Qualitative and Quantitative Measurement

From Chapter 7 of *Social Research Methods: Qualitative and Quantitative Approaches*, 7/e. W. Lawrence Neuman.
Copyright © 2011 by Pearson Education. Published by Allyn & Bacon. All rights reserved.

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 When the National Research Council examined the measure in 1995, members declared it outdated and said it should not be retained.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.

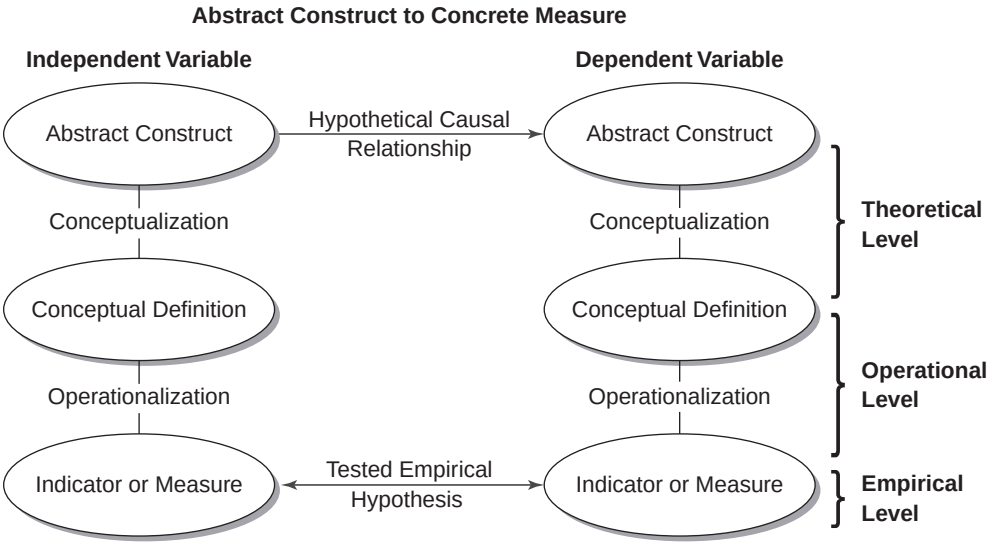

FIGURE 1 Conceptualization and Operationalization

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

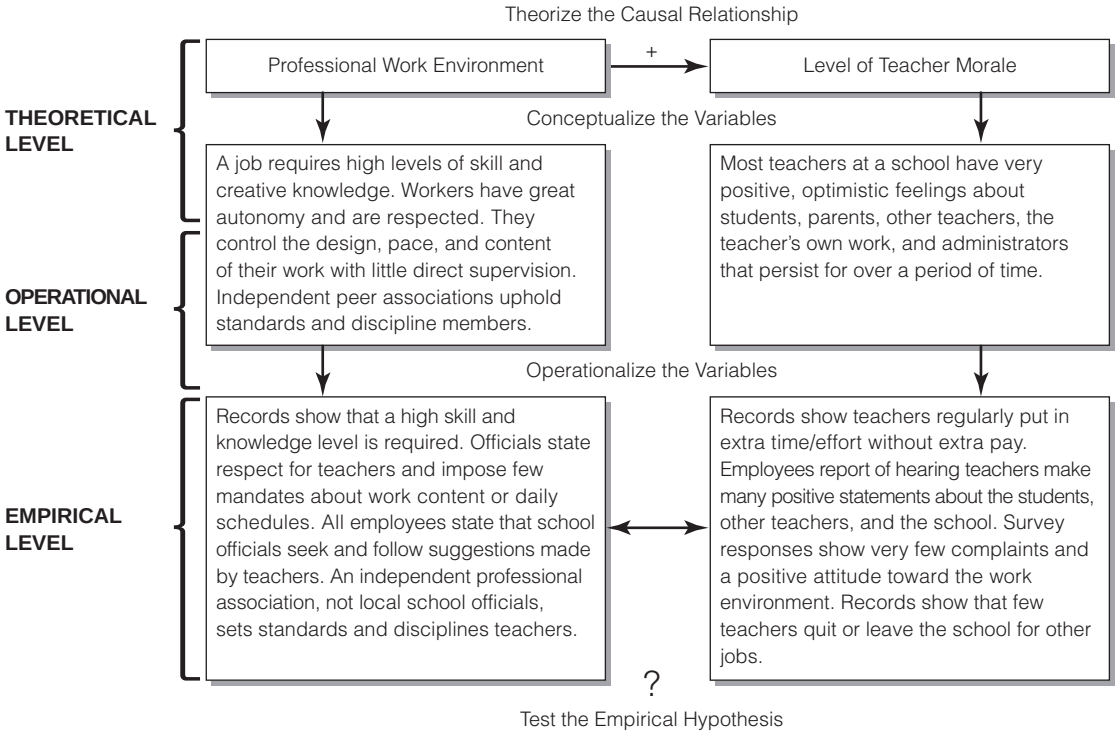

FIGURE 2 Example of the Deductive Measurement Process for the Hypothesis: A Professional Work Environment Increases the Level of Teacher Morale

(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

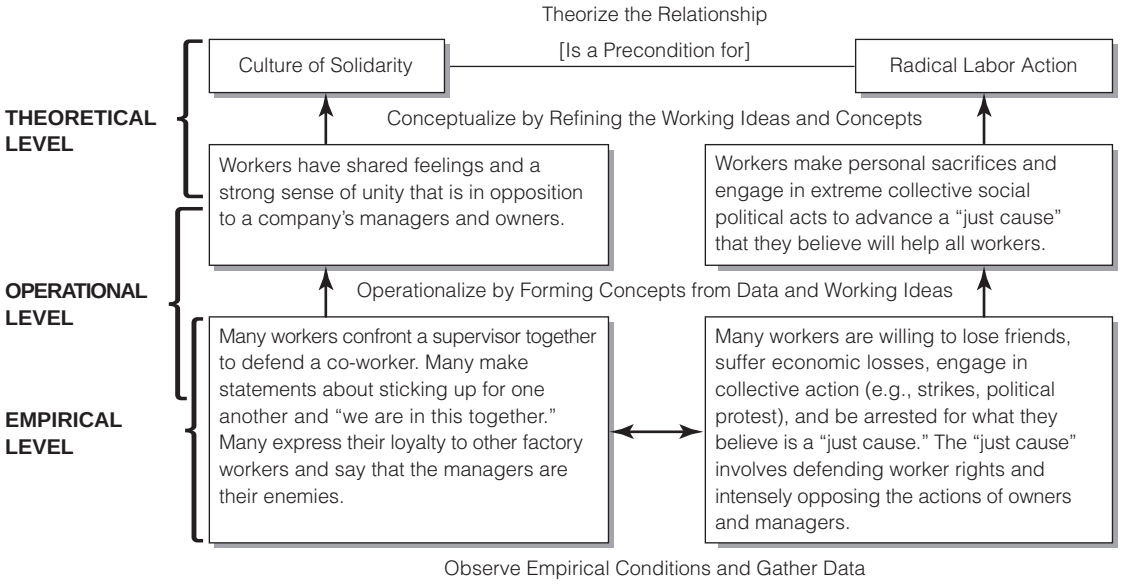

FIGURE 3 Example of the Inductive Measurement Process for the Proposition: Radical Labor Action Is Likely to Occur Where a Culture of Solidarity Has Been Created

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

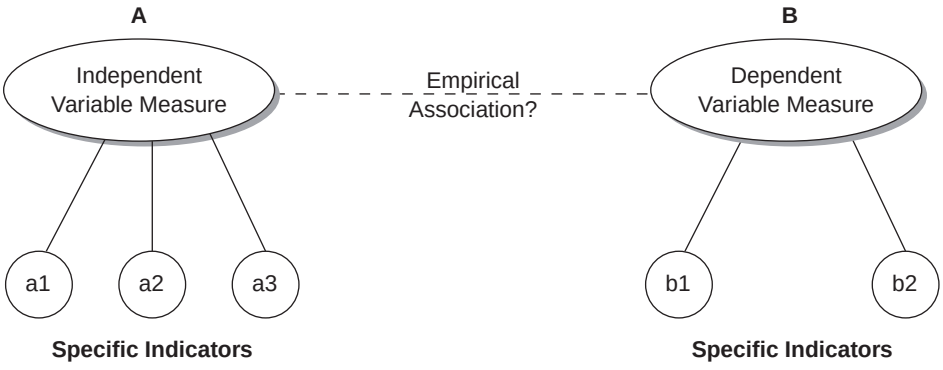

FIGURE 4 Measurement Using Multiple Indicators

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.

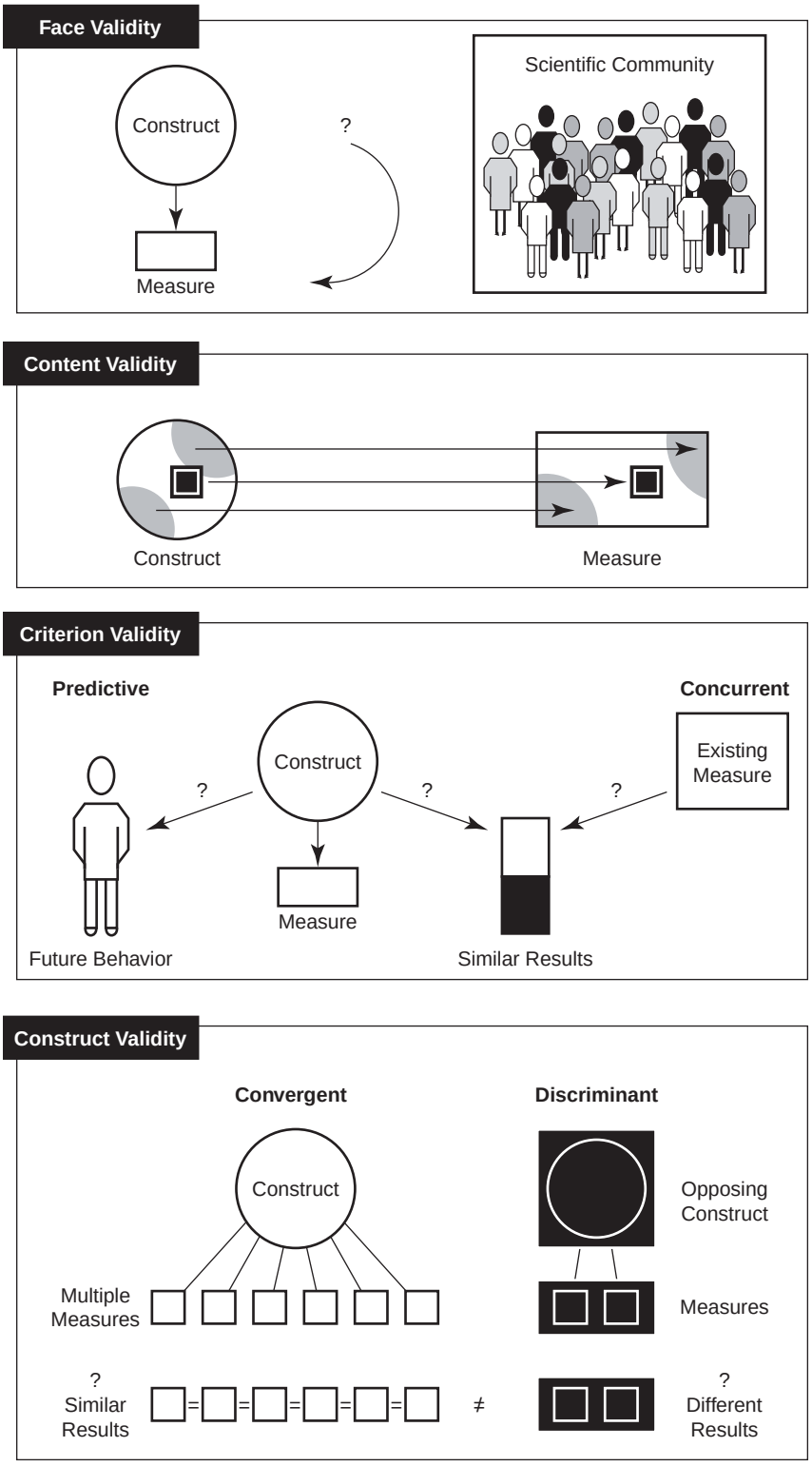

FIGURE 5 Types of Validity

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

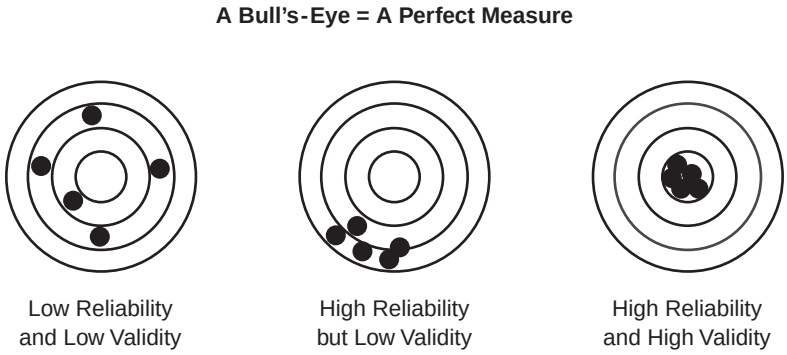

FIGURE 6 Illustration of Relationship between Reliability and Validity

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

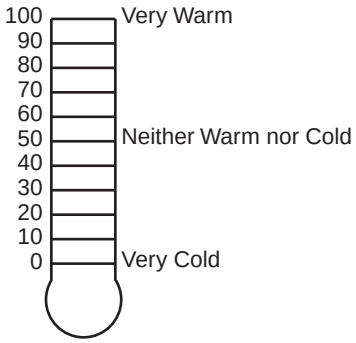

FIGURE 7 "Feeling Thermometer" Graphic Rating Scale

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.