusions.[13] In experimental research, we primarily talk about possible alternative causes of results that arise despite our attempts to institute controls.

External validity is also used primarily in experimental research. It refers to whether we can generalize a result that we found in a specific setting with a particular small group beyond that situation or externally to a wider range of settings and many different people. External validity addresses this question: If something happens in a laboratory or among a particular set of research participants (e.g., college students), does it also happen in the "real" (nonlaboratory) world or among the general population (nonstudents)? External validity has serioius implications for evaluating theory. If a general theory is true, it implies that we can generalize findings from a single test of the theory to many other situations and populations (see Lucas, 2003).

Statistical validity means that we used the proper statistical procedure for a particular purpose and have met the procedure's mathematical requirements. This validity arises because different statistical tests or procedures are appropriate for different situations as is discussed in textbooks on statistical procedures. All statistical procedures rest on assumptions about the mathematical properties of the numbers being used. A statistic will yield nonsense results if we use it for inappropriate situations or seriously violate its assumptions even if the computation of the numbers is correct. This is why we must know the purposes for which a statistical procedure is designed and its assumptions to use it. This is also why computers can do correct computations but produce output that is nonsense.

A GUIDE TO QUANTITATIVE MEASUREMENT

Thus far, we have discussed principles of measurement. Quantitative researchers have specialized measures that assist in the process of creating operational definitions for reliable and valid measures. This section of the chapter is a brief guide to these ideas and a few of the specific measures.

Levels of Measurement

We can array possible measures on a continuum. At one end are at "higher" ones. These measures contain a great amount of highly specific information with many exact and refined distinctions. At the

opposite end are "lower" ones. These are rough, less precise measures with minimal information and a few basic distinctions. The level of measurement affects how much we can learn when we measure features of the social world and limits the types of indicator we can use as we try to capture empirical details about a construct.

The **level of measurement** is determined by how refined, exact, and precise a construct is in our assumptions about it. This means that how we conceptualize a construct carries serious implications. It influences how we can measure the construct and restricts the range of statistical procedures that we can use after we have gathered data. Often we see a trade-off between the level of measurement and the ease of measuring. Measuring at a low level is simpler and easier than it is at a high level; however, a low level of measurement offers us the least refined information and allows the fewest statistical procedures during data analysis. We can look at the issue in two ways: (1) continuous versus discrete variable, and (2) the four levels of measurement.

Continuous and Discrete Variables. Variables can be continuous or discrete. **Continuous variables** contain a large number of values or attributes that flow along a continuum. We can divide a continuous variable into many smaller increments; in mathematical theory, the number of increments is infinite. Examples of continuous variables include temperature, age, income, crime rate, and amount of schooling. For example, we can measure the amount of your schooling as the years of schooling you completed. We can subdivide this into the total number of hours you have spent in classroom instruction and out-of-class assignments or preparation. We could further refine this into the number of minutes you devoted to acquiring and processing information and knowledge in school or due to school assignments. We could further refine this into all of the seconds that your brain was engaged in specific cognitive activities as you were acquiring and processing information.

Discrete variables have a relatively fixed set of separate values or variable attributes. Instead of a smooth continuum of numerous values, discrete variables contain a limited number of distinct categories. Examples of discrete variables include gender (male or female), religion (Protestant, Catholic, Jew, Muslim, atheist), marital status (never married single, married, divorced or separated, widowed), or academic degrees (high school diploma, or community college associate, four-year college, master's or doctoral degrees). Whether a variable is continuous or discrete affects its level of measurement.

Four Levels of Measurement. Levels of measurement build on the difference between continuous and discrete variables. Higher level measures are continuous and lower level ones are discrete. The four levels of measurement categorize its precision.[14]

Deciding on the appropriate level of measurement for a construct is not always easy. It depends on two things: how we understand a construct (its definition and assumptions), and the type of indicator or measurement procedure.

The way we conceptualize a construct can limit how precisely we can measure it. For example, we might reconceptualize some of the variables listed earlier as continuous to be discrete. We can think of temperature as a continuous variable with thousands of refined distinctions (e.g., degrees and fractions of degrees). Alternatively, we can think of it more crudely as five discrete categories (e.g., very hot, hot, cool, cold, very cold). We can think of age as continuous (in years, months, days, hours, minutes, or seconds) or discrete categories (infancy, childhood, adolescence, young adulthood, middle age, old age).

While we can convert continuous variables into discrete ones, we cannot go the other way around, that is, convert discrete variables into continuous

Levels of measurement A system for organizing information in the measurement of variables into four levels, from nominal level to ratio level.

Continuous variables Variables that are measured on a continuum in which an infinite number of finer gradations between variable attributes are possible.

Discrete variables Variables in which the attributes can be measured with only a limited number of distinct, separate categories.

TABLE 2 Characteristics of the Four Levels of Measurements

LEVEL	DIFFERENT CATEGORIES	RANKED	DISTANCE BETWEEN CATEGORIES MEASURED	TRUE ZERO
Nominal	Yes			
Ordinal	Yes	Yes		
Interval	Yes	Yes	Yes	
Ratio	Yes	Yes	Yes	Yes

ones. For example, we cannot turn sex, religion, and marital status into continuous variables. We can, however, treat related constructs with slightly different definitions and assumptions as being continuous (e.g., amount of masculinity or femininity, degree of religiousness, commitment to a marital relationship). There is a practical reason to conceptualize and measure at higher levels of measurement: We can collapse higher levels of measurement to lower levels, but the reverse is not true.

Distinguishing among the Four Levels. The four levels from lowest to highest precision are nominal, ordinal, interval, and ratio. Each level provides a different type of information (see Table 2). **Nominal-level measurement** indicates that a difference exists among categories (e.g., religion: Protestant, Catholic, Jew, Muslim; racial heritage: African, Asian, Caucasian, Hispanic, other). **Ordinal-level measurement** indicates a difference *and* allows us to rank order the categories (e.g., letter grades: A, B, C, D, F; opinion measures: strongly agree, agree, disagree, strongly disagree). **Interval-level measurement** does everything the first two do *and* allows us to specify the amount of distance between categories (e.g., Fahrenheit or celsius temperature: 5°, 45°, 90°; IQ scores: 95, 110, 125). **Ratio-level measurement** does everything the other levels do, *and* it has a true zero. This feature makes it possible to state relationships in terms of proportion or ratios (e.g., money income: $10, $100, $500; years of formal schooling: 1, 10, 13). In most practical situations, the distinction between interval and ratio levels makes little difference.

One source of confusion is that we sometimes use arbitrary zeros in interval measures but the zeros are only to help keep score. For example, a rise in temperature from 30 to 60 degrees is not really a doubling of the temperature, although the numbers appear to double. Zero degrees in Fahrenheit or centigrade is not the absence of any heat but is just a placeholder to make counting easier. For example, water freezes at 32° on a Fahrenheit temperature scale, 0° on a celsius or centigrade scale, and 273° on a Kelvin scale. Water boils at 212°, 100°, or 373.15°, respectively. If there were a true zero, the actual relation among temperature numbers would be a ratio. For example, 25° to 50° Fahrenheit would be "twice as warm," but this is not true because a ratio relationship does not exist without a true zero. We can see this in the ratio of boiling to freezing water temperatures. The ratio is 6.625 times higher in Fahrenheit, 100 times in Celsius, and 1.366 times

Nominal-level measurement The lowest, least precise level of measurement for which there is a difference in type only among the categories of a variable.

Ordinal-level measurement A level of measurement that identifies a difference among categories of a variable and allows the categories to be rank ordered as well.

Interval-level measurement A level of measurement that identifies differences among variable attributes, ranks categories, and measures distance between categories but has no true zero.

Ratio-level measurement The highest, most precise level of measurement; variable attributes can be rank ordered, the distance between them precisely measured, and there is an absolute zero.

in Kelvin. The Kelvin scale has an absolute zero (the absence of all heat), and its ratio corresponds to physical conditions. While this physical world example may be familiar, another example of arbitrary—not true—zeros occurs when measuring attitudes with numbers. We may assign a value to statements in a survey questionnaire (e.g., –1 = disagree, 0 = no opinion, +1 = agree). Just because our data are in the form of numbers does not allow us to use statistical procedures that require the mathematical assumption of a true zero.

Discrete variables are nominal and ordinal, whereas we can measure continuous variables at the interval or ratio level. There is an interesting unidirectional relationship among the four levels. We can convert a ratio-level measure into the interval, ordinal, or nominal level; an interval level into an ordinal or nominal level; and an ordinal into a nominal level; but the process does not work in the opposite way! This happens because higher levels of measurement contain more refined information than lower levels. We can always toss out or ignore the refined information of a high-level measure, but we cannot squeeze additional refined information out of a low-level measure.

For ordinal measures, we generally want to have at least five ordinal categories and try to obtain many observations for each. This is so because a distortion occurs as we collapse a continuous construct into few ordered categories. We minimize the distortion as the number of ordinal categories and the number of observations increase.[15] (See Example Box 2, Example of Four Levels of Measurement).

Before continuing, keep two things in mind. First, we can measure nearly any social phenomenon. We can measure some constructs directly and create precise numerical values (e.g., family income) while other constructs are less precise and require the use of surrogates or proxies to indirectly measure a variable (e.g., predisposition to commit a crime). Second, we can learn a great deal from the measures created by other researchers. We are fortunate to have the work of other researchers to draw on. It is not always necessary to start from scratch. We can use a past scale or index or modify it for our own purposes. Measuring aspects of social life is an ongoing process. We are constantly creating ideas, refining theoretical definitions, and improving measures of old or new constructs.

EXAMPLE BOX 2

Example of Four Levels of Measurement

VARIABLE (LEVEL OF MEASUREMENT)	HOW VARIABLE IS MEASURED
Religion (nominal)	Different religious denominations (Jewish, Catholic, Lutheran, Baptist) are not ranked but are only different (unless one belief is conceptualized as closer to heaven).
Attendance (ordinal)	"How often do you attend religious services? (0) Never, (1) less than once a year, (3) several times a year, (4) about once a month, (5) two or three times a week, or (8) several times a week." This might have been measured at a ratio level if the exact number of times a person attended were asked instead.
IQ score (interval)	Most intelligence tests are organized with 100 as average, middle, or normal. Scores higher or lower indicate distance from the average. Someone with a score of 115 has somewhat above average measured intelligence for people who took the test, whereas 90 is slightly below. Scores of below 65 or above 140 are rare.
Age (ratio)	Age is measured by years. There is a true zero (birth). Note that a 40-year-old has lived twice as long as a 20-year-old.

Principles of Good Measurement. Three features of good measurement whether we are considering using a single-indicator or a scale or index (discussed next) to measure a variable are that (1) the attributes or categories of a variable should be mutually exclusive, (2) they should also be exhaustive, and (3) the measurement should be unidimensional.

1. Mutually exclusive attributes means that an individual or a case will go into one and only one variable category. For example, we wish to measure the variable type of religion using the four attributes Christian, non-Christian, Jewish, and Muslim. Our measure is not mutually exclusive. Both Islam and Judaism are non-Christian religious faiths. A Jewish person and a Muslim fit into two categories: (1) the non-Christian and (2) Jewish or Muslim. Another example without mutually exclusive attributes is to measure the type of city using the three categories of river port city, state capital, and access to an international airport. A city could be all three (a river port state capital with an international airport), any combination of the three, or none of the three. To have mutually exclusive attitudes, we must create categories so that cases cannot be placed into more than one category.

2. Exhaustive attribute means that every case has a place to go or fits into at least one of a variable's categories. Returning to the example of the variable religion, with the four categorical attributes of Christian, non-Christian, Jewish, and Muslim, say we drop the non-Christian category to make the attributes mutually exclusive: Christian, Jewish, or Muslim. These are not exclusive attributes. The Buddhist, Hindu, atheist, and agnostic do not fit anywhere. We must create attributes to cover every possible situation. For example, Christian, Jewish, Muslim, or Other attributes for religion would be exclusive and mutually exclusive.

3. Unidimensionality means that a measure fits together or measures one single, coherent construct. Unidimensionality was hinted at in the previous discussions of construct and content validity. Unidimensionality states that if we combine several specific pieces of information into a single score or measure, all of the pieces should measure the same thing. We sometimes use a more advanced technique—factor analysis —to test for the unidimensionality of data.

We may see an apparent contradiction between the idea of using multiple indicators or a scale or index (see next section) to capture diverse parts of a complex construct and the criteria of unidimensionality. The contraction is apparent only because constructs vary theoretically by level of abstraction. We may define a complex, abstract construct using multiple subdimensions, each being a part of the complex construct's overall content. In contrast, simple, low-level constructs that are concrete typically have just one dimension. For example, "feminist ideology" is a highly abstract and complex construct. It includes specific beliefs and attitudes toward social, economic, political, family, and sexual relations. The ideology's belief areas are parts of the single, more abstract and general construct. The parts fit together as a whole. They are mutually reinforcing and collectively form one set of beliefs about the dignity, strength, and power of women. To create a unidimensional measure of feminist ideology requires us to conceptualize it as a unified belief system that might vary from very antifeminist to very profeminist. We can test the convergence validity of our measure with multiple indicators that tap the construct's subparts. If one belief area (e.g., sexual relations) is consistently distinct from all other areas in empirical tests, then we question its unidimensionality.

It is easy to become confused about unidimensionality because an indicator we use for a simple

Mutually exclusive attribute The principle that variable attributes or categories in a measure are organized so that responses fit into only one category and there is no overlap.

Exhaustive attributes The principle that attributes or categories in a measure should provide a category for all possible responses.

Unidimensionality The principle that when using multiple indicators to measure a construct, all indicators should consistently fit together and indicate a single construct.

construct in one situation might indicate one part of a different, complex construct in another situation. We can combine multiple simple, concrete constructs into a complex, more abstract construct. The principle of unidimensionality in measurement means that for us to measure a construct, we must conceptualize it as one coherent, integrated core idea *for its level of abstraction*. This shows the way that the processes of conceptualization and measurement are tightly interwoven.

Here is a specific example. A person's attitude about gender equality with regard to getting equal pay for work is a simpler, more specific and less abstract idea than gender ideology (i.e., a general set of beliefs about gender relations in all areas of life). We might measure attitude regarding equal pay as a unidimensional construct in its own or as a less abstract subpart of the complex, broader construct of gender ideology. This does not mean that gender ideology ceases to be unidimensional. It is a complex idea with several parts but can be unidimensional at a more abstract level.

SCALES AND INDEXES

In this section, we look at scales and indexes, specialized measures from among the hundreds created by researchers.[16] We have scales and indexes to measure many things: the degree of formalization in bureaucratic organizations, the prestige of occupations, the adjustment of people to a marriage, the intensity of group interaction, the level of social activity in a community, the degree to which a state's sexual assault laws reflect feminist values, and the level of socioeconomic development of a nation. We will examine principles of measurement, consider principles of index and scale construction, and then explore a few major types of index and scale.

You might find the terms *index* and *scale* confusing because people use them interchangeably. One researcher's scale is another's index. Both produce ordinal- or interval-level measures. To add to the confusion, we can combine scale and index techniques into a single measure. Nonetheless, scales and indexes are very valuable. They give us more information about a variable and expand the quality of measurement (i.e., increase reliability and validity) over using a simple, single indictor measure. Scales and indexes also aid in data reduction by condensing and simplifying information (see Expansion Box 3, Scales and Indexes: Are They Different?).

Index Construction

You hear about indexes all the time. For example, U.S. newspapers report the Federal Bureau of Investigation (FBI) crime index and the consumer price index (CPI). The FBI index is the sum of police reports on seven so-called index crimes (criminal homicide, aggravated assault, forcible rape, robbery, burglary, larceny of $50 or more, and auto theft). The index began as part of the Uniform Crime Report in 1930 (see Rosen, 1995). The CPI, which is a measure of inflation, is created by totaling the cost of buying a list of goods and services (e.g., food, rent, and utilities) and comparing the

EXPANSION BOX 3

Scales and Indexes: Are They Different?

For most purposes, researchers can treat scales and indexes as being interchangeable. Social researchers do not use a consistent nomenclature to distinguish between them.

A *scale* is a measure in which a researcher captures the intensity, direction, level, or potency of a variable construct and arranges responses or observations on a continuum. A scale can use a single indicator or multiple indicators. Most are at the ordinal level of measurement.

An *index* is a measure in which a researcher adds or combines several distinct indicators of a construct into a single score. This composite score is often a simple sum of the multiple indicators. It is used for content and convergent validity. Indexes are often measured at the interval or ratio level.

Researchers sometimes combine the features of scales and indexes in a single measure. This is common when a researcher has several indicators that are scales (i.e., that measure intensity or direction). He or she then adds these indicators together to yield a single score, thereby creating an index.

total to the cost of buying the same list in the previous period. The CPI has been used by the U.S. Bureau of Labor Statistics since 1919; wage increases, union contracts, and social security payments are based on it. An **index** is a combination of items into a single numerical score. Various components or subparts of a construct are each measured and then combined into one measure.

There are many types of indexes. For example, the total number of questions correct on an exam with 25 questions is a type of index. It is a composite measure in which each question measures a small piece of knowledge and all questions scored correct or incorrect are totaled to produce a single measure. Indexes measure the most desirable place to live (based on unemployment, commuting time, crime rate, recreation opportunities, weather, and so on), the degree of crime (based on combining the occurrence of different specific crimes), the mental health of a person (based on the person's adjustment in various areas of life), and the like.

Creating indexes is so easy that we must be careful to check that every item in an index has face validity and excludes any without face validity. We want to measure each part of the construct with at least one indicator. Of course, it is better to measure the parts of a construct with multiple indicators.

An example of an index is a college quality index (see Example Box 3, Example of Index). A theoretical definition says that a high-quality college has six distinguishing characteristics: (1) few students per faculty member, (2) a highly educated faculty, (3) high number of books in the library, (4) few students dropping out of college, (5) many students who go on to seek advanced degrees, and (6) faculty members who publish books or scholarly articles. We score 100 colleges on each item and then add the scores for each to create an index score of college quality that can be used to compare colleges.

We can combine indexes. For example, to strengthen my college quality index, I add a subindex on teaching quality. The index contains eight items: (1) average size of classes, (2) percentage of class time devoted to discussion, (3) number of different classes each faculty member teaches, (4) availability of faculty to students outside the classroom, (5) currency and amount of reading assigned, (6) degree to which assignments promote learning, (7) degree to which faculty get to know each student, and (8) student ratings of instruction. Similar subindex measures can be created for other parts of the college quality index. They can be combined into a more global measure of college quality. This further elaborates the definition of the construct "quality of college."

Next we look at three issues involved when we construct an index: weight of items, missing data, and the use of rates and standardization.

1. *Weighting* is an important issue in index construction. Unless otherwise stated, we assume that the items in an index are unweighted. Likewise, unless we have a good theoretical reason for assigning different weights to items, we use equal weights. An *unweighted index* gives each item equal weight. We simply sum the items without modification, as if each were multiplied by 1 (or –1 for items that are negative). A *weighted index* values or weights some items more than others. The size of weights can come from theoretical assumptions, the theoretical definition, or a statistical technique such as factor analysis.

For example, we can elaborate the theoretical definition of the college quality index. We decide that the student/faculty ratio and number of faculty with Ph.D.s are twice as important as the number of books in the library per student or the percentage of students pursuing advanced degrees. Also, the percentage of freshmen who drop out and the number of publications per faculty member are three times more important than books in the library or percentage of students pursuing an advanced degree. This is easier to see when it is expressed as a formula (refer to Example Box 3).

The number of students per faculty member and the percentage who drop out have negative signs because, as they increase, the quality of the college declines. The weighted and unweighted indexes can

Index The summing or combining of many separate measures of a construct or variable to create a single score.

EXAMPLE BOX 3

Example of Index

In symbolic form, where:

Q = overall college quality

A quality-of-college index is based on the following six items:

R = number of students per faculty member
F = percentage of faculty with Ph.D.s
B = number of books in library per student
D = percentage of freshmen who drop out or do not finish
A = percentage of graduates who seek an advanced degree
P = number of publications per faculty member

Unweighted formula: $(-1)\,R + (1)\,F + (1)\,B + (-1)\,D + (1)\,A + (1)\,P = Q$
Weighted formula: $(-2)\,R + (2)\,F + (1)\,B + (-3)\,D + (1)\,A + (3)\,P = Q$

Old Ivy College

Unweighted: $(-1)\,13 + (1)\,80 + (1)\,334 + (-1)\,14 + (1)\,28 + (1)\,4 = 419$
Weighted: $(-2)\,13 + (2)\,80 + (1)\,334 + (-3)\,14 + (1)\,28 + (3)\,4 = 466$

Local College

Unweighted: $(-1)\,20 + (1)\,82 + (1)\,365 + (-1)\,25 + (1)\,15 + (1)\,2 = 419$
Weighted: $(-2)\,20 + (2)\,82 + (1)\,365 + (-3)\,25 + (1)\,15 + (3)\,2 = 435$

Big University

Unweighted: $(-1)\,38 + (1)\,95 + (1)\,380 + (-1)\,48 + (1)\,24 + (1)\,6 = 419$
Weighted: $(-2)\,38 + (2)\,95 + (1)\,380 + (-3)\,48 + (1)\,24 + (3)\,6 = 392$

produce different results. Consider Old Ivy College, Local College, and Big University. All have identical unweighted index scores, but the colleges have different quality scores after weighting.

Weighting produces different index scores in this example, but in most cases, weighted and unweighted indexes yield similar results. Researchers are concerned with the relationship between variables, and weighted and unweighted indexes usually give similar results for the relationships between variables.[17]

2. *Missing data* can be a serious problem when constructing an index. Validity and reliability are threatened whenever data for some cases are missing. There are four ways to attempt to resolve the problem (see Expansion Box 4, Ways to Deal with Missing Data), but none fully solves it.

For example, I construct an index of the degree of societal development in 1985 for 50 nations. The index contains four items: life expectancy, percentage of homes with indoor plumbing, percentage of population that is literate, and number of telephones per 100 people. I locate a source of United Nations statistics for my information. The values for Belgium are 68 + 87 + 97 + 28 and for Turkey are 55 + 36 + 49 + 3; for Finland, however, I discover that literacy data are unavailable. I check other sources of information, but none has the data because they were not collected.

3. *Rates and standardization* are related ideas. You have heard of crime rates, rates of population growth, or the unemployment rate. Some indexes and single-indicator measures are expressed as rates. Rates involve standardizing the value of an item to make comparisons possible. The items in an

EXPANSION BOX 4

Ways to Deal with Missing Data

1. *Eliminate all cases for which any information is missing.* If one nation in the discussion is removed from the study, the index will be reliable for the nations on which information is available. This is a problem if other nations have missing information. A study of 50 nations may become a study of 20 nations. Also, the cases with missing information may be similar in some respect (e.g., all are in eastern Europe or in the Third World), which limits the generalizability of findings.
2. *Substitute the average score for cases in which data are present.* The average literacy score from the other nations is substituted. This "solution" keeps Finland in the study but gives it an incorrect value. For an index with few items or for a case that is not "average," this creates serious validity problems.
3. *Insert data based on nonquantitative information about the case.* Other information about Finland (e.g., percentage of 13- to 18-year-olds in high school) is used to make an informed guess about the literacy rate. This "solution" is marginally acceptable in this situation. It is not as good as measuring Finland's literacy, and it relies on an untested assumption—that one can predict the literacy rate from other countries' high school attendance rate.
4. *Insert a random value.* This is unwise for the development index example. It might be acceptable if the index had a very large number of items and the number of cases was very large. If that were the situation, however, then eliminating the case is probably a better "solution" that produces a more reliable measure.

Source: Allison (2001).

index frequently need to be standardized before they can be combined.

Standardization involves selecting a base and dividing a raw measure by the base. For example, City A had ten murders and City B had thirty murders in the same year. In order to compare murders in the two cities, we will need to standardize the raw number of murders by the city population. If the cities are the same size, City B is more dangerous. But City B may be safer if it is much larger. For example, if City A has 100,000 people and City B has 600,000, then the murder rate per 100,000 is ten for City A and five for City B.

Standardization makes it possible for us to compare different units on a common base. The process of standardization, also called *norming*, removes the effect of relevant but different characteristics in order to make the important differences visible. For example, there are two classes of students. An art class has twelve smokers and a biology class has twenty-two smokers. We can compare the rate or incidence of smokers by standardizing the number of smokers by the size of the classes. The art class has 32 students and the biology class has 143 students. One method of standardization that you already know is the use of percentages, whereby measures are standardized to a common base of 100. In terms of percentages, it is easy to see that the art class has more than twice the rate of smokers (37.5 percent) than the biology class (15.4 percent).

A critical question in standardization is deciding what base to use. In the examples given, how did I know to use city size or class size as the base? The choice is not always obvious; it depends on the theoretical definition of a construct. Different bases can produce different rates. For example, the unemployment rate can be defined as the number of people in the workforce who are out of work. The overall unemployment rate is

$$\text{unemployment rate} = \frac{\text{number of unemployed people}}{\text{total number of people working}}$$

We can divide the total population into subgroups to get rates for subgroups in the population such as

Standardization Procedures to adjust measures statistically to permit making an honest comparison by giving a common basis to measures of different units.

White males, African American females, African American males between the ages of 18 and 28, or people with college degrees. Rates for these subgroups may be more relevant to the theoretical definition or research problem. For example, we believe that unemployment is an experience that affects an entire household or family and that the base should be households, not individuals. The rate will look like this:

$$\text{unemployment rate} = \frac{\text{number of households with one unemployed person}}{\text{total number of households}}$$

Different conceptualizations suggest different bases and different ways to standardize. When combining several items into an index, it is best to standardize items on a common base (see Example Box 4, Standardization and the Real Winners at the 2000 Olympics).

Scales

We often use scales when we want to measure how an individual feels or thinks about something. Some call this the *hardness or potency of feelings*. Scales also help in the conceptualization and operationalization processes. For example, you believe a single ideological dimension underlies people's judgments about specific policies (e.g., housing, education, foreign affairs). Scaling can help you determine whether a single construct—for instance, "conservative/liberal ideology"—underlies the positions that people take on specific policies.

Scale A class of quantitative data measures often used in survey research that captures the intensity, direction, level, or potency of a variable construct along a continuum; most are at the ordinal level of measurement.

Likert scale A scale often used in survey research in which people express attitudes or other responses in terms of ordinal-level categories (e.g., agree, disagree) that are ranked along a continuum.

Scaling measures the intensity, direction, level, or potency of a variable. Graphic rating **scales** are an elementary form of scaling. People indicate a rating by checking a point on a line that runs from one extreme to another. This type of scale is easy to construct and use. It conveys the idea of a continuum, and assigning numbers helps people think about quantities. Scales assume that people with the same subjective feeling mark the graphic scale at the same place. Figure 7 is an example of a "feeling thermometer" scale that is used to find out how people feel about various groups in society (e.g., the National Organization of Women, the Ku Klux Klan, labor unions, physicians). Political scientists have used this type of measure in the national election study since 1964 to measure attitudes toward candidates, social groups, and issues.[18]

We next look at five commonly used social science scales: Likert, Thurstone, Borgadus social distance, semantic differential, and Guttman scale. Each illustrates a somewhat different logic of scaling.

1. *Likert scaling.* You have probably used **Likert scales**; they are widely used in survey research. They were developed in the 1930s by Rensis Likert to provide an ordinal-level measure of a person's attitude.[19] Likert scales are called *summated-rating* or *additive scales* because a person's score on the scale is computed by summing the number of responses he or she gives. Likert scales usually ask people to indicate whether they agree or

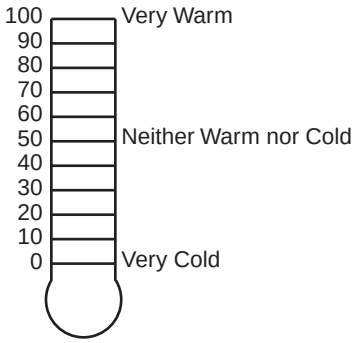


FIGURE 7 "Feeling Thermometer" Graphic Rating Scale

EXAMPLE BOX 4

Standardization and the Real Winners at the 2000 Olympics

Sports fans in the United States were jubilant about "winning" at the 2000 Olympics by carrying off the most gold medals. However, because they failed to *standardize,* the "win" is an illusion. Of course, the world's richest nation with the third largest population does well in one-on-one competition among all nations. To see what really happened, one must standardize on a base of the population or wealth. Standardization yields a more accurate picture by adjusting the results as if the nations had equal populations and wealth. The results show that the Bahamas, with fewer than 300,000 citizens (smaller than a medium-sized U.S. city), proportionately won the most gold. Adjusted for its population size or wealth, the United States is not even near the top; it appears to be the leader only because of its great size and wealth. Sports fans in the United States can perpetuate the illusion of being at the top only if they ignore the comparative advantage of the United States.

TOP TEN GOLD MEDAL WINNING COUNTRIES AT THE 2000 OLYMPICS IN SYDNEY

Unstandardized Rank			*Standardized Rank**			
RANK	**COUNTRY**	**TOTAL**	**COUNTRY**	**TOTAL**	**POPULATION**	**GDP**
1	USA	39	Bahamas	1.4	33.3	20.0
2	Russia	32	Slovenia	2	10	10.0
3	China	28	Cuba	11	9.9	50.0
4	Australia	16	Norway	4	9.1	2.6
5	Germany	14	Australia	16	8.6	4.1
6	France	13	Hungry	8	7.9	16.7
7	Italy	13	Netherlands	12	7.6	3.0
8	Netherlands	12	Estonia	1	7.1	20.0
9	Cuba	11	Bulgaria	5	6.0	41.7
10	Britain	11	Lithuania	2	5.4	18.2
	EU15**	80	EU15	80	2.1	0.9
			USA	39	1.4	0.4

*Population is gold medals per 10 million people and GDP is gold medals per $10 billion.

**EU15 is the 15 nations of the European Union treated as a single unit.

Source: Adapted from *The Economist,* October 7, 2000, p. 52. Copyright 2000 by Economist Newspaper Group. Reproduced with permission of Economist Newspaper Group in the format Textbook via Copyright Clearance Center.

disagree with a statement. Other modifications are possible; people might be asked whether they approve or disapprove or whether they believe something is "almost always true" (see Example Box 5, Examples of Types of Likert Scales).

To create a Likert scale, you need a minimum of two categories, such as "agree" and "disagree." Using only two choices creates a crude measure and forces distinctions into only two categories. It is usually better to use four to eight categories. You can combine or collapse categories after the data have been collected, but once you collect them using crude categories, you cannot make them more precise later. You can increase the number of categories at the end of a scale by adding "strongly agree," "somewhat agree," "very strongly agree," and so forth. You want to keep the number of choices to eight or nine at most. More distinctions than that are not meaningful, and people will become confused. The choices should be evenly

EXAMPLE BOX 5

Examples of Types of Likert Scales

THE ROSENBERG SELF-ESTEEM SCALE

All in all, I am inclined to feel that I am a failure:

(1) Almost always true
(2) Often true
(3) Sometimes true
(4) Seldom true
(5) Never true

A STUDENT EVALUATION OF INSTRUCTION SCALE

Overall, I rate the quality of instruction in this course as:

Excellent Good Average Fair Poor

A MARKET RESEARCH MOUTHWASH RATING SCALE

Brand	*Dislike Completely*	*Dislike Somewhat*	*Dislike a Little*	*Like a Little*	*Like Somewhat*	*Like Completely*
X	_____	_____	_____	_____	_____	_____
Y	_____	_____	_____	_____	_____	_____

WORK GROUP SUPERVISOR SCALE

My supervisor:

	Never	*Seldom*	*Sometimes*	*Often*	*Always*
Lets members know what is expected of them	1	2	3	4	5
Is friendly and approachable	1	2	3	4	5
Treats all unit members as equals	1	2	3	4	5

balanced (e.g., "strongly agree," "agree," "strongly disagree," "disagree"). Nunnally (1978:521) stated:

> *As the number of scale steps is increased from 2 up through 20, the increase in reliability is very rapid at first. It tends to level off at about 7, and after about 11 steps, there is little gain in reliability from increasing the number of steps.*

Researchers have debated about whether to offer a neutral category (e.g., "don't know," "undecided," "no opinion") in addition to the directional categories (e.g., "disagree," "agree"). A neutral category implies an odd number of categories.

Response set A tendency to agree with every question in a series rather than carefully thinking through one's answer to each.

We can combine several Likert scale items into a composite index if they all measure the same construct. Consider the Index of Equal Opportunity for Women and the Self-Esteem Index created by Sniderman and Hagen (1985) (see Example Box 6, Examples of Using the Likert Scale to Create Indexes). In the middle of large surveys, they asked respondents three questions about the position of women. The researchers later scored answers and combined items into an index that ranged from 3 to 15. Respondents also answered questions about self-esteem. Notice that when scoring these items, they scored one item (question 2) in reverse. The reason for switching directions in this way is to avoid the problem of the **response set**. The response

EXAMPLE BOX 6

Examples of Using the Likert Scale to Create Indexes

Sniderman and Hagen (1985) created indexes to measure beliefs about equal opportunity for women and self-esteem. For both indexes, scores were added to create an unweighted index.

INDEX OF EQUAL OPPORTUNITY FOR WOMEN

Questions

1. Women have less opportunity than men to get the education they need to be hired in top jobs.

Strongly Agree	Somewhat Agree	Somewhat Disagree	Disagree a Great Deal	Don't Know

2. Many qualified women cannot get good jobs; men with the same skills have less trouble.

Strongly Agree	Somewhat Agree	Somewhat Disagree	Disagree a Great Deal	Don't Know

3. Our society discriminates against women.

Strongly Agree	Somewhat Agree	Somewhat Disagree	Disagree a Great Deal	Don't Know

Scoring: For all items, Strongly Agree = 1, Somewhat Agree = 2, Somewhat Disagree = 4, Disagree a Great Deal = 5, Don't Know = 3.

Highest Possible Index Score = 15, respondent feels opportunities for women are equal
Lowest Possible Index Score = 3, respondent feels opportunities are not equal

SELF-ESTEEM INDEX

Questions

1. On the whole, I am satisfied with myself.	Agree	Disagree	Don't Know
2. At times, I think I am no good at all.	Agree	Disagree	Don't Know
3. I sometimes feel that (other) men do not take my opinion seriously.	Agree	Disagree	Don't Know

Scoring: Items 1 and 3: 1 = Disagree, 2 = Don't Know, 3 = Agree, Item 2: 1 = Disagree, 2 = Don't Know, 1 = Agree.

Highest Possible Index Score = 9, high self-esteem
Lowest Possible Index Score = 3, low self-esteem

set, also called *response style* and *response bias*, is the tendency of some people to answer a large number of items in the same way (usually agreeing) out of laziness or a psychological predisposition. For example, if items are worded so that saying "strongly agree" always indicates self-esteem, we would not know whether a person who always strongly agreed had high self-esteem or simply had a tendency to agree with questions. The person might be answering "strongly agree" out of habit or a tendency to agree. We word statements in alternative directions so that anyone who agrees all the

time appears to answer inconsistently or to have a contradictory opinion.

We often combine many Likert-scaled attitude indicators into an index. Scale and indexes can improve reliability and validity. An index uses multiple indicators, which improves reliability. The use of multiple indicators that measure several aspects of a construct or opinion improves content validity. Finally, the index scores give a more precise quantitative measure of a person's opinion. For example, we can measure a person's opinion with a number from 10 to 40 instead of in four categories: "strongly agree," "agree," "disagree," and "strongly disagree."

Instead of scoring Likert items, as in the previous example, we could use the scores –2, –1, +1, +2. This scoring has an advantage in that a zero implies neutrality or complete ambiguity whereas a high negative number means an attitude that opposes the opinion represented by a high positive number.

The numbers we assign to the response categories are arbitrary. Remember that the use of a zero does not give the scale or index a ratio level of measurement. Likert scale measures are at the ordinal level of measurement because responses indicate only a ranking. Instead of 1 to 4 or –2 to +2, the numbers 100, 70, 50, and 5 would have worked. Also, we should not be fooled into thinking that the distances between the ordinal categories are intervals just because numbers are assigned. The numbers are used for convenience only. The fundamental measurement is only ordinal.[20]

The real strength of the Likert Scale is its simplicity and ease of use. When we combine several ranked items, we get a more comprehensive multiple indicator measurement. The scale has two limitations: Different combinations of several scale items produce the same overall score, and the response set is a potential danger.

Thurstone scaling Measuring in which the researcher gives a group of judges many items and asks them to sort the items into categories along a continuum and then considers the sorting results to select items on which the judges agree.

2. *Thurstone scaling.* This scale is for situations when we are interested in something with many ordinal aspects but would like a measure that combines all information into a single interval-level continuum. For example, a dry cleaning business, Quick and Clean, contacts us; the company wants to identify its image in Greentown compared to that of its major competitor, Friendly Cleaners. We conceptualize a person's attitude toward the business as having four aspects: attitude toward location, hours, service, and cost. We learn that people see Quick and Clean as having more convenient hours and locations but higher costs and discourteous service. People see Friendly Cleaners as having low cost and friendly service but inconvenient hours and locations. Unless we know how the four aspects relate to the core attitude—image of the dry cleaner—we cannot say which business is generally viewed more favorably. During the late 1920s, Louis Thurstone developed scaling methods for assigning numerical values in such situations. These are now called **Thurstone scaling** or the *method of equal-appearing intervals.*[21]

Thurstone scaling uses the law of comparative judgment to address the issue of comparing ordinal attitudes when each person makes a unique judgment. The law anchors or fixes the position of one person's attitude relative to that of others as each makes an individual judgment. The law of comparative judgment states that we can identify the "most common response" for each object or concept being judged. Although different people arrive at different judgments, the individual judgments cluster around a single most common response. The dispersion of individual judgments around the common response follows a statistical pattern called the *normal distribution*. According to the law, if many people agree that two objects differ, then the most common responses for the two objects will be distant from each other. By contrast, if many people are confused or disagree, the common responses of the two objects will be closer to each other.

With Thurstone scaling, we develop many statements (e.g., more than 100) regarding the object of interest and then use judges to reduce the number to a smaller set (e.g., 20) by eliminating ambiguous

statements. Each judge rates the statements on an underlying continuum (e.g., favorable to unfavorable). We examine the ratings and keep some statements based on two factors: (1) agreement among the judges and (2) the statement's location on a range of possible values. The final set of statements is a measurement scale that spans a range of values.

Thurstone scaling begins with a large number of statements that cover all shades of opinion. Each statement should be clear and precise. "Good" statements refer to the present and are not capable of being interpreted as facts. They are unlikely to be endorsed by everyone, are stated as simple sentences, and avoid words such as *always* and *never.* We can get ideas for writing the statements from reviewing the literature, from the mass media, from personal experience, and from asking others. For example, statements about the dry cleaning business might include the four aspects listed before plus the following:

- I think X Cleaners dry cleans clothing in a prompt and timely manner.
- In my opinion, X Cleaners keeps its stores looking neat and attractive.
- I do not think that X Cleaners does a good job of removing stains.
- I believe that X Cleaners charges reasonable prices for cleaning coats.
- I believe that X Cleaners returns clothing clean and neatly pressed.
- I think that X Cleaners has poor delivery service.

We would next locate 50 to 300 judges who should be familiar with the object or concept in the statements. Each judge receives a set of statement cards and instructions. Each card has one statement on it, and the judges place each card in one of several piles. The number of piles is usually 7, 9, 11, or 13. The piles represent a range of values (e.g., favorable to neutral to unfavorable) with regard to the object or concept being evaluated. Each judge places cards in rating piles independently of the other judges.

After the judges place all cards in piles, we create a chart cross-classifying the piles and the statements. For example, 100 statements and 11 piles results in an 11×100 chart, or a chart with $11 \times 100 = 1{,}100$ boxes. The number of judges who assigned a rating to a given statement is written into each box. Statistical measures (beyond the present discussion) are used to compute the average rating of each statement and the degree to which the judges agree or disagree. We keep the statements with the highest between-judge agreement, or interrater reliability, as well as statements that represent the entire range of values. (See Example Box 7, Example of Thurstone Scaling.)

With Thurstone scaling, we can construct an attitude scale or select statements from a larger collection of attitude statements. The method has four limitations:

- It measures agreement or disagreement with statements but not the intensity of agreement or disagreement.
- It assumes that judges and others agree on where statements appear in a rating system.
- It is time consuming and costly.
- It is possible to get the same overall score in several ways because agreement or disagreement with different combinations of statements can produce the same average.

3. *Bogardus social distance scale.* A measure of the "social distance" that separates social groups from each other is the **Bogardus social distance scale**. We use it with one group to learn how much distance its members feel toward a target or "outgroup." Emory Bogardus developed this technique in the 1920s to measure the willingness of members of different ethnic groups to associate with each other. Since then it has been used to see how close or distant people in one group feel toward some other group (e.g., a religious minority or a deviant group).[22]

Bogardus social distance scale A scale measuring the social distance between two or more social groups by having members of one group indicate the limit of their comfort with various types of social interaction or closeness with members of the other group(s).

EXAMPLE BOX 7

Example of Thurstone Scaling

Variable Measured: Opinion with regard to the death penalty.

Step 1: Develop 120 statements about the death penalty using personal experience, the popular and professional literature, and statements by others.

Example Statements

1. I think that the death penalty is cruel and unnecessary punishment.
2. Without the death penalty, there would be many more violent crimes.
3. I believe that the death penalty should be used only for a few extremely violent crimes.
4. I do not think that anyone was ever prevented from committing a murder because of fear of the death penalty.
5. I do not think that people should be exempt from the death penalty if they committed a murder even if they are insane.
6. I believe that the Bible justifies the use of the death penalty.
7. The death penalty itself is not the problem for me, but I believe that electrocuting people is a cruel way to put them to death.

Step 2: Place each statement on a separate card or sheet of paper and make 100 sets of the 120 statements.

Step 3: Locate 100 persons who agree to serve as judges. Give each judge a set of the statements and instructions to place them in one of 11 piles, from 1 = highly unfavorable statement through 11 = highly favorable statement.

Step 4: The judges place each statement into one of the 11 piles (e.g., Judge 1 puts statement 1 into pile 2; Judge 2 puts the same statement into pile 1; Judge 3 also puts it into pile 2, Judge 4 puts it in pile 3, and so on).

Step 5: Collect piles from judges and create a chart summarizing their responses. See the example chart that follows.

NUMBER OF JUDGES RATING EACH STATEMENT RATING PILE

	Unfavorable					*Neutral*					*Favorable*	
Statement	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*Total*
1	23	60	12	5	0	0	0	0	0	0	0	100
2	0	0	0	0	2	12	18	41	19	8	0	100
3	2	8	7	13	31	19	12	6	2	0	0	100
4	9	11	62	10	4	4	0	0	0	0	0	100

Step 6: Compute the average rating and degree of agreement by judges. For example, the average for question 1 is about 2, so there is high agreement; the average for question 3 is closer to 5, and there is much less agreement.

Step 7: Choose the final 20 statements to include in the death penalty opinion scale. Choose statements if the judges showed agreement (most placed an item in the same or a nearby pile) and ones that reflect the entire range of opinion, from favorable to neutral to unfavorable.

Step 8: Prepare a 20-statement questionnaire, and ask people in a study whether they agree or disagree with the statements.

The scale has a simple logic. We ask people to respond to a series of ordered statements. We place more socially intimate or close situations at one end and the least socially threatening situations at the opposite end. The scale's logic assumes that a person who is uncomfortable with another social group and might accept a few nonthreatening (socially distant) situations will express discomfort or refusal regarding the more threatening (socially intimate) situations.

We can use the scale in several ways. For example, we give people a series of statements: People from Group X are entering your country, are in your town, work at your place of employment, live in your neighborhood, become your personal friends, and marry your brother or sister. We ask people whether they feel comfortable with the situation in the statement or the contact is acceptable. We ask people to respond to all statements until they are at a situation with which they do not feel comfortable. No set number of statements is required; the number usually ranges from five to nine.

We can use the Bogardus scale to see how distant people feel from one outgroup versus another (see Example Box 8, Example of Bogardus Social Distance Scale). We can use the measure of social distance as either an independent or a dependent variable. For example, we might believe that social distance from a group is highest for people who have some other characteristic, such as education. Our hypothesis might be that White people's feelings of social distance toward Vietnamese people is negatively associated with education; that is, the least educated Whites feel the most social distance. In this situation, social distance is the dependent variable, and amount of education is the independent variable.

The social distance scale has two potential limitations. First, we must tailor the categories to a specific outgroup and social setting. Second, it is not easy for us to compare how a respondent feels toward several different groups unless the respondent completes a similar social distance scale for all outgroups at the same time. Of course, how a respondent completes the scale and the respondent's actual behavior in specific social situations may differ.

4. *Semantic differential.* Developed in the 1950s as an indirect measure of a person's feelings about a concept, object, or other person, **semantic differential** measures subjective feelings by using many adjectives because people usually communicate evaluations through adjectives. Most adjectives have polar opposites (e.g., *light/dark, hard/soft, slow/fast*). The semantic differential attempts to capture evaluations by relying on the connotations of adjectives. In this way, it measures a person's feelings and evaluations in an indirect manner.

To use the semantic differential, we offer research participants a list of paired opposite adjectives with a continuum of 7 to 11 points between them. We ask participants to mark the spot on the continuum between the adjectives that best expresses their evaluation or feelings. The adjectives can be very diverse and should be mixed (e.g., positive items should not be located mostly on either the right or the left side). Adjectives in English tend to fall into three major classes of meaning: evaluation *(good–bad),* potency *(strong–weak),* and activity *(active–passive).* Of the three classes, evaluation is usually the most significant.

The most difficult part of the semantic differential is analyzing the results. We need to use advanced statistical procedures to do so. Results from the procedures inform us as to how a person perceives different concepts or how people view a concept, object, or person. For example, political analysts might discover that young voters perceive their candidate to be traditional, weak, and slow, and midway between good and bad. Elderly voters perceive the candidate as leaning toward strong, fast, and good, and midway between traditional and modern. In Example Box 9, Example of Semantic Differential, a person rated two concepts. The pattern of responses for each concept illustrates how

Semantic differential A scale that indirectly measures feelings or thoughts by presenting people a topic or object and a list of polar opposite adjectives or adverbs and then having them indicate feelings by marking one of several spaces between the two adjectives or adverbs.

EXAMPLE BOX 8

Example of Bogardus Social Distance Scale

A researcher wants to find out how socially distant freshmen college students feel from exchange students from two different countries: Nigeria and Germany. She wants to see whether students feel more distant from students coming from Africa or from Europe. She uses the following series of questions in an interview:

> Please give me your first reaction, yes or no, whether you personally would feel comfortable having an exchange student from (name of country):
>
> ———— As a visitor to your college for a week
>
> ———— As a full-time student enrolled at your college
>
> ———— Taking several of the same classes you are taking
>
> ———— Sitting next to you in class and studying with you for exams
>
> ———— Living a few doors down the hall on the same floor in your dormitory
>
> ———— As a same-sex roommate sharing your dorm room
>
> ———— As someone of the opposite sex who has asked you to go out on a date

Hypothetical Results

	Percentage of Freshmen Who Report Feeling Comfortable	
	Nigeria	*Germany*
Visitor	100%	100%
Enrolled	98	100
Same class	95	98
Study together	82	88
Same dorm	71	83
Roommate	50	76
Go on date	42	64

The results suggest that freshmen feel more distant from Nigerian students than from German students. Almost all feel comfortable having the international students as visitors, enrolled in the college, and taking classes. Feelings of distance increase as interpersonal contact increases, especially if the contact involves personal living settings or activities not directly related to the classroom.

this individual feels. This person views the two concepts differently and appears to feel negatively about divorce.

Guttman scaling index A scale that researchers use after data are collected to reveal whether a hierarchical pattern exists among responses so that people who give responses at a "higher level" also tend to give "lower level" ones.

Statistical techniques can create three-dimensional diagrams of results.[23] The three aspects are diagrammed in a three-dimensional "semantic space." In the diagram, "good" is up and "bad" is down, "active" is left and "passive" is right, "strong" is away from the viewer and "weak" is close.

5. *Guttman scaling.* Also called *cumulative scaling*, the **Guttman scaling index** differs from the previous scales or indexes in that we use it to

EXAMPLE BOX 9

Example of Semantic Differential

Please read each pair of adjectives below and then place a mark on the blank space that comes closest to your first impression feeling. There are no right or wrong answers.

How do you feel about the idea of divorce?

Bad	___	x	___	___	___	___	___	___	___	Good
Deep	___	___	___	___	___	___	___	x	___	Shallow
Weak	___	___	x	___	___	___	___	___	___	Strong
Fair	___	___	___	___	___	___	___	x	___	Unfair
Quiet	___	___	___	___	___	___	___	___	x	Loud
Modern	___	___	___	___	___	___	___	___	___	Traditional
Simple	___	___	___	___	___	x	___	___	___	Complex
Fast	___	x	___	___	___	___	___	___	___	Slow
Dirty	___	x	___	___	___	___	___	___	___	Clean

How do you feel about the idea of marriage?

Bad	___	___	___	___	___	___	___	___	x	Good
Deep	___	x	___	___	___	___	___	___	___	Shallow
Weak	___	___	___	___	___	___	___	x	___	Strong
Fair	___	x	___	___	___	___	___	___	___	Unfair
Quiet	___	___	x	___	___	___	___	___	___	Loud
Modern	___	___	___	___	___	___	___	___	x	Traditional
Simple	___	___	___	___	___	x	___	___	___	Complex
Fast	___	___	___	___	___	___	___	x	___	Slow
Dirty	___	___	___	___	___	___	x	___	___	Clean

evaluate data after collecting them. This means that we must design a study with the Guttman scaling technique in mind. Louis Guttman developed the scale in the 1940s to determine whether there was a structured relationship among a set of indicators. He wanted to learn whether multiple indicators about an issue had an underlying single dimension or cumulative intensity.[24]

To use Guttman scaling, we begin by measuring a set of indicators or items. These can be questionnaire items, votes, or observed characteristics. We usually measure three to twenty indicators in a simple yes/no or present/absent fashion. We select items for which we believe there could be a logical relationship among all of them. We place the results into a Guttman scale chart and next determine whether there is a hierarchical pattern among items.

After we have the data, we can consider all possible combinations of responses. For example, we have three items: whether a child knows (1) her age, (2) her telephone number, and (3) three local elected political officials. The little girl could know her age but no other answer, or all three, or only her age and telephone number. Three items have eight possible combinations of answers or patterns of responses from not knowing any through knowing all three. There is a mathematical way to compute the number of combinations (e.g., twenty-three); you can write down all combinations of yes or no for three questions and see the eight possibilities.

An application of Guttman scaling known as *scalogram analysis* allows us to test whether a patterned hierarchical relationship exists in the data. We can divide response patterns into scaled items

and errors (or nonscalable). A scaled pattern for the child's knowledge example would be as follows: not knowing any item, knowing age only, knowing only age plus phone number, and knowing all three. All other combinations of answers (e.g., knowing the political leaders but not her age) are logically possible but nonscalable. If we find a hierarchical relationship, then most answers fit into the scalable patterns. The items are scalable, or capable of forming a Guttman scale, if a hierarchical pattern exists. For higher order items, a smaller number would agree but all would also agree to the lower order ones but not vice versa. In other words, higher order items build on the middle-level ones, and middle-level build on lower ones.

Statistical procedures indicate the degree to which items fit the expected hierarchical pattern. Such procedures produce a coefficient that ranges from zero to 100 percent. A score of zero indicates a random pattern without hierarchical structure; one of 100 percent indicates that all responses fit the hierarchical pattern. Alternative statistics to measure scalability have also been suggested.[25] (See Example Box 10, Guttman Scale Example.)

EXAMPLE BOX 10

Guttman Scale Example

Crozat (1998) examined public responses to various forms of political protest. He looked at survey data on the public's acceptance of forms of protest in Great Britain, Germany, Italy, the Netherlands, and the United States in 1974 and 1990. He found that the pattern of the public's acceptance formed a Guttman scale. Those who accepted more intense forms of protest (e.g., strikes and sit-ins) almost always accepted more modest forms (e.g., petitions or demonstrations), but not all who accepted modest forms accepted the more intense forms. In addition to showing the usefulness of the Guttman scale, Crozat also found that people in different nations saw protest similarly and the degree of Guttman scalability increased over time. Thus, the pattern of acceptance of protest activities was Guttman "scalable" in both time periods, but it more closely followed the Guttman pattern in 1990 than in 1974.

	FORM OF PROTEST				
	Petitions	*Demonstrations*	*Boycotts*	*Strikes*	*Sit-Ins*
Guttman Patterns					
	N	N	N	N	N
	Y	N	N	N	N
	Y	Y	N	N	N
	Y	Y	Y	N	N
	Y	Y	Y	Y	N
	Y	Y	Y	Y	Y
Other Patterns (examples only)					
	N	Y	N	Y	N
	Y	N	Y	Y	N
	Y	N	Y	N	N
	N	Y	Y	N	N
	Y	N	N	Y	Y

Clogg and Sawyer (1981) studied U.S. attitudes toward abortion using Guttman scaling. They examined the different conditions under which people thought abortion was acceptable (e.g., mother's health in danger, pregnancy resulting from rape). They discovered that 84.2 percent of responses fit into a scaled response pattern.

CONCLUSION

This chapter dicussed the principles and processes of measurement. Central to measurement is how we conceptualize—or refine and clarify ideas into conceptual definitions and operationalize conceptual variables into specific measures—or develop procedures that link conceptual definitions to empirical reality. How we approach these processes varies depending on whether a study is primarily qualitative or quantitative. In a quantitative study, we usually adopt a more deductive path, whereas with a qualitative study, the path is more inductive. Nonetheless, they share the same goal to establish an unambiguous connection between abstract ideas and empirical data.

The chapter also discussed the principles of reliability and validity. *Reliability* refers to a measure's dependability; *validity* refers to its truthfulness or the fit between a construct and data. In both quantitative and qualitative studies, we try to measure in a consistent way and seek a tight fit between the abstract ideas and the empirical social world. In addition, the principles of measurement are applied in quantitative studies to build indexes and scales. The chapter also discussed some major scales in use.

Beyond the core ideas of reliability and validity, we now know principles of sound measurement: Create clear definitions for concepts, use multiple indicators, and, as appropriate, weigh and standardize the data. These principles hold across all fields of study (e.g., family, criminology, inequality, race relations) and across the many research techniques (e.g., experiments, surveys).

As you are probably beginning to realize, a sound research project involves doing a good job in each phase of research. Serious mistakes or sloppiness in any one phase can do irreparable damage to the results, even if the other phases of the research project were conducted in a flawless manner.

KEY TERMS

bogardus social distance scale
casing
conceptual definition
conceptual hypothesis
conceptualization
concurrent validity
construct validity
content validity
continuous variables
convergent validity
criterion validity
discrete variables
discriminant validity
empirical hypothesis
equivalence reliability
exhaustive attributes
face validity
guttman scaling index
index
interval-level measurement
level of measurement
likert scale
measurement reliability
measurement validity
multiple indicators
mutually exclusive attributes
nominal-level measurement
operational definition
operationalization
ordinal-level measurement
predictive validity
ratio-level measurement
representative reliability
response set
rules of correspondence
scale
semantic differential
stability reliability
standardization
thurstone scaling
unidimensionality

REVIEW QUESTIONS

1. What are the three basic parts of measurement, and how do they fit together?
2. What is the difference between reliability and validity, and how do they complement each other?
3. What are ways to improve the reliability of a measure?
4. How do the levels of measurement differ from each other?
5. What are the differences between convergent, content, and concurrent validity? Can you have all three at once? Explain your answer.
6. Why are multiple indicators usually better than one indicator?
7. What is the difference between the logic of a scale and that of an index?
8. Why is unidimensionality an important characteristic of a scale?
9. What are advantages and disadvantages of weighting indexes?
10. How does standardization make comparison easier?

NOTES

1. Duncan (1984:220–239) presented cautions from a positivist approach on the issue of measuring anything.
2. The terms *concept*, *construct*, and *idea* are used more or less interchangeably, but their meanings have some differences. An *idea* is any mental image, belief, or impression. It refers to any vague impression, opinion, or thought. A *concept* is a thought, a general notion, or a generalized idea about a class of objects. A *construct* is a thought that is systematically put together, an orderly arrangement of ideas, facts, and impressions. The term *construct* is used here because its emphasis is on taking vague concepts and turning them into systematically organized ideas.
3. See Grinnell (1987:5–18) for further discussion.
4. See Blalock (1982:25–27) and Costner (1985) on the rules of correspondence or the auxiliary theories that connect an abstract concept with empirical indicators. Also see Zeller and Carmines (1980:5) for a diagram that illustrates the place of the rules in the measurement process. In his presidential address to the American Sociological Association in 1979, Hubert Blalock (1979a:882) said, "I believe that the most serious and important problems that require our immediate and concerted attention are those of conceptualization and measurement."
5. See Bailey (1984, 1986) for a discussion of the three levels.
6. See Bohrnstedt (1992a,b) and Carmines and Zeller (1979) for discussions of reliability and its various types.
7. See Sullivan and Feldman (1979) on multiple indicators. A more technical discussion can be found in Herting (1985), Herting and Costner (1985), and Scott (1968).
8. See Carmines and Zeller (1979:17). For a discussion of the many types of validity, see Brinberg and McGrath (1982).
9. The epistemic correlation is discussed in Costner (1985) and in Zeller and Carmines (1980:50–51, 137–139).
10. Kidder (1982) discussed the issue of disagreements over face validity, such as acceptance of a measure's meaning by the scientific community but not the subjects being studied.
11. This was adapted from Carmines and Zeller (1979:20–21).
12. For a discussion of types of criterion validity, see Carmines and Zeller (1979:17–19) and Fiske (1982) for construct validity.
13. See Cook and Campbell (1979) for elaboration.
14. See Borgatta and Bohrnstedt (1980) and Duncan (1984:119–155) for a discussion and critique of the topic of levels of measurement.
15. Johnson and Creech (1983) examined the measurement errors that occur when variables that are conceptualized as continuous are operationalized in a series of ordinal categories. They argued that errors are not serious if more than four categories and large samples are used.

16. For compilations of indexes and scales used in social research, see Brodsky and Smitherman (1983), Miller (1991), Robinson and colleagues (1972), Robinson and Shaver (1969), and Schuessler (1982).
17. For a discussion of weighted and unweighted index scores, see Nunnally (1978:534).
18. Feeling thermometers are discussed in Wilcox and associates (1989).
19. For more information on Likert scales, see Anderson and associates (1983:252–255), Converse (1987:72–75), McIver and Carmines (1981:22–38), and Spector (1992).
20. Some researchers treat Likert scales as interval-level measures, but there is disagreement on this issue. Statistically, whether the Likert scale has at least five response categories and an approximately even proportion of people answer in each category makes little difference.
21. McIver and Carmines (1981:16–21) have an excellent discussion of Thurstone scaling. Also see discussions in Anderson and colleagues (1983:248–252), Converse (1987:66–77), and Edwards (1957). The example used here is partially borrowed from Churchill (1983:249–254), who described the formula for scoring Thurstone scaling.
22. The social distance scale is described in Converse (1987:62–69). The most complete discussion can be found in Bogardus (1959).
23. The semantic differential is discussed in Nunnally (1978:535–543). Also see Heise (1965, 1970) on the analysis of scaled data.
24. See Guttman (1950).
25. See Bailey (1987:349–351) for a discussion of an improved method for determining scalability called *minimal marginal reproducibility*. Guttman scaling can involve more than yes/no choices and a large number of items, but the complexity increases quickly. A more elaborate discussion of Guttman scaling can be found in Anderson and associates (1983:256–260), Converse (1987:189–195), McIver and Carmines (1981:40–71), and Nunnally (1978:63–66). Clogg and Sawyer (1981) presented alternatives to Guttman scaling.

Qualitative and Quantitative Sampling

Qualitative and Quantitative Sampling

Reasons for Sampling
Sampling Strategies
Conclusion

Sampling is a major problem for any type of research. We can't study every case of whatever we're interested in, nor should we want to. Every scientific enterprise tries to find out something that will apply to everything *of a certain kind by studying* a few examples, *the results of the study being, as we say, "generalizable."*
—Howard Becker, *Tricks of the Trade*, p. 67

In *Promises I Can Keep*, an in-depth study of low-income mothers, Edin and Kefalas (2005) first identified eight low-income neighborhoods in the Philadelphia, Pennsylvania, area through extensive qualitative fieldwork and quantitative analysis of census data. Each neighborhood met three selection criteria: at least 20 percent of householders were below the poverty line, at least 20 percent of all households had a single parent, and each had a large number of Black, White, and Hispanic residents. In each neighborhood, Edin and Kefalas recruited half of the mothers to interview through referrals from local experts (teachers, social workers, public nurses, clergy, business owners, and public housing officials) and half by posting fliers on public phone booths or personally contacting mothers on street corners. All mothers had incomes putting them below the poverty line in the previous year. Edin and Kefalas tried to get a mixture: 50 Whites, 50 African Americans, and 50 Puerto Ricans, and tried to get one-half over 25 and one-half under 25 years old. They eventually had 162 mothers, 52 whites, 63 African American, and 47 Puerto Rican. Only 40 were over 25 years old, but ages ranged from 15 to 56. They say, "The resulting sample is not random or representative but is quite heterogeneous" (238).

REASONS FOR SAMPLING

When we sample, we select some cases to examine in detail, and then we use what we learn from them to understand a much larger set of cases. Most, but not all, empirical studies use sampling. Depending on the study, the method we use for sampling can differ. Most books on sampling emphasize its use in quantitative research and contain applied mathematics and quantitative examples. The primary use of sampling in quantitative studies is to create a representative sample **(**i.e., a **sample**, a selected small collection of cases or units) that

Sample A small set of cases a researcher selects from a large pool and generalizes to the population.

closely reproduces or represents features of interest in a larger collection of cases, called the **population**.

We examine data in a sample in detail, and if we sampled correctly, we can generalize its results to the entire population. We need to use very precise sampling procedures to create representative samples in quantitative research. These procedures rely on the mathematics of probabilities and hence, are called *probability sampling*.

In most quantitative studies, we want to see how many cases of a population fall into various categories of interest. For example, we might ask how many in the population of all of Chicago's high school students fit into various categories (e.g., high-income family, single-parent family, illegal drug user, delinquent behavior arrestee, musically talented person). We use probability samples in quantitative research because they are very efficient. They save a lot of time and cost for the accuracy they deliver. A properly conducted probability sample may cost 1/1000 the cost and time of gathering information on an entire population, yet it will yield virtually identical results. Let us say we are interested in gathering data on the 18 million people in the United States diagnosed with diabetes. From a well-designed probability sample of 1,800, we can take what we learned and generalize it to all 18 million. It is more efficient to study 1,800 people to learn about 18 million than to study all 18 million people.

Probability samples can be highly accurate. For large populations, data from a well-designed, carefully executed probability sample are often equally if not more accurate than trying to reach every case in the population, but this fact confuses many people. Actually, when the U.S. government planned its 2000 census, all of the social researchers and statistically trained scientists agreed that probability sampling would produce more accurate data than the traditional census method of trying to count every person. A careful probability sample of 30,000 has a very tiny and known error rate. If we try to locate every single person of 300,000,000, systematic errors will slip in unless we take extraordinary efforts and expend huge amounts of time and money. By the way, the government actually conducted the census in the traditional way, but it was for political, not scientific, reasons.

Sampling proceeds differently in qualitative studies and often has a different purpose from quantitative studies. In fact, using the word *sampling* creates confusion in qualitative research because the term is closely associated with quantitative studies (see Luker, 2008:101). In qualitative studies, to allow us to make statements about categories in the population, we rarely sample to gather a small set of cases that is a mathematically accurate reproduction of the entire population. Instead, we sample to identify relevant categories at work in a few cases. In quantitative sampling, we select cases/units. We then treat them as carriers of aspects/features of the social world. A sample of cases/units "stands in" for the much larger population of cases/units. We pick a few to "stand in" for the many. In contrast, the logic of the qualitative sample is to sample aspects/features of the social world. The aspects/features of our sample highlight or "shine light into" key dimensions or processes in a complex social life. We pick a few to provide clarity, insight, and understanding about issues or relationships in the social world. In qualitative sampling, our goal is to deepen understanding about a larger process, relationship, or social scene. A sample gives us valuable information or new aspects. The aspects accentuate, enhance, or enrich key features or situations. We sample to open up new theoretical insights, reveal distinctive aspects of people or social settings, or deepen understanding of complex situations, events, or relationships. In qualitative research, "it is their relevance to the research topic rather than their representativeness which determines the way in which the people to be studied are selected" (Flick, 1998: 41).

We should not overdo the quantitative-qualitative distinction. In a few situations, a study that is primarily quantitative will use the qualitative-sampling

Population The abstract idea of a large group of many cases from which a researcher draws a sample and to which results from a sample are generalized.

strategy and vice versa. Nonetheless, most quantitative studies use probability or probability-like samples while most qualitative studies use a nonprobability method and nonrepresentative strategy.

SAMPLING STRATEGIES

We want to avoid two types of possible sampling mistakes. The first is to conduct sampling in a sloppy or improper manner; the second is to choose a type of sample inappropriate for a study's purpose. The first mistake reminds us to be very meticulous and systematic when we sample. To avoid the second mistake, we need a sampling strategy that matches our specific study's purpose and data. Sampling strategies fall into two broad types: a sample that will accurately represent the population of cases, and all others. We primarily use the first strategy in quantitative studies and the latter in qualitative studies.

Strategies When the Goal Is to Create a Representative Sample

In a representative sample, our goal is to create sample data that mirror or represent many other cases that we cannot directly examine. We can do this in two ways. The first is the preferred method and considered the "gold standard" for representative samples, the *probability sample*. It builds on more than a century of careful reasoning and applied mathematics plus thousands of studies in natural science and quantitative social science. With a probability sampling strategy, we try to create an accurate representative sample that has mathematically predictable errors (i.e., precisely known chances of being "off target"). This sampling approach is complex with several subtypes. Before we examine it, let us look at the second, simpler way to produce a representative sample: to use a nonprobability technique. It is a less accurate substitute when we want a representative sample; however, it is acceptable when probability sampling is impossible, too costly, time consuming, or impractical.

Convenience sampling A nonrandom sample in which the researcher selects anyone he or she happens to come across.

Nonprobability Sampling Techniques. Ideally, we would prefer probability samples when we want to create a representative sample, as a less demanding alternative there are two nonprobability alternatives: convenience and quota samples. In **convenience sampling** (also called *accidental, availability,* or *haphazard sampling*), our primary criteria for selecting cases are that they are easy to reach, convenient, or readily available. This sample type may be legitimate for a few exploratory preliminary studies and some qualitative research studies when our purpose is something other than creating a representative sample. Unfortunately, it often produces very nonrepresentative samples, so it is *not recommended* for creating an accurate sample to represent the population.

When we select cases based on convenience, our sample can seriously misrepresent features in the entire population.[1] You may ask why, if this method is so bad and samples can be seriously nonrepresentative, anyone would use it. The reason is simple: convenience samples are easy, cheap, and quick to obtain. Another reason might be that people are ignorant about how to create a good representative sample. An example of such sampling is the person-on-the-street interview conducted by television programs. Television interviewers go out on the street with camera and microphone to talk to a few people who are convenient to interview. The people walking past a television studio in the middle of the day do not represent everyone. Likewise, television interviewers tend to pick people who look "normal" to them and avoid people who are unattractive, disabled, impoverished, elderly, or inarticulate. Another example is a newspaper that asks readers to clip a questionnaire and mail it in, a Web site that asks users to click on a choice, or a television program that asks viewers to call in their choices. Such samples may have entertainment value, but they easily yield highly misleading data

that do not represent the population even when a large number of people respond.

Maybe you wonder what makes such a sample nonrepresentative. If you want to know about everyone in city XYZ that has a population of 1 million, only some read the newspaper, visit a Web site, or tuned into a program. Also, not everyone who is reading the newspaper, visiting the Web site, or has tuned in is equally interested in an issue. Some people will respond, and there may be many of them (e.g., 50,000), but they are self-selected. We cannot generalize accurately from self-selected people to the entire population. Many in the population do not read the newspaper, visit specific Web sites, or tune into certain television programs, and even if they did, they may lack the interest and motivation to participate. Two key ideas to remember about representative samples are that: (1) self-selection yields a nonrepresentative sample and (2) a big sample size alone is not enough to make a sample representative.

For many purposes, well-designed **quota sampling** is an acceptable nonprobability substitute method for producing a quasi-representative sample.[2] In quota sampling, we first identify relevant categories among the population we are sampling to capture diversity among units (e.g., male and female; or under age 30, ages 30 to 60, over age 60). Next we determine how many cases to get for each category—this is our "quota." Thus, we fix a number of cases in various categories of the sample at the start.

Let us return to the example of sampling residents from city XYZ. You select twenty-five males and twenty-five females under age 30 years of age, fifty males and fifty females aged 30 to 60, and fifteen males and fifteen females over age 60 for a 180-person sample. While this is a start as a population's diversity, it is difficult to represent all possible population characteristics accurately (see Figure 1). Nonetheless, quota sampling ensures that a sample has some diversity. In contrast, in convenience sampling, everyone in a sample might be of the same age, gender, or background. The description of sampling in the *Promises I Can Keep* study at the opening of this chapter used quota sampling (also see Example Box 1, Quota Samples).

Quota sampling is relatively easy. My students conducted an opinion survey of the undergraduate student body using quota sampling. We used three quota categories—gender, class, and minority/majority group status—and a convenience selection method (i.e., a student interviewer approached anyone in the library, a classroom, the cafeteria). We set the numbers to be interviewed in each quota category in advance: 50 percent males and 50 percent females; 35 percent freshman, 25 percent sophomores, 20 percent juniors, and 20 percent seniors; and 10 percent minority and 90 percent majority racially. We picked the proportions based on approximate representation in the student body according to university official records. In the study, a student interviewer approached a person, confirmed that he or she was a student, and verified his or her gender, class, and minority/majority status. If the person fit an unfilled quota (e.g., locate five freshman males who are racial-ethnic minorities), the person was included in the sample and the interviewer proceeded to ask survey questions. If the person did not fit the quota, the interviewer quickly thanked the person without asking survey questions and moved on to someone else.

Quota samples have three weaknesses. First, they capture only a few aspects (e.g., gender and age) of all population diversity and ignore others (e.g., race-ethnicity, area of residence in the city, income level). Second, the fixed number of cases in each category may not accurately reflect the proportion of cases in the total population for the category. Perhaps 20 percent of city residents are over 60 years old but are 10 percent of a quota. Lastly, we use convenience sampling selection for each

Quota sampling A nonrandom sample in which the researcher first identifies general categories into which cases or people will be placed and then selects cases to reach a predetermined number in each category.

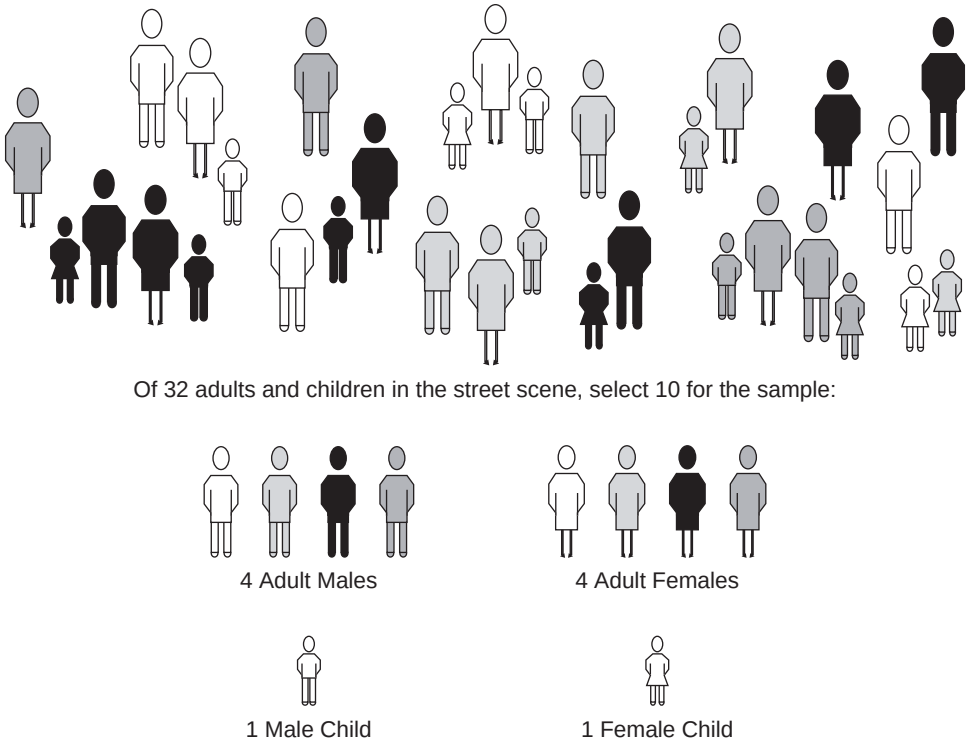


FIGURE 1 Quota Sampling

quota category. For example, we include the first twenty-five males under age 30 we encounter—even if all twenty-five are high-income White lawyers who just returned from a seminar on financial investments. Nothing prevents us from sampling only "friendly"-acting people who want us to pick them.

Probability Sampling Techniques. Probability sampling is the "gold standard" for creating a representative sample. It has a specialized vocabulary that may make it difficult to understand until you learn it, so next we will review some of its vocabulary.

Sampling element The name for a case or single unit to be sampled.

The Language of Probability Sampling. You draw a sample from a large collection of cases/units. Each case/unit is your **sampling element**. It is the unit of analysis or a case in a population. It could be a person, a family, a neighborhood, a nation, an organization, a written document, a symbolic message (television commercial, display of a flag), or a social action (e.g., an arrest, a divorce, or a kiss).

The large collection is the population, but sometimes the word *universe* is used. To define the population, you specify the elements and identify

EXAMPLE BOX 1

Quota Samples

Two studies illustrate different uses of quota sampling in quantitative research. In a study, McMahon, McAlaney, and Edgar (2007) wanted to examine public views of binge drinking in the United Kingdom. They noted that most past research was on young adults and campaigns to curb binge drinking had been ineffective. The authors wanted to learn about public perceptions of binge drinking among the entire adult population. They developed a survey that asked how people defined binge drinking, the extent to which they saw it as a concern, and reasons for and solutions to it. They combined quota sampling with another sampling method to interview 586 people in one city (Inverclyde, Scotland). For quota sampling, interviewers approached potential participants in the streets surrounding a shopping center and invited them to take part in the survey. The quota was based on getting a balance of gender and six age categories. The other method was to go door-to-door in several low-income neighborhoods. The authors learned useful information about views on binge drinking across age groups in both genders in one city. They found wide variation in definitions of binge drinking and support for a "false consensus effect" in which a small number of the heaviest drinkers see their behavior as normal and socially accepted. Nonetheless, the sample is not representative, so findings on the extent of binge drinking in the public and views about it may not reflect the behaviors or views within the city's overall population.

A second study in China by Bian, Breiger, Davis, and Galaskiewicz (2005) employed a targeted use of quota sampling. Their interest was in the difference between the social networks and social ties (e.g., friends, family) among people in different social classes in major Chinese cities. They selected households in four of China's largest metropolitan areas (Shanghai, Shenzhen, Tianjin, and Wuhan), identified a set of neighborhoods in each, and then sampled 100 people per city. They had a list of thirteen occupational titles that represented the full range of the class system in China and 88 percent of all working people in the four cities. Their quota was to get an equal number in each city and a sufficient number of households in each of the thirteen occupational categories for careful analysis. Thus, only 4 percent of the people held the position as manager, but nearly 10 percent of the sample were managers, and 40 percent of people held an industrial worker occupation, but close to 10 percent of people in the sample were industrial workers. The study goal was to test hypotheses about whether a household's social ties are with others of similar or different social classes. They asked households to maintain a written log of social visits (in person or via phone) with other people and recorded the occupation of visitors. This process lasted a year, and researchers interviewed people every three months. The primary interest in the study was to compare patterns of social networks across the various social classes. For example, did managers socialize only with other managers or with people from a wide range of classes? Did industrial workers socialize with industrial workers as well as people in various lower occupations but not in higher occupations? Because the study goal was to compare social network patterns across the various classes, not to have a representative sample that described the Chinese population, it was a highly effective use of quota sampling.

its geographical and temporal boundaries as well as any other relevant boundaries.

Most probability studies with large samples of the entire U.S. population have several boundaries. They include adults over 18 who are residents of the forty-eight continental states and exclude the institutionalized population (i.e., people in hospitals, assisted living and nursing homes, military housing, prisons and jails, homeless and battered women's shelters, college dormitories). Ignoring people in Alaska, Hawaii, and Puerto Rico and excluding the institutionalized population can throw off statistics—for example, the unemployment rate would be higher if the millions of people in prison were included in calculations (see Western and Pettit, 2005). Many studies include only English

speakers, yet as of 2007, roughly 5 percent of U.S. households were "linguistically isolated" (no one over 14 spoke English very well (U.S. Census Bureau, 2007).

To draw a probability sample we start with a population, but *population* is an abstract concept. We must conceptualize and define it more precisely in a process similar to conceptualization in the measurement process, for example, all people in Tampa, Florida, or all college students in the state of Nevada. A **target population** is the specific collection of elements we will study (e.g., noninstitutionalized persons 18 years of age and older with legal residences with the city limits of Tampa on May 15, 2011; students enrolled full-time in an accredited two- or four-year postsecondary educational facility in the state of Nevada in October 2010). In some ways, the target population is analogous to our use of a conceptual definition of the measurement process.

Populations are in constant motion, so we need a temporal boundary. For example, in a city at any given moment, people are dying, boarding or getting off airplanes, and driving across city boundaries in cars. Whom should we count? Do we exclude a long-time city resident who happens to be on vacation when the time is fixed? A population (e.g., persons over the age of 18 who are in the city limits of Milwaukee, Wisconsin, at 12:01 A.M. on March 1, 2011), is an abstract idea. It exists in the mind but is difficult to pinpoint concretely (see Example Box 2, Examples of Populations).

After we conceptualize our population, we need to create an operational definition for the abstract population idea in a way that is analogous to operationalization in the measurement process. We turn the abstract idea into an empirically concrete specific list that closely approximates all population elements. This is our **sampling frame**.

There are many types of sampling frames: telephone directories, tax records, driver's license records, and so on. Listing the elements in a population sounds simple, but it is often difficult because often there is no accurate, up-to-date list of all elements in a population.

A good sampling frame is crucial for accurate sampling. Any mismatch between a sampling frame and the conceptually defined population can create errors. Just as a mismatch between our theoretical and operational definitions of a variable weakens measurement validity, a mismatch between the abstract population and the sampling frame weakens our sampling validity. The most famous case in the history of sampling involved an issue of sampling frames.[3] (See Expansion Box 1, Sampling Frames and the History of Sampling.)

Let us say that our population is all adult residents in the Pacific coast region of the United States in 2010. We contact state departments of transportation to obtain lists of everyone with a driver's

Target population The concretely specified large group of many cases from which a researcher draws a sample and to which results from the sample are generalized.

Sampling frame A list of cases in a population, or the best approximation of them.

EXAMPLE BOX 2

Examples of Populations

1. All persons ages 16 or older living in Australia on December 2, 2009, who were not incarcerated in prison, asylums, and similar institutions
2. All business establishments employing more than 100 persons in Ontario Province, Canada, that operated in the month of July 2005
3. All admissions to public or private hospitals in the state of New Jersey between August 1, 1988, and July 31, 1993
4. All television commercials aired between 7:00 A.M. and 11:00 P.M. Eastern Standard Time on three major U.S. networks between November 1 and November 25, 2004
5. All currently practicing physicians in the United States who received medical degrees between January 1, 1960, and the present
6. All African American male heroin addicts in the Vancouver, British Columbia, or Seattle, Washington, metropolitan areas during 2004

EXPANSION BOX 1

Sampling Frames and the History of Sampling

A famous case in the history of sampling illustrates the limitations of quota sampling and of sampling frames. The *Literary Digest,* a major U.S. magazine, sent postcards to people before the 1920, 1924, 1928, and 1932 U.S. presidential elections. The magazine took the names for its sample from automobile registrations and telephone directories. People returned the postcards indicating for whom they would vote. The magazine correctly predicted all four election outcomes. The magazine's success with predictions was well known, and in 1936, it increased the sample from about 1 million to 10 million. 2.4 million people returned postcards they were sent. The magazine predicted a huge victory for Alf Landon over Franklin D. Roosevelt. But the *Literary Digest* was wrong; Roosevelt won by a landslide. Another random sample of 50,000 by George Gallup was accurate within 1percent of the result.

The prediction was wrong for several reasons, but the sampling mistakes were central. Although the magazine sampled a very large number of people, its sampling frame did not accurately represent the target population (i.e., all voters). It excluded people without telephones or automobiles, a sizable percentage of the population in 1936. The frame excluded as much as 65 percent of the population, particularly a section of the voting population (lower income) that tended to favor Roosevelt. The magazine had been accurate in earlier elections because people with higher and lower incomes did not differ in the way they voted. Also, during earlier elections before the Great Depression, more low-income people could afford to have telephones and automobiles.

The *Literary Digest* mistake teaches us two lessons. First, an accurate sampling frame is crucial. Second, the size of a sample is less important than how accurately it represents the population. A representative sample of 50,000 can give more accurate predictions about the U.S. population than a nonrepresentative sample of 10 million or 50 million.

license in California, Oregon, and Washington. We know some people do not have driver's licenses, although some people drive illegally without them or do not drive. The lists of people with licenses, even if updated regularly, quickly goes out of date as people move into or out of a state. This example shows that before we use official records, such as driver's licenses, as a sampling frame, we must know how officials produce such records. When the state of Oregon instituted a requirement that people show a social security number to obtain a driver's license, the number applying for licenses dropped by 10 percent (from 105,000 issued over three months of 2007 to 93,000 in the same three months of 2008). Thus, thousands disappeared from official records. We could try income tax records, but not everyone pays taxes. Some people cheat and do not pay, others have no income and do not have to file, others have died or have not begun to pay taxes, and still others have entered or left the area since taxes were due. Voter registration records exclude as much as half of the population. In the United States between 53 and 77 percent of eligible voters are registered (Table 401, Statistical Abstract of the United States, 2009). Telephone directories are worse. Many people are not listed in a telephone directory, some people have unlisted numbers, and others have recently moved. With a few exceptions (e.g., a list of all students enrolled at a university), it is difficult to get a perfectly accurate sampling frame. A sampling frame can include those outside the target population (e.g., a telephone directory that lists people who have moved away) or it may omit those within it (e.g., those without telephones). (See Example Box 3, Sampling Frame.)

The ratio of a sample size to the size of the target population is the **sampling ratio**. If the target

Sampling ratio The number of cases in the sample divided by the number of cases in the population or the sampling frame, or the proportion of the population in a sample.

EXAMPLE BOX 3

Sampling Frame

A study by Smith, Mitchell, Attebo, and Leeder (1997) in Australia shows how different sampling frames can influence a sample. The authors examined 2,557 people aged 49 and over living in a defined post code area recruited from a door-to-door census. Of all addresses, people in 80.9 percent were contacted and 87.9 percent of the people responded. The authors searched the telephone directory and the electoral roll for each person. The telephone directory listed 82.2 percent and the electoral roll contained 84.3 percent. Younger people, those who did not own their own homes, and those born outside of Australia were significantly less likely to be included in either sampling frame. The telephone directory was also likely to exclude people with higher occupational prestige while the electoral roll was likely to exclude unmarried persons and males.

population has 50,000 people and the sample has 150, then the sampling ratio is 150/50,000 = 0.003, or 0.3 percent. For a target population of 500 and sample of 100, the sampling ratio is 100/500 = 0.20, or 20 percent. Usually, we use the number of elements in a sampling frame as our best estimate of the size of the target population.

Except for small specialized populations (e.g., all students in a classroom), when we do not need to sample, we use data from a sample to estimate features in the larger population. Any characteristic of a population (e.g., the percentage of city residents who smoke cigarettes, the average height of all women over the age of 21, the percent of people who believe in UFOs) is a population **parameter**. It is the true characteristic of the population. We do not know the parameter with absolute certainty for large populations (e.g., an entire nation), so we can estimate it by using sample data. Information in the sample used to estimate a population parameter is called a **statistic**. (See Figure 2.)

Parameter A characteristic of the entire population that is estimated from a sample.

Statistic A word with several meanings including a numerical estimate of a population parameter computed from a sample.

Random Sampling

In applied mathematics, probability theory relies on random processes. The word *random* has several meanings. In daily life, it can mean unpredictable, unusual, unexpected, or haphazard. In mathematics, random has a specific meaning: a selection process without any pattern. In mathematics, random processes mean that each element will have an equal probability of being selected. We can mathematically calculate the probability of outcomes over many cases with great precision for true random processes.

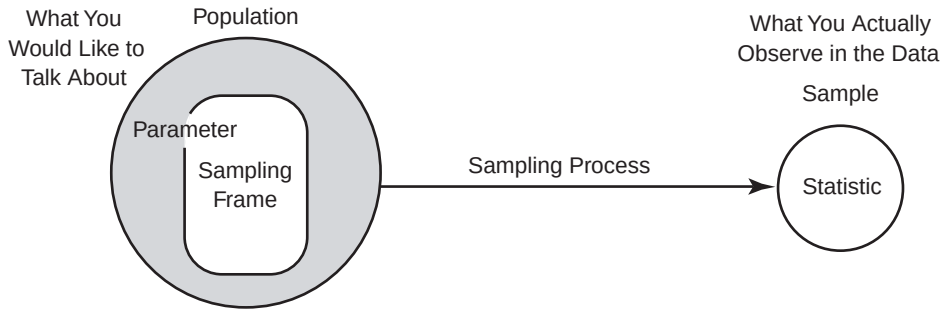


FIGURE 2 A Model of the Logic of Sampling

Random samples yield samples most likely to truly represent the entire population. They also allow us to calculate statistically the relationship between the sample and the population—that is, the size of the **sampling error**. The sampling error is the deviation between what is in the sample data and an ideal population parameter due to random processes.

Probability samples rely on random selection processes. Random selection for sampling requires more precision, time, and effort than samples with nonrandom selection. The formal mathematical procedure specifies exactly which person to pick for the sample, and it may be very difficult to locate that specific person! In sampling, random is not anyone, nor does it mean thoughtless or haphazard. For example, if we are using true random sampling in a telephone survey, we might have to call back six or seven times at different times of the days and on different days, trying to get a specific person whom the mathematically random process identified.[4]

This chapter does not cover all technical and statistical details of random sampling. Instead, we discuss the fundamentals of how probability sampling works, the difference between good and bad samples, how to draw a sample, and basic principles of sampling in social research. If you plan to pursue a career in quantitative research, you will need more mathematical and statistical background on probability and sampling than space permits here.

Five Ways to Sample Randomly

Simple Random. All probability samples are modeled on the **simple random sample** that first specifies the population and target population and identifies its specific sampling elements (e.g., all households in Prescott, Arizona, in March 2011). Next we create an accurate sampling frame and we then use a true random process (discussed later) to pick elements from the sampling frame. Beyond creating an accurate sampling frame, the next difficulty is that we must locate the specific sampled element selected by a random process. If the sampled element is a household, we may have to revisit or call back five times to contact that specific selected household.

To select elements from a sampling frame, we will need to create a list of random numbers that will tell us which elements on it to select. We will need as many random numbers as there are elements to be sampled. The random numbers should range from 1 (the first element on the sampling frame) to the highest number in our sampling frame. If the sampling frame lists 15,000 households, and we want to sample 150 from it, we need a list of 150 random numbers (i.e., numbers generated by a true random process, from 1 to 15,000).

There are two main ways to obtain a list of random numbers. The "old-fashioned" way is to use a **random-number table**. Such tables are available in most statistics and research methods books including this one (see Appendix). The numbers are generated by a pure random process so that any number has an equal probability of appearing in any position. Today most people use computer programs to produce lists of random numbers. Such programs are readily available and often free.

You may ask, once we select an element from the sampling frame, do we then return it to the sampling frame, or do we keep it separate? Unrestricted random sampling is called "random sampling with replacement"—that is, replacing an element after sampling it so it has a chance to be selected again. In simple random sampling without replacement, we "toss out" or ignore elements

Random sample A sample using a mathermatically random method, such as a random-number table or computer program, so that each sampling element of a population has an equal probablity of being selected into the sample.

Sampling error How much a sample deviates from being representative of the population.

Simple random sample A random sample in which a researcher creates a sampling frame and uses a pure random process to select cases so that each sampling element in the population will have an equal probability of being selected.

Random-number table A list of numbers that has no pattern and that researchers use to create a random process for selecting cases and other randomization purposes.

already selected for the sample. For almost all practical purposes in social science, random sampling is without replacement.

We can see the logic of simple random sampling with an elementary example: sampling marbles from a jar. Let us say I have a large jar full of 5,000 marbles, some red and some white. The marble is my sampling element, the 5,000 marbles are my population (both target and ideal), and my sample size is 100. I do not need a sampling frame because I am dealing with small physical objects. The population parameter I want to estimate is the percentage of red marbles in the jar.

I need a random process to select 100 marbles. For small objects, this is easy; I close my eyes, shake the jar, pick one marble, and repeat the procedure 100 times. I now have a random sample of marbles. I count the number of red marbles in my sample to estimate the percentage of red versus white marbles in the population. This is a lot easier than counting all 5,000 marbles. My sample has 52 white and 48 red marbles.

Does this mean that the population parameter is exactly 48 percent red marbles? Maybe or maybe not; because of random chance, my specific sample might be off. I can check my results by dumping the 100 marbles back in the jar, mixing the marbles, and drawing a second random sample of 100 marbles. On the second try, my sample has 49 white marbles and 51 red ones. Now I have a problem. Which is correct? You might ask how good this random sampling business is if different samples from the same population can yield different results. I repeat the procedure over and over until I have drawn 130 different samples of 100 marbles each (see Chart 1 for results). Most people might find it easier to empty the jar and count all 5,000 marbles, but I want to understand the process of sampling. The results of my 130 different samples reveal a clear pattern. The most common mix of red and white marbles is 50/50. Samples that are close to that split are more frequent than those with more uneven splits. The population parameter appears to be 50 percent white and 50 percent red marbles.

Mathematical proofs and empirical tests demonstrate that the pattern found in Chart 1 always appears. The set of many different samples is my **sampling distribution**. It is a distribution of different samples. It reveals the frequency of different sample outcomes from many separate random samples. This pattern appears if the sample size is 1,000 instead of 100, if there are 10 colors of marbles instead of 2, if the population has 100 marbles or 10 million marbles instead of 5,000, and if the sample elements are people, automobiles, or colleges instead of marbles. In fact, the "bell-shaped" sampling distribution pattern becomes clearer as I draw more and more independent random samples from a population.

The sampling distribution pattern tells us that over many separate samples, the true population parameter (i.e., the 50/50 split in the preceding example) is more common than any other outcome. Some samples may deviate from the population parameter, but they are less common. When we plot many random samples as in the graph in Chart 1, the sampling distribution always looks like a normal or bell-shaped curve. Such a curve is theoretically important and is used throughout statistics. The area under a bell-shaped curve is well known or, in this example, we can quickly figure out the odds that we will get a specific number of marbles. If the true population parameter is 50/50, standard statistical charts tell what the odds of getting 50/50 or a 40/50 or any other split in a random sample are.

The **central limit theorem** from mathematics tells us that as the number of different random samples in a sampling distribution increases toward infinity, the pattern of samples and of the population parameter becomes increasingly predictable. For a huge number of random samples, the sampling distribution always forms a normal curve, and the midpoint of the curve will be the population parameter.

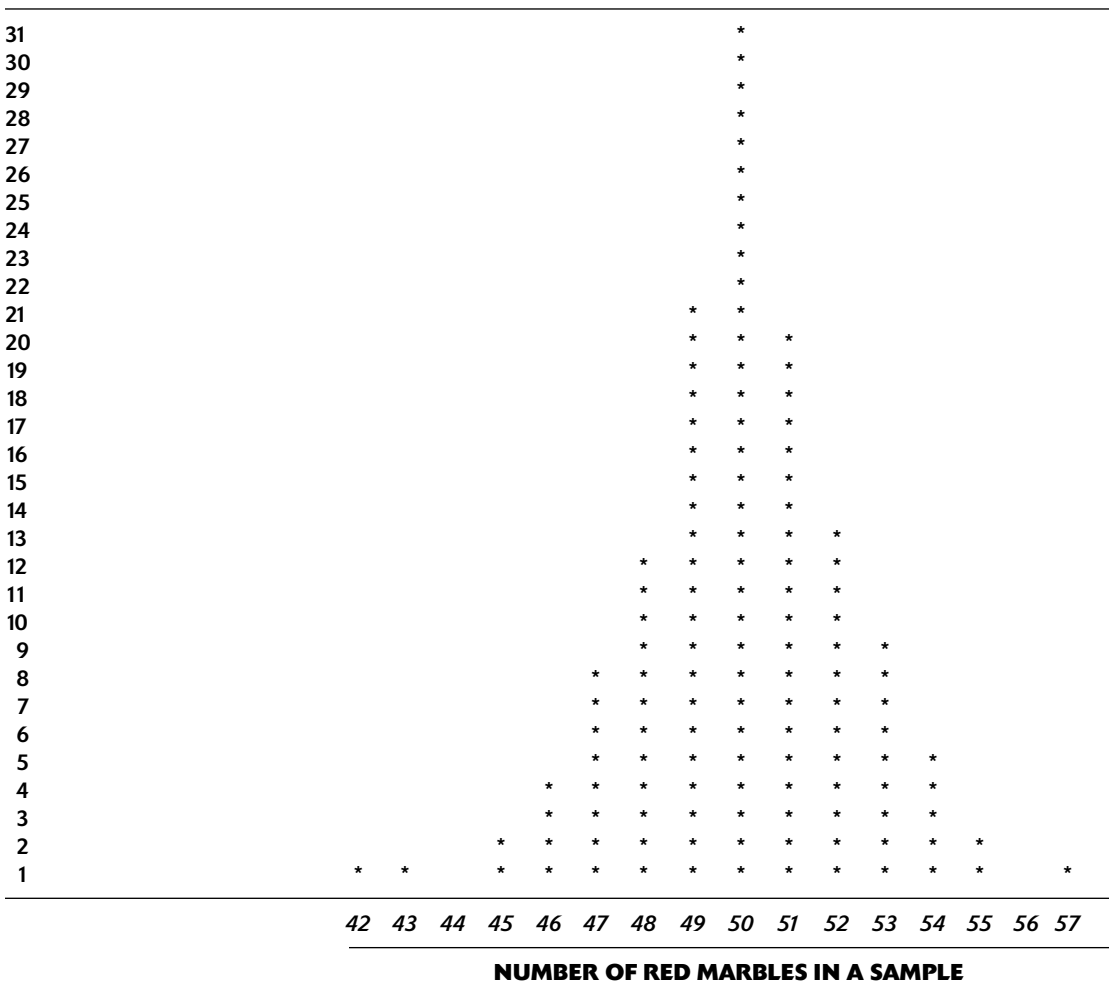
Sampling distribution A distribution created by drawing many random samples from the same population.

Central limit theorem A mathematical relationship that states when many random samples are drawn from a population, a normal distribution is formed, and the center of the distribution for a variable equals the population parameter.

CHART 1 Example of Sampling Distribution

RED	WHITE	NUMBER OF SAMPLES
42	58	1
43	57	1
45	55	2
46	54	4
47	53	8
48	52	12
49	51	21
50	50	31
51	49	20
52	48	13
53	47	9
54	46	5
55	45	2
57	43	1
	Total	130

Number of red and white marbles that were randomly drawn from a jar of 5,000 marbles with 100 drawn each time, repeated 130 times for 130 independent random samples.



You probably do not have the time or energy to draw many different samples and just want to draw one sample. You are not alone. We rarely draw many random samples except to verify the central limit theorem. We draw only one random sample, but the central limit theorem lets us generalize from one sample to the population. The theorem is about many samples, but it allows us to calculate the probability that a particular sample is off from the population parameter. We will not go into the calculations here.

The important point is that random sampling does not guarantee that every random sample perfectly represents the population. Instead, it means that most random samples will be close to the population parameter most of the time. In addition, we can calculate the precise probability that a particular sample is inaccurate. The central limit theorem lets us estimate the chance that a particular sample is unrepresentative or how much it deviates from the population parameter. It lets us estimate the size of the sampling error. We do this by using information from one sample to estimate the sampling distribution and then combine this information with knowledge of the central limit theorem and area under a normal curve. This lets us create something very important, **confidence intervals**.

The confidence interval is a simple but very powerful idea. When television or newspaper polls are reported, you may hear about what journalists call the "margin of error" being plus or minus 2 percentage points. This is a version of confidence interval, which is a range around a specific point that we use to estimate a population parameter.

Confidence intervals A range of values, usually a little higher and lower than a specific value found in a sample, within which a researcher has a specified and high degree of confidence that the population parameter lies.

Systematic sampling A random sample in which a researcher selects every *k*th (e.g., third or twelfth) case in the sampling frame using a sampling interval.

Sampling interval The inverse of the sampling ratio that is used when selecting cases in systematic sampling.

We use a range because the statistics of random processes are based on probability. They do not let us predict an exact point. They do allow us to say with a high level of confidence (e.g., 95 percent) that the true population parameter lies within a certain range (i.e., the confidence interval). The calculations for sampling errors or confidence intervals are beyond the level of the discussion here. Nonetheless, the sampling distribution is the key idea that tells us the sampling error and confidence interval. Thus, we cannot say, "This sample gives a perfect measure of the population parameter," but we can say, "We are 95 percent certain that the true population parameter is no more than 2 percent different from what was have found in the sample." (See Expansion Box 2, Confidence Intervals.)

Going back to the marble example, I cannot say, "There are precisely 2,500 red marbles in the jar based on a random sample." However, I can say, "I am 95 percent certain that the population parameter lies between 2,450 and 2,550." I combine the characteristics of my sample (e.g., its size, the variation in it) with the central limit theorem to predict specific ranges around the population parameter with a specific degree of confidence.

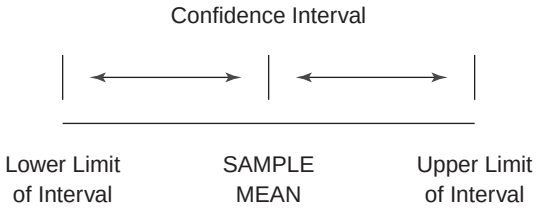
Systematic Sampling. **Systematic sampling** is a simple random sampling with a shortcut selection procedure. Everything is the same except that instead of using a list of random numbers, we first calculate a **sampling interval** to create a quasi-random selection method. The sampling interval (i.e., 1 in *k*, where *k* is some number) tells us how to select elements from a sampling frame by skipping elements in the frame before selecting one for the sample.

For instance, we want to sample 300 names from 900. After a random starting point, we select every third name of the 900 to get a sample of 300. The sampling interval is 3. Sampling intervals are easy to compute. We need the sample size and the population size (or sampling frame size as a best estimate). We can think of the sampling interval as the inverse of the sampling ratio. The sampling ratio for 300 names out of 900 is 300/900 = .333 = 33.3 percent. The sampling interval is 900/300 = 3.

EXPANSION BOX 2

Confidence Intervals

The confidence interval is a simple and very powerful idea; it has excellent mathematics behind it and some nice formulas. If you have a good mathematics background, this concept could be helpful. If you are nervous about complex mathematical formulas with many Greek symbols, here is a simple example with a simple formula (a minimum of Greek). The interval is a range that goes above and below an estimate of some characteristic of the population (i.e., population parameters), such as its average or statistical mean. The interval has an upper and lower limit. The example illustrates a simplified way to calculate a confidence interval and shows how sample size and sample homogeneity affect it.



Let us say you draw a sample of nine 12-year-old children. You weigh them and find that their average weight, the mean, is 90 pounds with a standard deviation of 36 pounds. You want to create a confidence interval around your best estimate of the population parameter (the mean weight for the population of all 12-year-olds). You symbolize the population parameter with the Greek letter μ.

Here is how to figure out a confidence interval for the population mean based on a simple random sample. You estimate a confidence level around μ by adding and subtracting a range above and below the sample mean, your best estimate of μ.

To calculate the confidence interval around the sample mean, you first calculate something called the *standard error of the mean*. Call it standard error for short. It is your estimate of variability in the sampling distribution. You use another Greek letter, σ, to symbolize the standard deviation and add the letter *m* as a subscript to it, indicating that it is your estimate of the standard deviation in the sampling distribution. Thus, the standard error comes from the standard deviation in the sampling distribution of all possible random samples from the population.

You estimate the standard deviation of the sampling distribution by getting the standard deviation of your sample and adjusting it slightly. To simplify this example, you skip the adjustment and assume that it equals the sample standard deviation. To get the standard error, you adjust it for your sample size symbolized by the letter *N*. The formula for it is:

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

Let us make the example more concrete. For the example, let us look at weight among nine 12-year-olds. For the sampling distribution of the mean you use a mean of 90 pounds and a standard deviation of 36/3 = 12 (note the square root of 9 = 3). The confidence interval has a low and upper limit. Here are formulas for them.

Lower limit $M - Z_{.95}\sigma_m$
Upper limit $M + Z_{.95}\sigma_m$

In addition to the σ_m there are two other symbols now:

M in the formula stands for mean in your sample.
$Z_{.95}$ stands for the z-score under a bell-shaped or normal curve at a 95 percent level of confidence (the most typical level). The z-score for a normal curve is a standard number (i.e., it is always the same for 95 percent level of confidence, and it happens to be 1.96). We could pick some confidence level other than 95 percent, but it is the most typical one used.

You now have everything you need to calculate upper and lower limits of the confidence interval. You compute them by adding and subtracting 1.96 standard deviations to/from the mean of 90 as follows:

Lower limit $90 - (1.96)(12) = 66.48$
Upper limit $90 + (1.96)(12) = 113.52$

(*continued*)

EXPANSION BOX 2
(continued)

This says you can be 95 percent confident that the population parameter lies somewhere between 66.48 and 113.52 pounds. You determined the upper and lower limits by adding and subtracting an amount to the sample mean (90 pounds in your example). You use 1.96 because it is the *z*-score when you want to be 95 percent confident. You calculated 12 as the standard error of the mean based on your sample size and the standard deviation of your sample.

You might see the wide range of 66 to 113 pounds and think it is large, and you might ask why is the sample small, with just nine children?

Here is how increasing the sample size affects the confidence interval. Let us say that instead of a sample of nine children you had 900 12-year-olds (luckily the square root of 900 is easy to figure out: 30). If everything remained the same, your σ_m with a sample of 900 is 36/30 = 1.2. Now your confidence interval is as follows

Lower limit $90 - (1.96)(1.2) = 87.765$
Upper limit $90 + (1.96)(1.2) = 92.352$

With the much larger sample size, you can be 95 percent confident that the population parameter of average weight is somewhere between 87.765 and 92.352 pounds.

Here is how having a very homogeneous sample affects the confidence interval. Let us say that you had a standard deviation of 3.6 pounds, not 36 pounds. If everything else remained the same, your σ_m with a standard deviation of 3.6 is 3.6/9 = 0.4

Now your confidence interval is as follows

Lower limit $90 - (1.96)(0.4) = 89.215$
Upper limit $90 + (1.96)(0.4) = 90.784$

With the very homogeneous sample, you can be 95 percent confident that the population parameter of average weight is somewhere between 89.215 and 90.784 pounds.

Let us review the confidence intervals as sample size and standard deviation change:

Sample size = 9, standard deviation = 36. Confidence interval is 66 to 113 pounds.
Sample size = 900, standard deviation = 36. Confidence interval is 87.765 to 92.352 pounds.
Sample size = 9, standard deviation = 3.6 pounds. Confidence interval is 89.215 to 90.784 pounds.

In most cases, a simple random sample and a systematic sample yield equivalent results. One important situation in which systematic sampling cannot be substituted for simple random sampling occurs when the elements in a sample are organized in some kind of cycle or pattern. For example, our sampling frame is organized as a list of married couples with the male first and the female second (see Table 1). Such a pattern gives us an unrepresentative sample if systematic sampling is used. Our systematic sample can be nonrepresentative and include only wives because of the organization of the cases. When our sample frame is organized as couples, even-numbered sampling intervals result in samples with all husbands or all wives.

Figure 3 illustrates simple random sampling and systematic sampling. Notice that different names were drawn in each sample. For example, H. Adams appears in both samples, but C. Droullard

TABLE 1 Problems with Systematic Sampling of Cyclical Data

CASE	
1	Husband
2[a]	Wife
3	Husband
4	Wife
5	Husband
6[a]	Wife
7	Husband
8	Wife
9	Husband
10[a]	Wife
11	Husband
12	Wife

Random start = 2; Sampling interval = 4.
[a]Selected into sample.

FIGURE 3 How to Draw Simple Random and Systematic Samples

1. Number each case in the sampling frame in sequence. The list of 40 names is in alphabetical order, numbered from 1 to 40.
2. Decide on a sample size. We will draw two 25 percent (10-name) samples.
3. For a *simple random sample,* locate a random-number table (see excerpt to this figure). Before using the random-number table, count the largest number of digits needed for the sample (e.g., with 40 names, two digits are needed; for 100 to 999, three digits; for 1,000 to 9,999, four digits). Begin anywhere on the random-number table (we will begin in the upper left) and take a set of digits (we will take the last two). Mark the number on the sampling frame that corresponds to the chosen random number to indicate that the case is in the sample. If the number is too large (over 40), ignore it. If the number appears more than once (10 and 21 occurred twice in the example), ignore the second occurrence. Continue until the number of cases in the sample (10 in our example) is reached.
4. For a *systematic sample,* begin with a random start. The easiest way to do this is to point blindly at the random-number table, then take the closest number that appears on the sampling frame. In the example, 18 was chosen. Start with the random number and then count the sampling interval, or 4 in our example, to come to the first number. Mark it, and then count the sampling interval for the next number. Continue to the end of the list. Continue counting the sampling interval as if the beginning of the list were attached to the end of the list (like a circle). Keep counting until ending close to the start, or on the start if the sampling interval divides evenly into the total of the sampling frame.

No.	*Name (Gender)*	*Simple Random*	*Systematic*	*No.*	*Name (Gender)*	*Simple Random*	*Systematic*
01	Abrams, J. (M)			21	Hjelmhaug, N. (M)	Yes*	
02	Adams, H. (F)	Yes	Yes (6)	22	Huang, J. (F)	Yes	Yes (1)
03	Anderson, H. (M)			23	Ivono, V. (F)		
04	Arminond, L. (M)			24	Jaquees, J. (M)		
05	Boorstein, A. (M)			25	Johnson, A. (F)		
06	Breitsprecher, P. (M)	Yes	Yes (7)	26	Kennedy, M. (F)		Yes (2)
07	Brown, D. (F)			27	Koschoreck, L. (F)		
08	Cattelino, J. (F)			28	Koykkar, J. (M)		
09	Cidoni, S. (M)			29	Kozlowski, C. (F)	Yes	
10	Davis, L. (F)	Yes*	Yes (8)	30	Laurent, J. (M)		Yes (3)
11	Droullard, C. (M)	Yes		31	Lee, R. (F)		
12	Durette, R. (F)			32	Ling, C. (M)		
13	Elsnau, K. (F)	Yes		33	McKinnon, K. (F)		
14	Falconer, T. (M)		Yes (9)	34	Min, H. (F)	Yes	Yes (4)
15	Fuerstenberg, J. (M)			35	Moini, A. (F)		
16	Fulton, P. (F)			36	Navarre, H. (M)		
17	Gnewuch, S. (F)			37	O'Sullivan, C. (M)		
18	Green, C. (M)		START, Yes (10)	38	Oh, J. (M)		Yes (5)
				39	Olson, J. (M)		
19	Goodwanda, T. (F)	Yes		40	Ortiz y Garcia, L. (F)		
20	Harris, B. (M)						

Excerpt from a Random-Number Table (for Simple Random Sample)

15010	18590	00102	42210	94174	22099
90122	38221	21529	00013	04734	60457
67256	13887	94119	11077	01061	27779
13761	23390	12947	21280	44506	36457
81994	66611	16597	44457	07621	51949
79180	25992	46178	23992	62108	43232
07984	47169	88094	82752	15318	11921

*Numbers that appeared twice in random numbers selected.

is in only the simple random sample. This is because it is rare for any two random samples to be identical.

The sampling frame contains twenty males and twenty females (gender is in parentheses after each name). The simple random sample yielded three males and seven females, and the systematic sample yielded five males and five females. Does this mean that systematic sampling is more accurate? No. To check this, we draw a new sample using different random numbers, taking the first two digits and beginning at the end (e.g., 11 from 11921 and then 43 from 43232). Also, we draw a new systematic sample with a different random start. The last time the random start was 18, but we now try a random start of 11. What did we find? How many of each gender?[5]

Stratified Sampling. When we use **stratified sampling**, we first divide the population into subpopulations (strata) on the basis of supplementary information.[6] After dividing the population into strata, we draw a random sample from each subpopulation. In stratified sampling, we control the relative size of each stratum rather than letting random processes control it. This guarantees representativeness or fixes the proportion of different strata within a sample. Of course, the necessary information about strata is not always available.

In general, if the stratum information is accurate, stratified sampling produces samples that are more representative of the population than those of simple random sampling. A simple example illustrates why this is so. Imagine a population that is 51 percent female and 49 percent male; the population parameter is a gender ratio of 51 to 49. With stratified sampling, we draw random samples among females and among males so that the sample contains a 51 to 49 percent gender ratio. If we had used simple random sampling, it would be possible for a random sample to be off from the true gender ratio in the population. Thus, we have fewer errors representing the population and a smaller sampling error with stratified sampling.

Stratified sampling A random sample in which the researcher first identifies a set of mutually exclusive and exhaustive categories, divides the sampling frame by the categories, and then uses random selection to select cases from each category.

We use stratified sampling when a stratum of interest is a small percentage of a population and random processes could miss the stratum by chance. For example, we draw a sample of 200 from 20,000 college students using information from the college registrar's office. It indicates that 2 percent of the 20,000 students, or 400, are divorced women with children under the age of 5. For our study, this group is important to include in the sample. There would be four such students (2 percent of 200) in a representative sample, but we could miss them by chance in one simple random sample. With stratified sampling, we obtain a list of the 400 such students from the registrar and randomly select four from it. This guarantees that the sample represents the population with regard to the important strata (see Example Box 4, Illustration of Stratified Sampling).

In special situations, we may want the proportion of a stratum in a sample to differ from its true proportion in the population. For example, the population contains 0.5 percent Aleuts, but we want to examine Aleuts in particular. We oversample so that Aleuts make up 10 percent of the sample. With this type of disproportionate stratified sample, we cannot generalize directly from the sample to the population without special adjustments.

In some situations, we want the proportion of a stratum or subgroup to differ from its true proportion in the population. For example, Davis and Smith (1992) reported that the 1987 General Social Survey oversampled African Americans. A random sample of the U.S. population yielded 191 Blacks. Davis and Smith conducted a separate sample of African Americans to increase it to 544. The 191 Black respondents are about 13 percent of the random sample, roughly equal to the percentage of Blacks in the U.S. population. The 544 Blacks are 30 percent of the disproportionate sample. The researcher who wants to use the entire sample must adjust it to reduce the number of sampled African Americans before generalizing to the U.S. population. Disproportionate sampling helps the researcher who wants to focus on issues

EXAMPLE BOX 4

Illustration of Stratified Sampling

Sample of 100 Staff of General Hospital, Stratified by Position

POSITION	POPULATION		SIMPLE RANDOM SAMPLE	STRATIFIED SAMPLE	ERRORS COMPARED TO THE POPULATION
	N	*Percent*	*n*	*n*	
Administrators	15	2.88	1	3	–2
Staff physicians	25	4.81	2	5	–3
Intern physicians	25	4.81	6	5	+1
Registered nurses	100	19.23	22	19	+3
Nurse assistants	100	19.23	21	19	+2
Medical technicians	75	14.42	9	14	+5
Orderlies	50	9.62	8	10	–2
Clerks	75	14.42	5	14	+1
Maintenance staff	30	5.77	3	6	–3
Cleaning staff	25	4.81	3	5	–2
Total	520	100.00	100	100	

Randomly select 3 of 15 administrators, 5 of 25 staff physicians, and so on.

Note: Traditionally, *N* symbolizes the number in the population and *n* represents the number in the sample.

The simple random sample overrepresents nurses, nursing assistants, and medical technicians but underrepresents administrators, staff physicians, maintenance staff, and cleaning staff. The stratified sample gives an accurate representation of each position.

most relevant to a subpopulation. In this case, he or she can more accurately generalize to African Americans using the 544 respondents instead of a sample of only 191. The larger sample is more likely to reflect the full diversity of the African American subpopulation.

Cluster Sampling. We use **cluster sampling** to address two problems: the lack of a good sampling frame for a dispersed population and the high cost to reach a sampled element.[7] For example, there is no single list of all automobile mechanics in North America. Even if we had an accurate sampling frame, it would cost too much to reach the sampled mechanics who are geographically spread out. Instead of using a single sampling frame, we use a sampling design that involves multiple stages and clusters.

A *cluster* is a unit that contains final sampling elements but can be treated temporarily as a sampling element itself. First we sample clusters, and then we draw a second sample from within the clusters selected in the first stage of sampling. We randomly sample clusters and then randomly sample elements from within the selected clusters. This has a significant practical advantage when we can create a good sampling frame of clusters even if it is impossible to create one for sampling elements. Once we have a sample of clusters, creating a sampling frame for elements within each cluster becomes manageable. A second advantage for geographically dispersed populations is that elements within each cluster are physically closer to one another, which can produce a savings in locating or reaching each element.

Cluster sampling A type of random sample that uses multiple stages and is often used to cover wide geographic areas in which aggregated units are randomly selected and then samples are drawn from the sampled aggregated units or clusters.

We draw several samples in stages in cluster sampling. In a three-stage sample, stage 1 is a random sampling of large clusters; stage 2 is a random sampling of small clusters within each selected large cluster; and the last stage is a sampling of elements from within the sampled small clusters. For example, we want a sample of individuals from Mapleville. First, we randomly sample city blocks, then households within blocks, and then individuals within households (see Chart 2). Although there is no accurate list of all residents of Mapleville, there is an accurate list of blocks in the city. After selecting a random sample of blocks, we count all households on the selected blocks to create a sample frame for each block. Then we use the list of households to draw a random sample at the stage of sampling households. Finally, we choose a specific individual within each sampled household.

Cluster sampling is usually less expensive than simple random sampling, but it is less accurate. Each stage in cluster sampling introduces sampling errors, so a multistage cluster sample has more sampling errors than a one-stage random sample.[8]

When we use cluster sampling, we must decide the number of clusters and the number of elements within clusters. For example, in a two-stage cluster sample of 240 people from Mapleville, we could randomly select 120 clusters and select 2 elements from each or randomly select two clusters and select 120 elements in each. Which is better? A design with more clusters is better because elements within clusters (e.g., people living on the same block) tend to be similar to each other (e.g., people on the same block tend to be more alike than those on different blocks). If few clusters are chosen, many similar elements could be selected, which would be less representative of the total population. For example, we could select two blocks with relatively wealthy people and draw 120 people from each block. This would be less representative than a sample with 120 different city blocks and 2 individuals chosen from each.

When we sample from a large geographical area and must travel to each element, cluster sampling significantly reduces travel costs. As usual, there is a trade-off between accuracy and cost. For example, Alan, Ricardo, and Barbara each personally interview a sample of 1,500 students who represent the population of all college students in North America. Alan obtains an accurate sampling frame of all students and uses simple random sampling. He travels to 1,000 different locations to interview one or two students at each. Ricardo draws a random sample of three colleges from a list of all 3,000 colleges and then visits the three and selects 500 students from each. Barbara draws a random sample of 300 colleges. She visits the 300 and selects 5 students at each. If travel costs average $250 per location, Alan's travel bill is $250,000, Ricardo's is $750, and Barbara's is $75,000. Alan's sample is highly accurate, but Barbara's is only slightly less accurate for one-third the cost. Ricardo's sample is the cheapest, but it is not representative.

Within-Household Sampling. Once we sample a household or similar unit (e.g., family or dwelling unit) in cluster sampling, the question arises as to whom we should choose. A potential source of bias is introduced if the first person who answers the telephone, the door, or the mail is used in the sample. The first person who answers should be selected only if his or her answering is the result of a truly random process. This is rarely the case. Certain people are unlikely to be at home, and in some households one person (e.g., a husband) is more likely than another to answer the telephone or door. Researchers use within-household sampling to ensure that after a random household is chosen, the individual within the household is also selected randomly.

We can randomly select a person within a household in several ways.[9] The most common method is to use a selection table specifying whom you should pick (e.g., oldest male, youngest female) after determining the size and composition of the household (see Table 2). This removes any bias that might arise from choosing the first person to answer the door or telephone or from the interviewer's selection of the person who appears to be friendliest.

Probability Proportionate to Size (PPS). There are two ways we can draw cluster samples. The method just described is proportionate or unweighted

CHART 2 Illustration of Cluster Sampling

Goal: Draw a random sample of 240 people in Mapleville.

Step 1: Mapleville has 55 districts. Randomly select 6 districts.

1 2 3* 4 5 6 7 8 9 10 11 12 13 14 15* 16 17 18 19 20 21 22 23 24 25 26
27* 28 29 30 31* 32 33 34 35 36 37 38 39 40* 41 42 43 44 45 46 47 48
49 50 51 52 53 54* 55

* = Randomly selected.

Step 2: Divide the selected districts into blocks. Each district contains 20 blocks. Randomly select 4 blocks from the district.

Example of District 3 (selected in step 1):

1 2 3 4* 5 6 7 8 9 10* 11 12 13* 14 15 16 17* 18 19 20

* = Randomly selected.

Step 3: Divide blocks into households. Randomly select households.

Example of Block 4 of District 3 (selected in step 2):

Block 4 contains a mix of single-family homes, duplexes, and four-unit apartment buildings. It is bounded by Oak Street, River Road, South Avenue, and Greenview Drive. There are 45 households on the block. Randomly select 10 households from the 45.

1	#1 Oak Street	16	"	31*	"
2	#3 Oak Street	17*	#154 River Road	32*	"
3*	#5 Oak Street	18	#156 River Road	33	"
4	"	19*	#158 River Road	34	#156 Greenview Drive
5	"	20*	"	35*	"
6	"	21	#13 South Avenue	36	"
7	#7 Oak Street	22	"	37	"
8	"	23	#11 South Avenue	38	"
9*	#150 River Road	24	#9 South Avenue	39	#158 Greenview Drive
10*	"	25	#7 South Avenue	40	"
11	"	26	#5 South Avenue	41	"
12	"	27	#3 South Avenue	42	"
13	#152 River Road	28	#1 South Avenue	43	#160 Greenview Drive
14	"	29*	"	44	"
15	"	30	#152 Greenview Drive	45	"

* = Randomly selected.

Step 4: Select a respondent within each household.

Summary of cluster sampling:
1 person randomly selected per household
10 households randomly selected per block
4 blocks randomly selected per district
6 districts randomly selected in the city
1 x 10 x 4 x 6 = 240 people in sample

TABLE 2 Within-Household Sampling

Selecting individuals within sampled households. Number selected is the household chosen in Chart 2.

NUMBER	*LAST NAME*	*ADULTS (OVER AGE 18)*	*SELECTED RESPONDENT*
3	Able	1 male, 1 female	Female
9	Bharadwaj	2 females	Youngest female
10	DiPiazza	1 male, 2 females	Oldest female
17	Wucivic	2 males, 1 female	Youngest male
19	Cseri	2 females	Youngest female
20	Taylor	1 male, 3 females	Second oldest female
29	Velu	2 males, 2 females	Oldest male
31	Wong	1 male, 1 female	Female
32	Gray	1 male	Male
35	Mall-Krinke	1 male, 2 females	Oldest female

EXAMPLE SELECTION TABLE (ONLY ADULTS COUNTED)

MALES	*FEMALES*	*WHOM TO SELECT*	*MALES*	*FEMALES*	*WHOM TO SELECT*
1	0	Male	2	2	Oldest male
2	0	Oldest male	2	3	Youngest female
3	0	Youngest male	3	2	Second oldest male
4+	0	Second oldest male	3	3	Second oldest female
0	1	Female	3	4	Third oldest female
0	2	Youngest female	4	3	Second oldest male
0	3	Second oldest female	4	4	Third oldest male
0	4+	Oldest female	4	5+	Youngest female
1	1	Female	5+	4	Second oldest male
1	2	Oldest female	5+	5+	Fourth oldest female
1	3	Second oldest female			
2	1	Youngest male			
3	1	Second oldest male			

+ = or more

cluster sampling. It is proportionate because the size of each cluster (or number of elements at each stage) is the same. The more common situation is for the cluster groups to be of different sizes. When this is the case, we must adjust the probability for each stage in sampling.

The foregoing example with Alan, Barbara, and Ricardo illustrates the problem with unweighted cluster sampling. Barbara drew a simple random sample of 300 colleges from a list of all 3,000 colleges, but she made a mistake—unless every college has an identical number of students. Her method gave each college an equal chance of being selected—a 300/3,000, or 10 percent chance. But colleges have different numbers of students, so each student does not have an equal chance to end up in her sample.

Barbara listed every college and sampled from the list. A large university with 40,000 students and a small college with 400 students had an equal chance of being selected. But if she chose the large university, the chance of a given student

at that college being selected was 5 in 40,000 (5/40,000 = 0.0125 percent), whereas a student at the small college had a 5 in 400 (5/400 = 1.25 percent) chance of being selected. The small-college student was 100 times more likely to be in her sample. The total probability of a student from the large university being selected was 0.125 percent (10×0.0125) while it was 12.5 percent (10×1.25) for the small-college student. Barbara violated a principle of random sampling: that each element has an equal chance to be selected into the sample.

If Barbara uses **probability proportionate to size (PPS)** and samples correctly, then each final sampling element or student will have an equal probability of being selected. She does this by adjusting the chances of selecting a college in the first stage of sampling. She must give large colleges with more students a greater chance of being selected and small colleges a smaller chance. She adjusts the probability of selecting a college on the basis of the proportion of all students in the population who attend it. Thus, a college with 40,000 students will be 100 times more likely to be selected than one with 400 students. (See Example Box 5, Probability Proportionate to Size (PPS) Sampling.)

Random-Digit Dialing. **Random-digit dialing (RDD)** is a sampling technique used in research projects in which the general public is interviewed by telephone.[10] It does not use the published telephone directory as the sampling frame. Using a telephone directory as the sampling frame misses three kinds of people: those without telephones, those who have recently moved, and those with unlisted numbers. Those without phones (e.g., the poor, the uneducated, and transients) are missed in any telephone interview study, but 95 percent of people in advanced industrialized nations have a telephone. Several types of people have unlisted numbers: those who want to avoid collection agencies; those who are very wealthy; and those who want to have privacy and to avoid obscene calls, salespeople, and prank calls. In some urban areas in the United States, the percentage of unlisted numbers is 50 percent. In addition, people change their residences, so annual directories have numbers for people who have moved away and do not list those who have recently moved into an area.

If we use RDD, we randomly select telephone numbers, thereby avoiding the problems of telephone directories. The population is telephone numbers, not people with telephones. RDD is not difficult, but it takes time and can frustrate the person doing the calling.

Here is how RDD works in the United States. Telephone numbers have three parts: a three-digit area code, a three-digit exchange number or central office code, and a four-digit number. For example, the area code for Madison, Wisconsin, is 608, and there are many exchanges within the area code (e.g., 221, 993, 767, 455), but not all of the 999 possible three-digit exchanges (from 001 to 999) are active. Likewise, not all of the 9,999 possible four-digit numbers in an exchange (from 0000 to 9999) are being used. Some numbers are reserved for future expansion, are disconnected, or are temporarily withdrawn after someone moves. Thus, a possible U.S. telephone number consists of an active area code, an active exchange number, and a four-digit number in an exchange.

In RDD, a researcher identifies active area codes and exchanges and then randomly selects four-digit numbers. A problem is that the researcher can select any number in an exchange. This means that some selected numbers are out of service, disconnected, pay phones, or numbers for businesses; only some numbers are what the researcher wants: working residential phone numbers. Until the researcher calls, it is not possible to know whether the number is a working residential number. This means spending much time reaching numbers that are disconnected, are for businesses, and so forth. Research organizations often use

Probability proportionate to size (PPS) An adjustment made in cluster sampling when each cluster does not have the same number of sampling elements.

Random-digit dialing (RDD) A method of randomly selecting cases for telephone interviews that uses all possible telephone numbers as a sampling frame.

EXAMPLE BOX 5

Probability Proportionate to Size (PPS) Sampling

Henry wants to conduct one-hour, in-person interviews with people living in the city of Riverdale, which is spread out over a large area. Henry wants to reduce his travel time and expenses, so he uses a *cluster sampling design*. The last census reported that the city had about 490,000 people. Henry can interview only about 220 people, or about 0.05 percent of the city population. He first gathers maps from the city tax office and fire department, and retrieves census information on city blocks. He learns that there are 2,182 city blocks. At first, he thinks he can randomly select 10 percent of the blocks (i.e., 218), go to a block and count housing units, and then locate one person to interview in each housing unit (house, apartment, etc.), but the blocks are of unequal geographic and population size. He studies the population density of the blocks and estimates the number of people in each, and then develops a five-part classification based on the average size of a block as in the following chart.

Block Type	*Number of Clusters*	*Average Number People per Block*
Very high density	20	2,000
High density	200	800
Medium density	800	300
Low density	1,000	50
Semirural	162	10

Henry realizes that randomly selecting city blocks without adjustment will not give each person an equal chance of being selected. For example, 1 very high-density block has the same number of people as 40 low-density blocks. Henry adjusts proportionately to the block size. The easiest way to do this is to convert all city blocks to equal-size units based on the smallest cluster, or the semirural city blocks. For example, there are 2,000/10 or 200 times more people in a high-density block than a semirural block, so Henry increases the odds of selecting such a block to make its probability 200 times higher than a semirural block. Essentially, Henry creates adjusted cluster units of 10 persons each (because that is how many there are in the semirural blocks) and substitutes them for city blocks in the first stage of sampling. The 162 semirural blocks are unchanged, but after adjustment, he has 20 X 200 = 4,000 units for the very high density blocks, 200 X 80 = 16,000 units for the high-density blocks, and so forth, for a total of 49,162 such units. Henry now numbers each block, using the adjusted cluster units, with many blocks getting multiple numbers. For example, he assigns numbers 1 to 200 to the first very high density block, and so forth, as follows:

1 Very high density block #1
2 Very high density block #1
3 Very high density block #1

. . . and so forth

3,999 Very high density block #20
4,000 Very high density block #20
4,001 High-density block #1
4,002 High-density block #2

. . . and so forth

49,160 Semirural block #160
49,161 Semirural block #161
49,162 Semirural block #162

Henry still wants to interview about 220 people and wants to select one person from each adjusted cluster unit. He uses simple random sample methods to select 220 of the 49,162 adjusted cluster units. He can then convert the cluster units back to city blocks. For example, if Henry randomly selected numbers 25 and 184, both are in very high density block #1, telling him to select two people from that block. If he randomly picked the number 49,161, he selects one person in semirural block #161. Henry now goes to each selected block, identifies all housing units in that block, and randomly selects among housing units. Of course, Henry may use within-household sampling after he selects a housing unit.

computers to select random digits and dial the phone automatically. This speeds the process, but a human must still listen and find out whether the number is a working residential one (see Expansion Box 3, Random Digit Dialing.)

The sampling element in RDD is the phone number, not the person or the household. Several families or individuals can share the same phone number, and in other situations, each person may have a separate phone number. This means that after a working residential phone is reached, a second stage of sampling, within household sampling, is necessary to select the person to be interviewed.

Example Box 6, (Example Sample, the 2006 General Social Survey) illustrates how the many sampling terms and ideas can be used together in a specific real-life situation.

EXPANSION BOX 3

Random-Digit Dialing (RDD)

During the past decade, participation in RDD surveys has declined. This is due to factors such as new call-screening technologies, heightened privacy concerns due to increased telemarketing calls, a proliferation of nonhousehold telephone numbers, and increased cell telephone users (most RDD samples include only landline numbers). When they compared a new technique, address-based sampling (ABS), to RDD for the U.S. adult population, Link et al. (2008) estimated that RDD sampling frames may be missing 15–19 percent of the population. Although the alternative was superior to RDD in some respects, ABS had other limitations including overrepresentation of English-speaking non-Hispanics and more educated persons than RDD. One issue in RDD sampling involves reaching someone by phone. A researcher might call a phone number dozens of times that is never answered. Does the nonanswer mean an eligible person is not answering or that the number is not really connected with a person? A study (Kennedy, Keeter, and Dimock, 2008) of this issue estimates that about half (47 percent) of unanswered calls in which there are six call-back attempts have an eligible person who is not being reached.

Decision Regarding Sample Size

New social researchers often ask, "How large does my sample have to be?" The best answer is, "It depends." It depends on population characteristics, the type of data analysis to be employed, and the degree of confidence in sample accuracy needed for research purposes. As noted, a large sample size alone does not guarantee a representative sample. A large sample without random sampling or with a poor sampling frame creates a less representative sample than a smaller one that has careful random sampling and an excellent sampling frame.

We can address the question of sample size in two ways. One method is to make assumptions about the population and use statistical equations about random sampling processes. The calculation of sample size by this method requires a statistical discussion that goes beyond the level of this text.[11] We must make assumptions about the degree of confidence (or number of errors) that is acceptable and the degree of variation in the population. In general, the more diverse a population, the more precise is the statistical analysis, the more variables will be examined simultaneously, and the greater confidence is required in sample accuracy (e.g., it makes a difference in critical health outcomes, huge financial loss, or the freedom or incarceration of innocent people), the larger the required sample size. The flip side is that samples from homogeneous populations with simple data analysis of one or a few variables that are used for low-risk decisions can be equally effective when they are smaller.

A second method to decide a sample size is a rule of thumb, a conventional or commonly accepted amount. We use rules of thumb because we rarely have the information required by the statistical estimation method. Also, these rules give sample sizes close to those of the statistical method. Rules of thumb are based on past experience with samples that have met the requirements of the statistical method.

A major principle of sample size is that the smaller the population, the larger the sampling ratio has to be for a sample that has a high probability of yielding the same results as the entire population. Larger populations permit smaller sampling ratios for equally good samples because as the population

EXAMPLE BOX 6

Example Sample, the 2006 General Social Survey

Sampling has many terms for the different types of samples. A complex sample illustrates how researchers use them. We can look at the 2006 sample for the best-known national U.S. survey in sociology, the General Social Survey (GSS). It has been conducted since 1972. Its sampling has been updated several times over the years based on the most sophisticated social science sampling techniques to produce a representative population within practical cost limits. The *population* consists of all resident adults (18 years of age or older) in the United States for the *universe* of all Americans. The *target population* consists of all English- or Spanish-speaking mentally competent adults who live in households but excludes people living in institutional settings. The researchers used a complex multistage area probability sample to the block or segment level. At the block level, they used *quota sampling* with quotas based on gender, age, and employment status. They selected equal numbers of men and women as well as persons over and under 35 years of age.

The sample design combined a *cluster sample* and a *stratified sample*. U.S. territory was divided into standard metropolitan statistical areas (SMSAs, a U.S. Census Bureau classification) and nonmetropolitan counties. The SMSAs and counties were stratified by region, age, and race before selection. Researchers adjusted clusters using *probability proportionate to size (PPS)* based on the number of housing units in each county or SMSA.

The sampling design had three basic stages. *Stage 1*: Randomly select a "primary sampling unit" (a U.S. census tract, a part of a SMSA, or a county) from among the stratified "primary sampling units." Researchers also classified units by whether there were stable mailing addresses in a geographic area or others. *Stage 2*: Randomly select smaller geographic units (e.g., a census tract, parts of a county), and *Stage 3*: Randomly select housing units on blocks or similar geographic units. As a final stage, researchers used the household as the sampling element and randomly selected households from the addresses in the block. After selecting an address, an interviewer contacted the household and chose an eligible respondent from it. The interviewer looked at a quota selection table for possible respondents and interviewed a type of respondent (e.g., second oldest) based on the table. Interviewers used computer-assisted personal interviewing (CAPI).

In the 2006 sample, researchers first identified 9,535 possible household addresses or locations. However, this number dropped to 7,987 after they eliminated vacant addresses and ones where no one who spoke either English or Spanish lived. After taking into account people who refused to participate, were too ill, were ineligible, or did not finish an interview (23.3%), the final sample included 4,510 persons (for details, see http://publicdata.norc.org:41000/gss/Documents/Codebook/A.pdf)

size grows, the returns in accuracy for sample size decrease.

In practical terms, this means for small populations (under 500), we need a large sampling ratio (about 30 percent) or 150 people, while for large populations (over 150,000), we can obtain equally good accuracy with a smaller sampling ratio (1 percent), and samples of about 1,500 can be equally accurate, all things being the same. Notice that the population of 150,000 is 30 times larger but the sample is just 10 times larger. Turning to very large populations (more than 10 million), we can achieve accuracy with tiny sampling ratios (0.025 percent), or samples of about 2,500. The size of the population ceases to be relevant once the sampling ratio is very small, and samples of about 2,500 are as accurate for populations of 200 million as for 10 million. These are approximate sizes, and practical limitations (e.g., cost) also play a role.

A related principle is that for small samples, a small increase in sample size produces a big gain in accuracy. Equal increases in sample size produce an increase in accuracy more for small than for large samples. For example, an increase in sample size from 50 to 100 reduces errors from 7.1 percent to 2.1 percent, but an increase from 1,000 to 2,000

TABLE 3 Sample Size of a Random Sample for Different Populations with a 99 Percent Confidence Level

POPULATION SIZE	SAMPLE SIZE	% POPULATION IN SAMPLE
200	171	85.5%
500	352	70.4%
1,000	543	54.3%
2,000	745	37.2%
5,000	960	19.2%
10,000	1,061	10.6%
20,000	1,121	5.6%
50,000	1,160	2.3%
100,000	1,173	1.2%

decreases errors from only 1.6 percent to 1.1 percent.[12] (See Table 3.)

Notice that our plans for data analysis influence the required sample size. If we want to analyze many small subgroups within the population, we need a larger sample. Let us say we want to see how elderly Black females living in cities compare with other subgroups (elderly males, females of other ages and races, and so forth). We will need a large sample because the subgroup is a small proportion (e.g., 10 percent) of the entire sample. A rule of thumb is to have about 50 cases for each subgroup we wish to analyze. If we want to analyze a group that is only 10 percent of our sample, then we will need a sample 10 times 50 (500 cases) in the sample for the subgroup analysis. You may ask how you would know that the subgroup of interest is only 10 percent of the sample until you gather sample data? This is a legitimate question. We often must use various other sources of information (e.g., past studies, official statistics about people in an area), then make an estimate, and then plan our sample size requirements from the estimate.

Making Inferences. The reason we draw probability samples is to make inferences from the sample to the population. In fact, a subfield of statistical data analysis is called *inferential statistics*. We directly observe data in the sample but are not interested in a sample alone. If we had a sample of 300 from 10,000 students on a college campus, we are less interested in the 300 students than in using information from them to infer to the population of 10,000 students. Thus, a gap exists between what we concretely have (variables measured in sample data) and what is of real interest (population parameters) (see Figure 4).

We can express the logic of measurement in terms of a gap between abstract constructs and concrete indicators. Measures of concrete, observable data are approximations for abstract constructs. We use the approximations to estimate what is of real interest (i.e., constructs and causal laws). Conceptualization and operationalization bridge the gap in measurement just as the use of sampling frames, the sampling process, and inference bridge the gap in sampling.

We can integrate the logic of sampling with the logic of measurement by directly observing measures of constructs and empirical relationships in samples (see Figure 4). We infer or generalize from what we observe empirically in samples to the abstract causal laws and parameters in the population. Likewise, there is an analogy between the logic of sampling and the logic of measurement for validity. In measurement, we want valid indicators of constructs: that is, concrete observable indicators that accurately represent unseen abstract constructs. In sampling, we want samples that have little sampling error: that is, concrete collections of cases that accurately represent unseen and abstract populations. A valid measure deviates little from the construct it represents. A good sample has little sampling error, and it permits estimates that deviate little from population parameters.

We want to reduce sampling errors. For equally good sampling frames and precise random selection processes, the sampling error is based on two factors: the sample size and the population diversity. Everything else being equal, the larger the sample size, the smaller the sampling error. Likewise, populations with a great deal of homogeneity will have smaller sampling errors. We can think of it this way: if we had a choice between

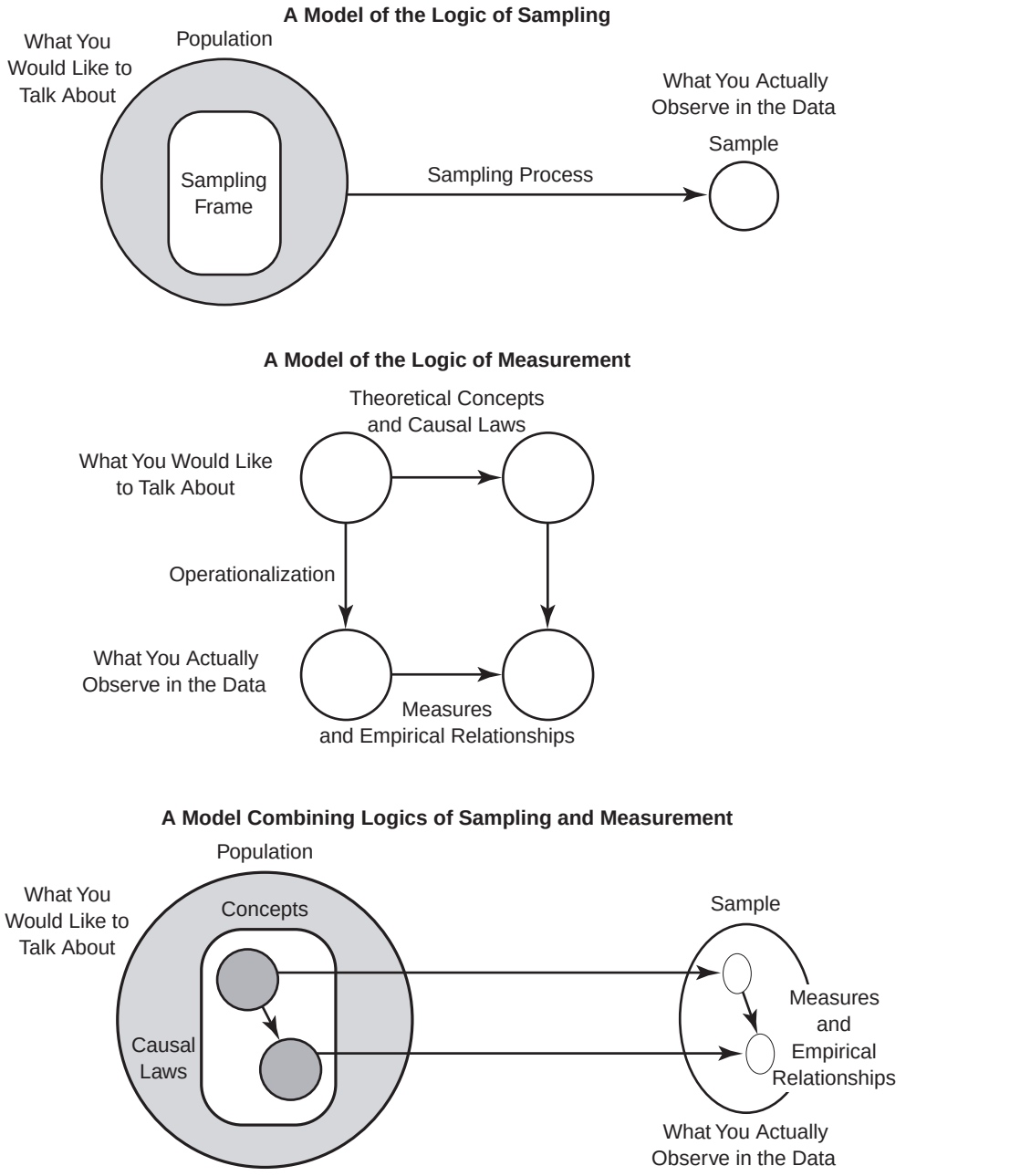


FIGURE 4 Model of the Logic of Sampling and of Measurement

sampling/picking 10 or 50 marbles out of a jar of 1000 red and white marbles to determine the number of red marbles, it would be better to pick 50. Likewise, if there are ten colors of marbles in a jar, we are less able to predict accurately the number of red marbles than if there were only two colors of marbles.

Sampling error is related to confidence intervals. If two samples are identical except one is much larger, the larger one will have a smaller sampling error and narrower confidence intervals. Likewise, if two samples are identical except that the cases in one are more similar to each other, the one with greater homogeneity will have a smaller sampling error and narrower confidence intervals. A narrow confidence interval means that we are able to estimate more precisely the population parameter for a given level of confidence.

Here is an example: You want to estimate the annual income of bricklayers. You have two samples. Sample 1 gives a confidence interval of $30,000 to $36,000 around the estimated population parameter of $33,000 for an 80 percent level of confidence. However, you want a 95 percent level of confidence. Now the range is $25,000 to $45,000. A sample that has a smaller sampling error (because it is much larger) might give the $30,000 to $36,000 range for a 95 percent confidence level.

Strategies When the Goal Differs from Creating a Representative Sample

In qualitative research, the purpose of research may not require having a representative sample from a huge number of cases. Instead, a nonprobability sample often better fits the purposes of a study. In nonprobability samples, you do not have to determine the sample size in advance and have limited knowledge about the larger group or population from which the sample is taken. Unlike a probability sample that required a preplanned approach based on mathematical theory, nonprobability sampling often gradually selects cases with the specific content of a case determining whether it is chosen. Table 4 shows a variety of nonprobability sampling techniques.

TABLE 4 Types of Nonprobability Samples

TYPE OF SAMPLE	PRINCIPLE
Convenience	Get any cases in any manner that is convenient.
Quota	Get a preset number of cases in each of several predetermined categories that will reflect the diversity of the population, using haphazard methods.
Purposive	Get all possible cases that fit particular criteria, using various methods.
Snowball	Get cases using referrals from one or a few cases, then referrals from those cases, and so forth.
Deviant case	Get cases that substantially differ from the dominant pattern (a special type of purposive sample).
Sequential	Get cases until there is no additional information or new characteristics (often used with other sampling methods).
Theoretical	Get cases that will help reveal features that are theoretically important about a particular setting/topic.
Adaptive	Get cases based on multiple stages, such as snowball followed by purposive. This sample is used for hidden populations.

Purposive or Judgmental Sampling

Purposive sampling (also known as *judgmental sampling*) is a valuable sampling type for special situations. It is used in exploratory research or in field research.[12] It uses the judgment of an expert in

Purposive sampling A nonrandom sample in which the researcher uses a wide range of methods to locate all possible cases of a highly specific and difficult-to-reach population.

selecting cases, or it selects cases with a specific purpose in mind. It is inappropriate if the goal is to have a representative sample or to pick the "average" or the "typical" case. In purposive sampling, cases selected rarely represent the entire population.

Purposive sampling is appropriate to select unique cases that are especially informative. For example, we want to use content analysis to study magazines to find cultural themes. We can use three specific popular women's magazines to study because they are trend setting. In the study *Promises I Can Keep* that opened this chapter, the researchers selected eight neighborhoods using purposive sampling. We often use purposive sampling to select members of a difficult-to-reach, specialized population, such as prostitutes. It is impossible to list all prostitutes and sample randomly from the list. Instead, to locate persons who are prostitutes, a researcher will use local knowledge (e.g., locations where prostitutes solicit, social groups with whom prostitutes associate) and local experts (e.g., police who work on vice units, other prostitutes) to locate possible prostitutes for inclusion in the research project. A researcher will use many different methods to identify the cases because the goal is to locate as many cases as possible.

We also use purposive sampling to identify particular types of cases for in-depth investigation to gain a deeper understanding of types (see Example Box 7, Purposive Sampling).

Snowball Sampling

We are often interested in an interconnected network of people or organizations.[13] The network could be scientists around the world investigating the same problem, the elites of a medium-size city, members of an organized crime family, persons who sit on the boards of directors of major banks and corporations, or people on a college campus who

EXAMPLE BOX 7

Purposive Sampling

In her study *Inside Organized Racism,* Kathleen Blee (2002) used purposive sampling to study women who belong to racist hate organizations. The purpose of her study was to learn why and how women became actively involved in racist hate organizations (e.g., neo-Nazi, Ku Klux Klan). She wanted "to create a broadly based, national sample of women racist group members" (p. 198). A probability sample was not possible because no list of all organizations exists, and the organizations keep membership lists secret.

Blee avoided using snowball sampling because she wanted to interview women who were not connected to one another. To sample women for the study, she began by studying the communication (videotapes, books, newsletters, magazines, flyers, Web sites) "distributed by every self-proclaimed racist, anti-Semitic, white supremacist, Christian Identity, neo-Nazi, white power skinhead, and white separatist organization in the United States for a one-year period" (p. 198). She also obtained lists from antiracist organizations that monitor racist groups and examined the archives at the libraries of Tulane University and the University of Kansas for right-wing extremism. She identified more than one hundred active organizations. From these, she found those that had women members or activists and narrowed the list to thirty racist organizations. She then tried to locate women who belonged to organizations that differed in ideological emphasis and organizational form in fifteen different states in four major regions of the United States.

In a type of cluster sampling, she first located organizations and then women active in them. To find women to interview, she used personal contacts and referrals from informed persons: "parole officers, correctional officials, newspaper reporters and journalists, other racist activists and former activists, federal and state task forces on gangs, attorneys, and other researchers" (p. 200). She eventually located thirty-four women aged 16 to 90 years of age and conducted two 6-hour life history interviews with each.

Source: Excerpt from page 198 of *Inside Organized Racism*: *Women in the Hate Movement,* by Kathleen M. Blee. © 2002 by the Regents of the University of California. Published by the University of California Press.

have had sexual relations with each other. The crucial feature is that each person or unit is connected with another through a direct or indirect linkage. This does not mean that each person directly knows, interacts with, or is influenced by every other person in the network. Rather, taken as a whole, with direct and indirect links, most people are within an interconnected web of linkages.

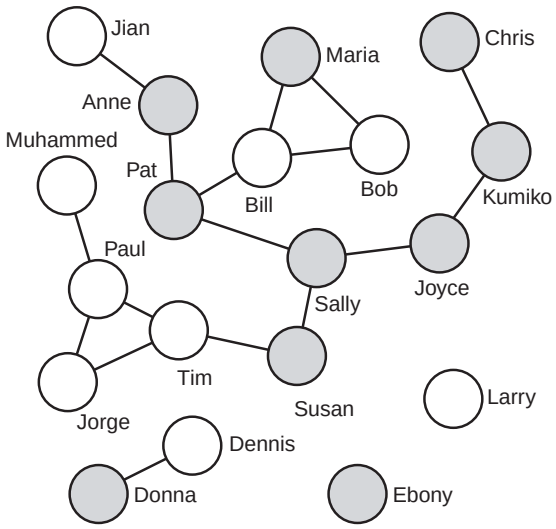
For example, Sally and Tim do not know each other directly, but each has a good friend, Susan, so they have an indirect connection. All three are part of the same friendship network. Researchers represent such a network by drawing a *sociogram*, a diagram of circles connected with lines. The circles represent each person or case, and the lines represent friendship or other linkages (see Figure 5).

Snowball sampling (also called *network, chain referral, reputational,* and *respondent-driven sampling*) is a method for sampling (or selecting) the cases in a network. The method uses an analogy to a snowball, which begins small but becomes larger as we roll it on wet snow and it picks up additional snow. Snowball sampling is a multistage technique. It begins with one or a few people or cases and spreads out based on links to the initial cases.

For example, we want to study friendship networks among the teenagers in our community. We might start with three teenagers who do not know each other. We ask each teen to name four close friends. Next we go to each set of four friends and ask each person to name four close friends. This continues to the next round of four people and repeats again. Before long, a large number of people have been identified. Each person in the sample is directly or indirectly tied to the original teenagers, and several people may have named the same person. The process stops, either because no new names are given, indicating a closed network, or because the network is so large that it is at the limit of what can be studied. The sample includes those named by at least one other person in the network as being a close friend.

Deviant Case Sampling

We use **deviant case sampling** (also called *extreme case sampling*) when we are interested in cases that



Note: Shading indicates various skin tones.

FIGURE 5 Sociogram of Friendship Relations

differ from the dominant pattern, mainstream, or predominant characteristics of other cases. Similar to purposive sampling, we use a variety of techniques to locate cases with specific characteristics. The goal is to locate a collection of unusual, different, or peculiar cases that are not representative of the whole. We select cases because they are unusual. We can sometimes learn more about social life by considering cases that fall outside the general pattern or including what is beyond the main flow of events.

For example, we want to study high school dropouts. Let us say that previous research suggested that a majority of dropouts come from low-income,

Snowball sampling A nonrandom sample in which the researcher begins with one case and then, based on information about interrelationships from that case, identifies other cases and repeats the process again and again.

Deviant case sampling A nonrandom sample, especially used by qualitative researchers, in which a researcher selects unusual or nonconforming cases purposely as a way to provide increased insight into social processes or a setting.

single-parent families and tend to be racial minorities. The family environment is one in which parents and/or siblings have low education or are themselves dropouts. In addition, many dropouts engage in illegal behavior. We might seek dropouts who are members of the majority racial group, who have no record of illegal activities, and who are from stable two-parent, upper-middle-income families. By looking at atypical dropouts we might learn more about the reasons for dropping out.

Sequential Sampling

Sequential sampling is also similar to purposive sampling. We use purposive sampling to try to locate as many relevant cases as possible. Sequential sampling differs because we continue to gather cases until the amount of new information ends or a certain diversity of cases is reached. The principle is to gather cases until we reach a saturation point. In economic terms, information is gathered until the marginal utility, or incremental benefit for additional cases, levels off or drops significantly. It requires that we continuously evaluate all collected cases. For example, we locate and plan in-depth interviews of sixty widows over 70 years of age who have been living without a spouse for 10 or more years. Depending on our purposes, getting an additional twenty widows whose life experiences, social backgrounds, and worldviews differ little from the first sixty may be unnecessary.

Sequential sampling A nonrandom sample in which a researcher tries to find as many relevant cases as possible until time, financial resources, or his or her energy is exhausted or until there is no new information or diversity from the cases.

Theoretical sampling A nonrandom sample in which the researcher selects specific times, locations, or events to observe in order to develop a social theory or evaluate theoretical ideas.

Hidden population A population of people who engage in clandestine, socially disapproved of, or concealed activities and who are difficult to locate and study.

Adaptive sampling A nonprobability sampling technique used for hidden populations in which several approaches to identify and recruit, including a snowball or referral method, may be used.

Theoretical Sampling

In **theoretical sampling**, what we sample (e.g., people, situations, events, time periods) comes from grounded theory. A growing theoretical interest guides the selection of sample cases. The researcher selects cases based on new insights that the sample could provide. For example, a field researcher could be observing a site and a group of people during weekdays. Theoretically, the researcher may question whether the people act the same at other times or aspects of the site change. He or she could then sample other time periods (e.g., nights and weekends) to have a fuller picture and learn whether important conditions are the same.

Adaptive Sampling and Hidden Populations

In contrast to sampling the general population or visible and accessible people, sampling **hidden populations** (i.e., people who engage in clandestine or concealed activities) is a recurrent issue in the studies of deviant or stigmatized behavior (such as victims of sexual violence, illegal drug users). This method illustrates the creative application of sampling principles, mixing qualitative and quantitative styles of research and combining probability with nonprobability techniques.

Adaptive sampling is a design that adjusts based on early observations.[15] For example, we ask illegal drug users to refer other drug users as in snowball sampling. However, we adjust the way that we trace through the network based on our research topic. We might identify a geographic area, divide it into sections randomly, and then select participants in that area through strategies such as random-digit dialing or by posting recruitment fliers. Once we identify members of the targeted hidden population, we use them in a snowball technique to find others. AIDS researchers or studies of illegal drug users that have sampled "hidden populations" are instructive, often relying on modified snowball techniques. (See Example Box 8, Hidden Populations).

EXAMPLE BOX 8

Hidden Populations

Three studies of hidden populations illustrate the difficulties of sampling. Martin and Dean (1993) sampled gay men from New York City. The men had to live in the city, be over age 18, not be diagnosed as having AIDS, and engage in sex with other men. The authors began with a purposive sample using five diverse sources to recruit 291 respondents. They first contacted 150 New York City organizations with predominately homosexual or bisexual members. They next screened these to 90 organizations that had men appropriate for the study. From the 90, the researchers drew a stratified random sample of 52 organizations by membership size. They randomly selected five members from each of the organizations. Reports of Martin and Dean's study appeared in local news sources. This brought calls from forty-one unsolicited volunteers. They also found thirty-two men as referrals from respondents who had participated in a small pilot study, seventy-two men from an annual New York City Gay Pride Parade, and fifteen eligible men whom they contacted at a New York City clinic and asked to participate. They next used snowball sampling by asking each of the 291 men to give a recruitment packet to three gay male friends. Each friend who agreed to participate was also asked to give packets to three friends. This continued until it had gone five levels out from the initial 291 men. Eventually, Dean recruited 746 men into the study. The researchers checked their sample against two random samples of gay men in San Francisco, a random-digit dialing sample of 500, and a cluster sample of 823 using San Francisco census tracts. Their sample paralleled those from San Francisco on race, age, and the percent being "out of the closet."

Heckathorn (1997, 2002) studied active drug injectors in two small Connecticut cities and the surrounding area. As of July 1996, medical personnel had diagnosed 390 AIDS cases in the towns; about half of the cases involved drug injection. The sampling was purposive in that each sampled element had to meet certain criteria. Heckathorn also used a modified snowball sampling with a "dual reward system." He gave each person who completed an interview a monetary reward and a second monetary reward for recruiting a new respondent. The first person was asked not to identify the new person to the researcher, a practice sometimes referred to as *masking* (i.e., protecting friends). This avoids the "snitching" issue and "war on drugs" stigma, especially strong in the U.S. context. This modified snowball sampling is like sequential sampling in that after a period of time, fewer and fewer new recruits are found until the researcher comes to saturation or an equilibrium.

Wang et al. (2006) used a respondent-driven sampling method to recruit 249 illicit drug users in three rural Ohio counties to examine substance abuse and health care needs. To be eligible for the sample, participants had to be over 18 years of age, not be in drug abuse treatment, and not have used cocaine or methamphetamines in the past month. After locating an eligible participant, the researchers paid him or her $50 dollars to participate. The participant could earn an additional $10 by recruiting eligible peers. In a snowball process, each subsequent participant was also asked to make referrals. The authors identified nineteen people to start. Only a little more than half (eleven of the nineteen) referred peers for the study who were eligible and participated. Over roughly 18 months, the researchers were able to identify 249 participants for their study. They compared the study sample with characteristics of estimates of the illegal drug–using population and found that the racial composition of the originally identified participants (White) led to overrepresentation of that racial category. Otherwise, it appeared that the method was able to draw a reasonable sample of the hidden population.

CONCLUSION

This chapter discussed probability and nonprobability sampling (see Summary Review Box 1, Types of Samples). A key point is that a sampling strategy should match in a specific study's purpose. In general, probability sampling is preferred for a representative sample; it allows for using statistical tests in data analysis. In addition to simple random sampling, the chapter referred to other probability samples: systematic, stratified, RDD, and cluster sampling. The

SUMMARY REVIEW BOX 1

Types of Samples

EIGHT TYPES OF NONPROBABILITY SAMPLES

Type of Sample	*Principle*
Adaptive	Get a few cases using knowledge of likely locations of a hidden population, use random techniques or recruit, and then use a snowball sample to expand from a few cases.
Convenience	Get any cases in any manner that is convenient.
Deviant case	Get cases that substantially differ from the dominant pattern (a special type of purposive sample).
Purposive	Get all possible cases that fit particular criteria using various methods.
Quota	Using haphazard methods, get a preset number of cases in each of several predetermined categories that will reflect the diversity of the population.
Sequential	Get cases until there is no additional information or new characteristics (often used with other sampling methods).
Snowball	Get cases using referrals from one or a few cases, then referrals from those cases, and so forth.
Theoretical	Get cases that will help reveal features that are theoretically important about a particular setting/topic.

FOUR TYPES OF PROBABILITY SAMPLES

Type of Sample	*Technique*
Cluster	Create a sampling frame for large cluster units, draw a random sample of the cluster units, create a sampling frame for cases within each selected cluster unit, then draw a random sample of cases, and so forth.
Simple random	Create a sampling frame for all cases and then select cases using a purely random process (e.g., random-number table or computer program).
Stratified	Create a sampling frame for each of several categories of cases, draw a random sample from each category, and then combine the several samples.
Systematic	Create a sampling frame, calculate the sampling interval 1/k, choose a random starting place, and then take every 1/k case.

discussions of sampling error, the central limit theorem, and sample size indicated that probability sampling produces most accurate sampling when the goal is creating a representative sample.

The chapter also discussed several types of nonprobability samples: convenience, deviant case quota, sequential, snowball, and theoretical. Except for convenience, these types are best suited for studies in which the purpose is other than creating a sample that is highly representative of a population.

Before you move on, it may be useful to restate a fundamental principle of all social research: Do not compartmentalize the steps of the research process; rather, learn to see the interconnections among the steps. Research design, measurement, sampling, and specific research techniques are interdependent. In practice, we need to think about data collection as we design research and develop measures. Likewise, sampling issues influence research design, measurement, and data collection strategies. As you will see, good social research depends on simultaneously controlling quality at several different steps: research design, conceptualization, measurement, sampling, and data collection and handling. Making serious errors at any one stage could make an entire research project worthless.

KEY TERMS

adaptive sampling
central limit theorem
cluster sampling
confidence intervals
convenience sampling
deviant case sampling
hidden populations
parameter
population
probability proportionate to size (PPS)
purposive sampling
quota sampling
random-digit dialing (RDD)
random-number table
random sample
sample
sampling distribution
sampling element
sampling error
sampling frame
sampling interval
sampling ratio
sequential sampling
simple random sample
snowball sampling
statistic
stratified sampling
systematic sampling
target population
theoretical sampling

REVIEW QUESTIONS

1. When is purposive sampling used?
2. When is the snowball sampling technique appropriate?
3. What is a sampling frame and why is it important?
4. Which sampling method is best when the population has several groups and a researcher wants to ensure that each group is in the sample?
5. How can researchers determine a sampling interval from a sampling ratio?
6. When should a researcher consider using probability proportionate to size?
7. What is the population in random-digit dialing? Does this type avoid sampling frame problems? Explain.
8. How do researchers decide how large a sample to use?
9. How are the logic of sampling and the logic of measurement related?
10. When is random-digit dialing used, and what are its advantages and disadvantages?

NOTES

1. See Stern (1979:77–81) and Beck (1983) on biased samples.
2. Babbie (1998:196), Kalton (1983:91–93), and Sudman (1976a:191–200) discuss quota sampling.
3. For a discussion of the *Literary Digest* sampling error, see Babbie (1998:192–194), Dillman (1978:9–10), Frey (1983:18–19), and Singleton and colleagues (1988: 132–133).
4. See Traugott (1987) on the importance of persistence in reaching sampled respondents for a representative sample. Also see Kalton (1983:63–69) on the importance of nonresponse.
5. Only one name appears in both. The stratified sample has six males and four females; the simple random sample has five males and five females. (Complete the lower block of numbers and then begin at the far right of the top block.)
6. Stratified sampling techniques are discussed in more detail in Frankel (1983:37–46), Kalton (1983:19–28), Mendenhall and associates (1971:53–88), Sudman (1976a:107–130), and Williams (1978:162–175).
7. Cluster sampling is discussed in Frankel (1983: 47–57), Kalton (1983:28–38), Kish (1965), Mendenhall and associates (1971:121–141, 171–183), Sudman (1976a: 69–84), and Williams (1978:144–161).
8. For a discussion, see Frankel (1983:57–62), Kalton (1983:38–47), Sudman (1976a:131–170), and Williams (1978:239–241).
9. Czaja and associates (1982) and Groves and Kahn (1979:32–36) discuss within-household sampling.
10. For more on random-digit dialing issues, see Dillman (1978:238–242), Frey (1983:69–77), Glasser and Metzger (1972), Groves and Kahn (1979:20–21, 45–63), Kalton (1983:86–90), and Waksberg (1978). Kviz (1984) reported that telephone directories can produce relatively accurate sampling frames in rural areas, at least for mail questionnaire surveys. Also see Keeter (1995).
11. See Grosof and Sardy (1985:181–185), Kalton (1983: 82–90), Kraemer and Thiemann (1987), Sudman (1976a:85–105), and Williams (1978: 211–227) for a technical discussion of selecting a sample size.
12. For further discussion on purposive sampling, see Babbie (1998:195), Grosof and Sardy (1985:172–173), and Singleton and associates (1988:153–154, 306). Bailey (1987:94–95) describes "dimensional" sampling, which is a variation of purposive sampling.
13. Snowball sampling is discussed in Babbie (1998:194–196), Bailey (1987:97), and Sudman (1976a:210–211). For discussions of sociometry and sociograms, also see Bailey (1987:366–367), Dooley (1984:86–87), Kidder and Judd (1986:240–241), Lindzey and Byrne (1968:452–525), and Singleton and associates (1988:372–373). Network sampling issues are discussed in Galaskiewicz (1985), Granovetter (1976), and Hoffmann-Lange (1987).
14. On adaptive sampling, see Martsolf et al. (2006), Thompson and Geber (1996), Thompson (2002), and Thompson and Collins (2002).