CHAPTER 5

Conceptualization, Operationalization, and Measurement

CHAPTER OVERVIEW

The interrelated steps of conceptualization, operationalization, and measurement allow researchers to turn a general idea for a research topic into useful and valid measurements in the real world. An essential part of this process involves transforming the relatively vague terms of ordinary language into precise objects of study with well-defined and measurable meanings.

Introduction

This chapter and the next deal with how researchers move from a general idea about what they want to study to effective and well-defined measurements in the real world. This chapter discusses the interrelated processes of conceptualization, operationalization, and measurement. Chapter 6 builds on this foundation to discuss types of measurements that are more complex.

Consider a notion such as "satisfaction with college." I'm sure you know some people who are very satisfied, some who are very dissatisfied, and many who are between those extremes. Moreover, you can probably place yourself somewhere along that satisfaction spectrum. While this probably makes sense to you as a general matter, how would you go about measuring how different students were in this regard, so you could place them along that spectrum?

There are some comments students make in conversations (such as "This place sucks") that would tip you off as to where they stood. Or, in a more active effort, you can probably think of questions you might ask students to learn about their satisfaction (such as "How satisfied are you with . . . ?"). Perhaps there are certain behaviors (class attendance, use of campus facilities, setting the dean's office on fire) that would suggest different levels of satisfaction. As you think about ways of measuring satisfaction with college, you are engaging in the subject matter of this chapter.

We begin by confronting the hidden concern people sometimes have about whether it's truly possible to measure the stuff of life: love, hate, prejudice, religiosity, radicalism, alienation. The answer is yes, but it will take a few pages to see how. Once we establish that researchers can measure anything that exists, we'll turn to the steps involved in doing just that.

Measuring Anything That Exists

Earlier in this book, I said that one of the two pillars of science is observation. Because this word can suggest a casual, passive activity, scientists often use the term *measurement* instead, meaning careful, deliberate observations of the real world for the purpose of describing objects and events in terms of the attributes composing a variable.

You may have some reservations about the ability of science to measure the really important aspects of human social existence. If you've read research reports dealing with something like liberalism or religion or prejudice, you may have been dissatisfied with the way the researchers measured whatever they were studying. You may have felt that they were too superficial, that they missed the aspects that really matter most. Maybe they measured religiosity as the number of times a person went to religious services, or maybe they measured liberalism by how people voted in a single election. Your dissatisfaction would surely have increased if you had found yourself being misclassified by the measurement system.

Your feeling of dissatisfaction reflects an important fact about social research: Most of the variables we want to study don't actually exist in the way that, say, rocks exist. Indeed, they are made up. Moreover, they seldom have a single, unambiguous meaning.

To see what I mean, suppose we want to study *political party affiliation.* To measure this variable, we might consult the list of registered voters to note whether the people we were studying were registered as Democrats or Republicans and take that as a measure of their party affiliation. But we could also simply ask someone what party they identify with and take their response as our measure. Notice that these two different measurement possibilities reflect somewhat different definitions of *political party affiliation.* They might even produce different results: Someone may have registered as a Democrat years ago but gravitated more and more toward a Republican philosophy over time. Or someone who is registered with neither political party may, when asked, say she is affiliated with the one she feels the most kinship with.

Similar points apply to *religious affiliation.* Sometimes this variable refers to official membership in a particular church, temple, mosque, and so forth; other times it simply means whatever religion, if any, you identify yourself with.

Perhaps to you it means something else, such as attendance at religious services.

The truth is that neither *party affiliation* nor *religious affiliation* has any real meaning, if by "real" we mean corresponding to some objective aspect of reality. These variables do not exist in nature. They are merely terms we've made up and assigned specific meanings to for some purpose, such as doing social research.

But, you might object, *political affiliation* and *religious affiliation*—and a host of other things social researchers are interested in, such as prejudice or compassion—have some reality. After all, researchers make statements about them, such as "In Happytown, 55 percent of the adults affiliate with the Republican Party, and 45 percent of them are Episcopalians. Overall, people in Happytown are low in prejudice and high in compassion." Even ordinary people, not just social researchers, have been known to make statements like that. If these things don't exist in reality, what is it that we're measuring and talking about?

What indeed? Let's take a closer look by considering a variable of interest to many social researchers (and many other people as well)—*prejudice*.

Conceptions, Concepts, and Reality

As we wander down the road of life, we observe a lot of things and know they are real through our observations, and we hear reports from other people that seem real. For example:

- We personally hear people say nasty things about minority groups.
- We hear people say that women are inferior to men.
- We read that women and minorities earn less for the same work.
- We learned about "ethnic cleansing" and wars in which one ethnic group tries to eradicate another.

With additional experience, we notice something more. A lot of the people who call African Americans ugly names also seem to want women to "stay in their place." They are also likely to think minorities are inferior to the majority and that women are inferior to men. These several tendencies often appear together in the same people and also have something in common. At some point, someone had a bright idea: "Let's use the word *prejudiced* as a shorthand notation for people like that. We can use the term even if they don't do all those things—as long as they're pretty much like that."

Being basically agreeable and interested in efficiency, we went along with the system. That's where "prejudice" came from. We never observed it. We just agreed to use it as a shortcut, a name that represents a collection of apparently related phenomena that we've each observed in the course of life. In short, we made it up.

Here's another clue that prejudice isn't something that exists apart from our rough agreement to use the term in a certain way. Each of us develops our own mental image of what the set of real phenomena we've observed represents in general and what these phenomena have in common. When I say the word *prejudice,* it evokes a mental image in your mind, just as it evokes one in mine. It's as though file drawers in our minds contained thousands of sheets of paper, with each sheet of paper labeled in the upper right-hand corner. A sheet of paper in each of our minds has the term *prejudice* on it. On your sheet are all the things you've been told about prejudice and everything you've observed that seems to be an example of it. My sheet has what I've been told about it plus all the things I've observed that seem examples of it—and mine isn't the same as yours.

The technical term for those mental images, those sheets of paper in our mental file drawers, is *conception.* That is, I have a conception of prejudice, and so do you. We can't communicate these mental images directly, so we use the terms written in the upper right-hand corner of our own mental sheets of paper as a way of communicating about our conceptions and the things we observe that are related to those conceptions. These terms make it possible for us to communicate and eventually agree on what we specifically mean by those terms. In social research, the process of coming to an agreement about what terms mean is **conceptualization**, and the result

conceptualization The mental process whereby fuzzy and imprecise notions (concepts) are made more specific and precise. So you want to study prejudice. What do you mean by "prejudice"? Are there different kinds of prejudice? What are they?

Research in Real Life

Gender and Race in City Streets

In the early 1970s, Elijah Anderson spent three years observing life in a black, working-class neighborhood in South Chicago, focusing on Jelly's, a combination bar and liquor store. While some people still believe that impoverished neighborhoods in the inner city are socially chaotic and disorganized, Anderson's study and others like it have clearly demonstrated a definite social structure there that guides the behavior of its participants. Much of his interest centered on systems of social status and how the 55 or so regulars at Jelly's worked those systems to establish themselves among their peers.

In the second edition (2003) of this classic study of urban life, Elijah Anderson returned to Jelly's and the surrounding neighborhood. There he found several changes, largely due to the outsourcing of manufacturing jobs overseas that has brought economic and mental depression to many of the residents. These changes, in turn, had also altered the nature of social organization.

For a research methods student, the book offers many insights into the process of establishing rapport with people being observed in their natural surroundings. Further, Anderson offers excellent examples of how concepts are established in qualitative research.

is called a *concept.* See the Research in Real Life box, "Gender and Race in City Streets," for a glimpse at a project that reveals a lot about conceptualization.

Perhaps you've heard some reference to the many words Eskimos have for *snow,* as an example of how environment can shape language. Here's an exercise you might enjoy when you're ready to take a break from reading. Search the web for "Eskimo words for snow." You may be surprised by what you find. You're likely to discover wide disagreement on the number of, say, Inuit, words—ranging from 1 to 400. Several sources, moreover, will suggest that if the Inuit have several words for *snow,* so does English. Cecil Adams, for example, lists "snow, slush, sleet, hail, powder, hard pack, blizzard, flurries, flake, dusting, crust, avalanche, drift, frost, and iceberg," on his website, Straight Dope. This illustrates the ambiguities in the field with regard to the concepts and words that we use in everyday communications and that also serve as the grounding for social research.

Let's take another example of a conception. Suppose that I'm going to meet someone named Pat, whom you already know. I ask you what Pat is like. Now suppose that you've seen Pat help lost children find their parents and put a tiny bird back in its nest. Pat got you to take turkeys to poor families on Thanksgiving and to visit a children's hospital on Christmas. You've seen Pat weep through a movie about a mother overcoming adversities to save and protect her child. As you search through your mental files, you may find all or most of those phenomena recorded on a single sheet labeled "compassionate." You look over the other entries on the page, and you find they seem to provide an accurate description of Pat. So you say, "Pat is compassionate."

Now I leaf through my own mental file drawer until I find a sheet marked "compassionate." I then look over the things written on my sheet, and I say, "Oh, that's nice." I now feel I know what Pat is like, but my expectations reflect the entries on *my* file sheet, not yours. Later, when I meet Pat, I happen to find that my own experiences correspond to the entries I have on my "compassionate" file sheet, and I say that you sure were right.

But suppose my observations of Pat contradict the things I have on my file sheet. I tell you that I don't think Pat is very compassionate, and we begin to compare notes.

You say, "I once saw Pat weep through a movie about a mother overcoming adversity to save and protect her child." I look at my "compassionate sheet" and can't find anything like that. Looking elsewhere in my file, I locate that sort of phenomenon on a sheet labeled "sentimental." I retort, "That's not compassion. That's just sentimentality."

To further strengthen my case, I tell you that I saw Pat refuse to give money to an organization dedicated to saving whales from extinction. "That represents a lack of compassion," I argue. You search through your files and find saving the whales on two sheets—"environmental activism" and "cross-species dating"—and you say so. Eventually, we set about comparing the entries we have on our respective sheets labeled

"compassionate." We then discover that many of our mental images corresponding to that term differ.

In the big picture, language and communication work only to the extent that you and I have considerable overlap in the kinds of entries we have on our corresponding mental file sheets. The similarities we have on those sheets represent the agreements existing in our society. As we grow up, we're told approximately the same thing when we're first introduced to a particular term, though our nationality, gender, race, ethnicity, region, language, or other cultural factors may shade our understanding of concepts.

Dictionaries formalize the agreements our society has about such terms. Each of us, then, shapes his or her mental images to correspond with such agreements. But because all of us have different experiences and observations, no two people end up with exactly the same set of entries on any sheet in their file systems. If we want to measure "prejudice" or "compassion," we must first stipulate what, exactly, counts as prejudice or compassion for our purposes.

Returning to the assertion made at the outset of this chapter, we can measure anything that's real. We can measure, for example, whether Pat actually puts the little bird back in its nest, visits the hospital on Christmas, weeps at the movie, or refuses to contribute to saving the whales. All of those behaviors exist, so we can measure them. But is Pat really compassionate? We can't answer that question; we can't measure compassion in any objective sense, because compassion doesn't exist in the way that those things I just described exist. Compassion exists only in the form of the agreements we have about how to use the term in communicating about things that are real.

Concepts as Constructs

If you recall the discussions of postmodernism in Chapter 2, you'll recognize that some people would object to the degree of "reality" I've allowed in the preceding comments. Did Pat "really" visit the hospital on Christmas? Does the hospital "really" exist? Does Christmas? Though we aren't going to be radically postmodern in this chapter, I think you'll recognize the importance of an intellectually tough view of what's real and what's not. (When the intellectual going gets tough, the tough become social scientists.)

In this context, Abraham Kaplan (1964) distinguishes three classes of things that scientists measure. The first class is *direct observables*: those things we can observe rather simply and directly, like the color of an apple or the check mark on a questionnaire. The second class, *indirect observables,* require "relatively more subtle, complex, or indirect observations" (1964: 55). We note a person's check mark beside "female" in a questionnaire and have indirectly observed that person's gender. History books or minutes of corporate board meetings provide indirect observations of past social actions. Finally, the third class of observables consists of *constructs*—theoretical creations that are based on observations but that cannot be observed directly or indirectly. A good example is intelligence quotient, or IQ. It is constructed mathematically from observations of the answers given to a large number of questions on an IQ test. No one can directly or indirectly observe IQ. It is no more a "real" characteristic of people than is compassion or prejudice. See Table 5-1 for more examples of what social scientists measure.

Kaplan (1964: 49) defines *concept* as a "family of conceptions." A concept is, as Kaplan notes, a construct, something we create. Concepts such as compassion and prejudice are constructs created from your conception of them, my conception of them, and the conceptions of all those who have ever used these terms. They cannot be observed directly or indirectly, because they don't exist. We made them up.

To summarize, *concepts* are constructs derived by mutual agreement from mental images (conceptions). Our *conceptions* summarize collections

TABLE 5-1
What Social Scientists Measure

	Examples
Direct observables	Physical characteristics (sex, height, skin color) of a person being observed and/or interviewed
Indirect observables	Characteristics of a person as indicated by answers given in a self-administered questionnaire
Constructs	Level of alienation, as measured by a scale that is created by combining several direct and/or indirect observables

of seemingly related observations and experiences. Although the observations and experiences are real, at least subjectively, conceptions, and the concepts derived from them, are only mental creations. The terms associated with concepts are merely devices created for the purposes of filing and communication. A term such as *prejudice* is, objectively speaking, only a collection of letters. It has no intrinsic reality beyond that. Is has only the meaning we agree to give it.

Usually, however, we fall into the trap of believing that terms for constructs do have intrinsic meaning, that they name real entities in the world. That danger seems to grow stronger when we begin to take terms seriously and attempt to use them precisely. Further, the danger is all the greater in the presence of experts who appear to know more than we do about what the terms really mean: It's easy to yield to authority in such a situation.

Once we assume that terms like *prejudice* and *compassion* have real meanings, we begin the tortured task of discovering what those real meanings are and what constitutes a genuine measurement of them. Regarding constructs as real is called *reification*. The reification of concepts in day-to-day life is quite common. In science, we want to be quite clear about what it is we are actually measuring, but this aim brings a pitfall with it. Settling on the "best" way of measuring a variable in a particular study may imply that we've discovered the "real" meaning of the concept involved. In fact, concepts have no real, true, or objective meanings—only those we agree are best for a particular purpose.

Does this discussion imply that compassion, prejudice, and similar constructs can't be measured? Interestingly, the answer is no. (And a good thing, too, or a lot of us social researcher types would be out of work.) I've said that we can measure anything that's real. Constructs aren't real in the way that trees are real, but they do have another important virtue: They are useful. That is, they help us organize, communicate about, and understand things that are real. They help us make predictions about real things. Some of those predictions even turn out to be true. Constructs can work this way because, although not real or observable in themselves, they have a definite relationship to things that are real and observable. The bridge from direct and indirect observables to useful constructs is the process called conceptualization.

indicator An observation that we choose to consider as a reflection of a variable we wish to study. Thus, for example, attending religious services might be considered an indicator of *religiosity*.

Conceptualization

As we've seen, day-to-day communication usually occurs through a system of vague and general agreements about the use of terms. Although you and I do not agree completely about the use of the term *compassionate*, I'm probably safe in assuming that Pat won't pull the wings off flies. A wide range of misunderstandings and conflict—from the interpersonal to the international—is the price we pay for our imprecision, but somehow we muddle through. Science, however, aims at more than muddling; it cannot operate in a context of such imprecision.

The process through which we specify what we mean when we use particular terms in research is called *conceptualization*. Suppose we want to find out, for example, whether women are more compassionate than men. I suspect many people assume this is the case, but it might be interesting to find out if it's really so. We can't meaningfully study the question, let alone agree on the answer, without some working agreements about the meaning of compassion. They are "working" agreements in the sense that they allow us to work on the question. We don't need to agree or even pretend to agree that a particular specification is ultimately the best one.

Conceptualization, then, produces a specific, agreed-on meaning for a concept for the purposes of research. This process of specifying exact meaning involves describing the indicators we'll be using to measure our concept and the different aspects of the concept, called dimensions.

Indicators and Dimensions

Conceptualization gives definite meaning to a concept by specifying one or more indicators of what we have in mind. An **indicator** is a sign of the presence or absence of the concept we are studying. Here's an example.

We might agree that visiting children's hospitals during Christmas and Hanukkah is an indicator of compassion. Putting little birds back in their nests might be agreed on as another indicator, and so forth. If the unit of analysis for our study is the individual person, we can then observe the presence or absence of each indicator for each person under study. Going beyond that, we can add up the number of indicators of compassion observed for each individual. We might agree on ten specific indicators, for example, and find six present in our study of Pat, three for John, nine for Mary, and so forth.

Returning to our question about whether men or women are more compassionate, we might calculate that the women we studied displayed an average of 6.5 indicators of compassion, the men an average of 3.2. On the basis of our quantitative analysis of group difference, we might therefore conclude that women are, on the whole, more compassionate than men.

Usually, though, it's not that simple. Imagine you're interested in understanding a small fundamentalist religious cult, particularly their harsh views on various groups: gays, nonbelievers, feminists, and others. In fact, they suggest that anyone who refuses to join their group and abide by its teachings will "burn in hell." In the context of your interest in compassion, they don't seem to have much. And yet, the group's literature often speaks of their compassion for others. You want to explore this seeming paradox.

To pursue this research interest, you might arrange to interact with cult members, getting to know them and learning more about their views. You could tell them you were a social researcher interested in learning about their group, or perhaps you would just express an interest in learning more, without saying why.

In the course of your conversations with group members and perhaps attendance of religious services, you would put yourself in situations where you could come to understand what the cult members mean by compassion. You might learn, for example, that members of the group were so deeply concerned about sinners burning in hell that they were willing to be aggressive, even violent, to make people change their sinful ways. Within their own paradigm, then, cult members would see beating up gays, prostitutes, and abortion doctors as acts of compassion.

Social researchers focus their attention on the meanings that the people under study give to words and actions. Doing so can often clarify the behaviors observed: At least now you understand how the cult can see violent acts as compassionate. On the other hand, paying attention to what words and actions mean to the people under study almost always complicates the concepts researchers are interested in. (We'll return to this issue when we discuss the validity of measures, toward the end of this chapter.)

Whenever we take our concepts seriously and set about specifying what we mean by them, we discover disagreements and inconsistencies. Not only do you and I disagree, but each of us is likely to find a good deal of muddiness within our own mental images. If you take a moment to look at what you mean by compassion, you'll probably find that your image contains several kinds of compassion. That is, the entries on your mental file sheet can be combined into groups and subgroups, say, compassion toward friends, co-religionists, humans, and birds. You may also find several different strategies for making combinations. For example, you might group the entries into feelings and actions.

The technical term for such groupings is **dimension**, a specifiable aspect of a concept. For instance, we might speak of the "feeling dimension" of compassion and the "action dimension" of compassion. In a different grouping scheme, we might distinguish "compassion for humans" from "compassion for animals." Or we might see compassion as helping people have what we want for them versus what they want for themselves. Still differently, we might distinguish compassion as forgiveness from compassion as pity.

Thus, we could subdivide compassion into several clearly defined dimensions. A complete conceptualization involves both specifying dimensions and identifying the various indicators for each.

dimension A specifiable aspect of a concept. "Religiosity," for example, might be specified in terms of a belief dimension, a ritual dimension, a devotional dimension, a knowledge dimension, and so forth.

When Jonathan Jackson (2005: 301) set out to measure "fear of crime," he considered seven different dimensions:

- The frequency of worry about becoming a victim of three personal crimes and two property crimes in the immediate neighbourhood
- Estimates of likelihood of falling victim to each crime locally
- Perceptions of control over the possibility of becoming a victim of each crime locally
- Perceptions of the seriousness of the consequences of each crime
- Beliefs about the incidence of each crime locally
- Perceptions of the extent of social physical incivilities in the neighbourhood
- Perceptions of community cohesion, including informal social control and trust/social capital

Sometimes conceptualization aimed at identifying different dimensions of a variable leads to a different kind of distinction. We may conclude that we've been using the same word for meaningfully distinguishable concepts. In the following example, the researchers find (1) that "violence" is not a sufficient description of "genocide" and (2) that the concept "genocide" itself comprises several distinct phenomena. Let's look at the process they went through to come to this conclusion.

When Daniel Chirot and Jennifer Edwards attempted to define the concept of "genocide," they found existing assumptions were not precise enough for their purposes:

> The United Nations originally defined it as an attempt to destroy "in whole or in part, a national, ethnic, racial, or religious group." If genocide is distinct from other types of violence, it requires its own unique explanation. *(2003: 14)*

Notice the final comment in this excerpt, as it provides an important insight into why researchers are so careful in specifying the concepts they study. If genocide, such as the Holocaust, were simply another example of violence, like assaults and homicides, then what we know about violence in general might explain genocide. If it differs from other forms of violence, then we may need a different explanation for it. So, the researchers began by suggesting that "genocide" was a concept distinct from "violence" for their purposes.

Then, as Chirot and Edwards examined historical instances of genocide, they began concluding that the motivations for launching genocidal mayhem differed sufficiently to represent four distinct phenomena that were all called "genocide" (2003: 15–18).

1. *Convenience:* Sometimes the attempt to eradicate a group of people serves a function for the eradicators, such as Julius Caesar's attempt to eradicate tribes defeated in battle, fearing they would be difficult to rule. Or when gold was discovered on Cherokee land in the Southeastern United States in the early nineteenth century, the Cherokee were forcibly relocated to Oklahoma in an event known as the "Trail of Tears," which ultimately killed as many as half of those forced to leave.
2. *Revenge:* When the Chinese of Nanking bravely resisted the Japanese invaders in the early years of World War II, the conquerors felt they had been insulted by those they regarded as inferior beings. Tens of thousands were slaughtered in the "Rape of Nanking" in 1937–1938.
3. *Fear:* The ethnic cleansing that recently occurred in the former Yugoslavia was at least partly motivated by economic competition and worries that the growing Albanian population of Kosovo was gaining political strength through numbers. Similarly, the Hutu attempt to eradicate the Tutsis of Rwanda grew out of a fear that returning Tutsi refugees would seize control of the country. Often intergroup fears such as these grow out of long histories of atrocities, often inflicted in both directions.
4. *Purification:* The Nazi Holocaust, probably the most publicized case of genocide, was intended as a purification of the "Aryan race." While Jews were the main target, gypsies, homosexuals, and many other groups were also included. Other examples include the Indonesian witch hunt against Communists in 1965–1966, and the attempt to eradicate all non-Khmer Cambodians under Pol Pot in the 1970s.

No single theory of genocide could explain these various forms of mayhem. Indeed, this act of conceptualization suggests four distinct phenomena, each needing a different set of explanations.

Specifying the different dimensions of a concept often paves the way for a more sophisticated understanding of what we're studying. We might observe, for example, that women are more compassionate in terms of feelings, and men more so in terms of actions—or vice versa. Whichever turned out to be the case, we would not be able to say whether men or women are really more compassionate. Our research would have shown that there is no single answer to the question. That alone represents an advance in our understanding of reality.

The Interchangeability of Indicators

There is another way that the notion of indicators can help us in our attempts to understand reality by means of "unreal" constructs. Suppose, for the moment, that you and I have compiled a list of 100 indicators of compassion and its various dimensions. Suppose further that we disagree widely on which indicators give the clearest evidence of compassion or its absence. If we pretty much agree on some indicators, we could focus our attention on those, and we would probably agree on the answer they provided. We would then be able to say that some people are more compassionate than others in some dimension. But suppose we don't really agree on any of the possible indicators. Surprisingly, we can still reach an agreement on whether men or women are the more compassionate. How we do that has to do with the interchangeability of indicators.

The logic works like this. If we disagree totally on the value of the indicators, one solution would be to study all of them. Suppose that women turn out to be more compassionate than men on all 100 indicators—on all the indicators you favor and on all of mine. Then we would be able to agree that women are more compassionate than men, even though we still disagree on exactly what compassion means in general.

The interchangeability of indicators means that if several different indicators all represent, to some degree, the same concept, then all of them will behave the same way that the concept would behave if it were real and could be observed. Thus, given a basic agreement about what "compassion" is, if women are generally more compassionate than men, we should be able to observe that difference by using any reasonable measure of compassion. If, on the other hand, women are more compassionate than men on some indicators but not on others, we should see if the two sets of indicators represent different dimensions of compassion.

You have now seen the fundamental logic of conceptualization and measurement. The discussions that follow are mainly refinements and extensions of what you've just read. Before turning to a technical elaboration of measurement, however, we need to fill out the picture of conceptualization by looking at some of the ways social researchers provide standards, consistency, and commonality for the meanings of terms.

Real, Nominal, and Operational Definitions

As we have seen, the design and execution of social research requires us to clear away the confusion over concepts and reality. To this end, logicians and scientists have found it useful to distinguish three kinds of definitions: real, nominal, and operational.

The first of these reflects the reification of terms. As Carl Hempel cautions,

> A "real" definition, according to traditional logic, is not a stipulation determining the meaning of some expression but a statement of the "essential nature" or the "essential attributes" of some entity. The notion of essential nature, however, is so vague as to render this characterization useless for the purposes of rigorous inquiry.
>
> *(1952: 6)*

In other words, trying to specify the "real" meaning of concepts only leads to a quagmire: It mistakes a construct for a real entity.

The **specification** of concepts in scientific inquiry depends instead on nominal and operational definitions. A nominal definition is one

specification The process through which concepts are made more specific.

that is simply assigned to a term without any claim that the definition represents a "real" entity. Nominal definitions are arbitrary—I could define compassion as "plucking feathers off helpless birds" if I wanted to—but as definitions they can be more or less useful. For most purposes, especially communication, that last definition of compassion would be pretty useless. Most nominal definitions represent some consensus, or convention, about how a particular term is to be used.

An operational definition, as you may remember from Chapter 4, specifies precisely how a concept will be measured—that is, the operations we'll perform. An operational definition is nominal rather than real, but it has the advantage of achieving maximum clarity about what a concept means in the context of a given study. In the midst of disagreement and confusion over what a term "really" means, we can specify a working definition for the purposes of an inquiry. Wishing to examine socioeconomic status (SES) in a study, for example, we may simply specify that we are going to treat SES as a combination of income and educational attainment. In this decision, we rule out other possible aspects of SES: occupational status, money in the bank, property, lineage, lifestyle, and so forth. Our findings will then be interesting to the extent that our definition of SES is useful for our purpose.

Creating Conceptual Order

The clarification of concepts is a continuing process in social research. Catherine Marshall and Gretchen Rossman (1995: 18) speak of a "conceptual funnel" through which a researcher's interest becomes increasingly focused. Thus, a general interest in social activism could narrow to "individuals who are committed to empowerment and social change" and further focus on discovering "what experiences shaped the development of fully committed social activists." This focusing process is inescapably linked to the language we use.

In some forms of qualitative research, the clarification of concepts is a key element in the collection of data. Suppose you were conducting interviews and observations in a radical political group devoted to combating oppression in U.S. society. Imagine how the meaning of oppression would shift as you delved more and more deeply into the members' experiences and worldviews. For example, you might start out thinking of oppression in physical and perhaps economic terms. The more you learned about the group, however, the more you might appreciate the possibility of psychological oppression.

The same point applies even to contexts where meanings might seem more fixed. In the analysis of textual materials, for example, social researchers sometimes speak of the "hermeneutic circle," a cyclical process of ever-deeper understanding.

> The understanding of a text takes place through a process in which the meaning of the separate parts is determined by the global meaning of the text as it is anticipated. The closer determination of the meaning of the separate parts may eventually change the originally anticipated meaning of the totality, which again influences the meaning of the separate parts, and so on.
>
> *(Kvale 1996: 47)*

Consider the concept "prejudice." Suppose you needed to write a definition of the term. You might start out thinking about racial/ethnic prejudice. At some point you would realize you should probably allow for gender prejudice, religious prejudice, antigay prejudice, and the like in your definition. Examining each of these specific types of prejudice would affect your overall understanding of the general concept. As your general understanding changed, however, you would likely see each of the individual forms somewhat differently.

The continual refinement of concepts occurs in all social research methods. Often you will find yourself refining the meaning of important concepts even as you write up your final report.

Although conceptualization is a continuing process, it is vital to address it specifically at the beginning of any study design, especially rigorously structured research designs such as surveys and experiments. In a survey, for example, operationalization results in a commitment to a specific set of questionnaire items that will represent the concepts under study. Without that commitment, the study could not proceed.

Even in less-structured research methods, however, it's important to begin with an initial

set of anticipated meanings that can be refined during data collection and interpretation. No one seriously believes we can observe life with no preconceptions; for this reason, scientific observers must be conscious of and explicit about these conceptual starting points.

Let's explore initial conceptualization the way it applies to structured inquiries such as surveys and experiments. Though specifying nominal definitions focuses our observational strategy, it does not allow us to observe. As a next step we must specify exactly what we are going to observe, how we will do it, and what interpretations we are going to place on various possible observations. All these further specifications make up the operational definition of the concept.

In the example of socioeconomic status, we might decide to ask survey respondents two questions, corresponding to the decision to measure SES in terms of income and educational attainment:

1. What was your total family income during the past 12 months?
2. What is the highest level of school you completed?

To organize our data, we'd probably want to specify a system for categorizing the answers people give us. For income, we might use categories such as "under $10,000," "$10,000 to $25,000," and so on. Educational attainment might be similarly grouped in categories: less than high school, high school, college, graduate degree. Finally, we would specify the way a person's responses to these two questions would be combined in creating a measure of SES.

In this way we would create a working and workable definition of SES. Although others might disagree with our conceptualization and operationalization, the definition would have one essential scientific virtue: It would be absolutely specific and unambiguous. Even if someone disagreed with our definition, that person would have a good idea how to interpret our research results, because what we meant by SES—reflected in our analyses and conclusions—would be precise and clear.

Table 5-2 shows the progression of measurement steps from our vague sense of what a term means to specific measurements in a fully structured scientific study.

TABLE 5-2
Progression of Measurement

Measurement Step	*Example: Social Class*
Conceptualization	What are the different meanings and dimensions of the concept "social class"?
Nominal definition	For our study, we will define "social class" as representing economic differences: specifically, income.
Operational definition	We will measure economic differences via responses to the survey question "What was your annual income, before taxes, last year?"
Measurements in the real world	The interviewer will ask, "What was your annual income, before taxes, last year?"

An Example of Conceptualization: The Concept of Anomie

To bring this discussion of conceptualization in research together, let's look briefly at the history of a specific social science concept. Researchers studying urban riots are often interested in the part played by feelings of powerlessness. Social scientists sometimes use the word *anomie* in this context. This term was first introduced into social science by Emile Durkheim, the great French sociologist, in his classic 1897 study, *Suicide*.

Using only government publications on suicide rates in different regions and countries, Durkheim produced a work of analytic genius. To determine the effects of religion on suicide, he compared the suicide rates of predominantly Protestant countries with those of predominantly Catholic ones, Protestant regions of Catholic countries with Catholic regions of Protestant countries, and so forth. To determine the possible effects of the weather, he compared suicide rates in northern and southern countries and regions, and he examined the different suicide rates across the months and seasons of the year. Thus, he could draw conclusions about a supremely individualistic and personal act without having any data about the individuals engaging in it.

At a more general level, Durkheim suggested that suicide also reflects the extent to which a

society's agreements are clear and stable. Noting that times of social upheaval and change often present individuals with grave uncertainties about what is expected of them, Durkheim suggested that such uncertainties cause confusion, anxiety, and even self-destruction. To describe this societal condition of normlessness, Durkheim chose the term *anomie.* Durkheim did not make up this word. Used in both German and French, it literally means "without law." The English term *anomy* had been used for at least three centuries before Durkheim to mean disregard for divine law. However, Durkheim created the social science concept of *anomie*.

In the years that have followed the publication of *Suicide*, social scientists have found anomie a useful concept, and many have expanded on Durkheim's use. Robert Merton, in a classic article entitled "Social Structure and Anomie" (1938), concluded that anomie results from a disparity between the goals and means prescribed by a society. Monetary success, for example, is a widely shared goal in our society, yet not all individuals have the resources to achieve it through acceptable means. An emphasis on the goal itself, Merton suggested, produces normlessness, because those denied the traditional avenues to wealth go about getting it through illegitimate means. Merton's discussion, then, could be considered a further conceptualization of the concept of anomie.

Although Durkheim originally used the concept of anomie as a characteristic of societies, as did Merton after him, other social scientists have used it to describe individuals. To clarify this distinction, some scholars have chosen to use *anomie* in reference to its original, societal meaning and to use the term *anomia* in reference to the individual characteristic. In a given society, then, some individuals experience anomia, and others do not. Elwin Powell, writing 20 years after Merton, provided the following conceptualization of anomia (though using the term *anomie*) as a characteristic of individuals:

> When the ends of action become contradictory, inaccessible or insignificant, a condition of anomie arises. Characterized by a general loss of orientation and accompanied by feelings of "emptiness" and apathy, anomie can be simply conceived as meaninglessness.
>
> *(1958: 132)*

Powell went on to suggest there were two distinct kinds of anomia and to examine how the two rose out of different occupational experiences to result at times in suicide. In his study, however, Powell did not measure anomia per se; he studied the relationship between suicide and occupation, making inferences about the two kinds of anomia. Thus, the study did not provide an operational definition of anomia, only a further conceptualization.

Although many researchers have offered operational definitions of anomia, one name stands out over all. Two years before Powell's article appeared, Leo Srole (1956) published a set of questionnaire items that he said provided a good measure of anomia as experienced by individuals. It consists of five statements that subjects were asked to agree or disagree with:

> 1. In spite of what some people say, the lot of the average man is getting worse.
> 2. It's hardly fair to bring children into the world with the way things look for the future.
> 3. Nowadays a person has to live pretty much for today and let tomorrow take care of itself.
> 4. These days a person doesn't really know who he can count on.
> 5. There's little use writing to public officials because they aren't really interested in the problems of the average man.
>
> *(1956: 713)*

In the half-century following its publication, the Srole scale has become a research staple for social scientists. You'll likely find this particular operationalization of anomia used in many of the research projects reported in academic journals.

This abbreviated history of anomie and anomia as social science concepts illustrates several points. First, it's a good example of the process through which general concepts become operationalized measurements. This is not to say that the issue of how to operationalize anomie/anomia has been resolved once and for all. Scholars will surely continue to reconceptualize and reoperationalize these concepts for years to come, continually seeking more-useful measures.

The Srole scale illustrates another important point. Letting conceptualization and

operationalization be open-ended does not necessarily produce anarchy and chaos, as you might expect. Order often emerges. For one thing, although we could define anomia any way we chose—in terms of, say, shoe size—we're likely to define it in ways not too different from other people's mental images. If you were to use a really offbeat definition, people would probably ignore you.

A second source of order is that, as researchers discover the utility of a particular conceptualization and operationalization of a concept, they're likely to adopt it, which leads to standardized definitions of concepts. Besides the Srole scale, examples include IQ tests and a host of demographic and economic measures developed by the U.S. Census Bureau. Using such established measures has two advantages: They have been extensively pretested and debugged, and studies using the same scales can be compared. If you and I do separate studies of two different groups and use the Srole scale, we can compare our two groups on the basis of anomia.

Social scientists, then, can measure anything that's real; through conceptualization and operationalization, they can even do a pretty good job of measuring things that aren't. Granting that such concepts as socioeconomic status, prejudice, compassion, and anomia aren't ultimately real, social scientists can create order in handling them. It is an order based on utility, however, not on ultimate truth.

Definitions in Descriptive and Explanatory Studies

As you'll recall from Chapter 4, two general purposes of research are description and explanation. The distinction between them has important implications for definition and measurement. If it seems that description is simpler than explanation, you may be surprised to learn that definitions are more problematic for descriptive research than for explanatory research. Before we turn to other aspects of measurement, you'll need a basic understanding of why this is so (we'll discuss this point more fully in Part 4).

It's easy to see the importance of clear and precise definitions for descriptive research. If we want to describe and report the unemployment rate in a city, our definition of being unemployed is obviously critical. That definition will depend on our definition of another term: the labor force. If it seems patently absurd to regard a three-year-old child as being unemployed, it is because such a child is not considered a member of the labor force. Thus, we might follow the U.S. Census Bureau's convention and exclude all people under 14 years of age from the labor force.

This convention alone, however, would not give us a satisfactory definition, because it would count as unemployed such people as high school students, the retired, the disabled, and homemakers who don't want to work outside the home. We might follow the census convention further by defining the labor force as "all persons 14 years of age and over who are employed, looking for work, or waiting to be called back to a job from which they have been laid off or furloughed." If a student, homemaker, or retired person is not looking for work, such a person would not be included in the labor force. Unemployed people, then, would be those members of the labor force, as defined, who are not employed.

But what does "looking for work" mean? Must a person register with the state employment service or go from door to door asking for employment? Or would it be sufficient to want a job or be open to an offer of employment? Conventionally, "looking for work" is defined operationally as saying yes in response to an interviewer's asking "Have you been looking for a job during the past seven days?" (Seven days is the period most often specified, but for some research purposes it might make more sense to shorten or lengthen it.)

As you can see, the conclusion of a descriptive study about the unemployment rate depends directly on how each issue of definition is resolved. Increasing the period during which people are counted as looking for work would add more unemployed people to the labor force as defined, thereby increasing the reported unemployment rate. If we follow another convention and speak of the civilian labor force and the civilian unemployment rate, we're excluding military personnel; that, too, increases the reported unemployment rate, because military personnel would be employed—by definition. Thus, the descriptive statement that the unemployment rate

in a city is 3 percent, or 9 percent, or whatever it might be, depends directly on the operational definitions used.

This example is relatively clear because there are several accepted conventions relating to the labor force and unemployment. Now, consider how difficult it would be to get agreement about the definitions you would need in order to say, "Forty-five percent of the students at this institution are politically conservative." Like the unemployment rate, this percentage would depend directly on the definition of what is being measured—in this case, political conservatism. A different definition might result in the conclusion "Five percent of the student body are politically conservative."

What percentage of the population do you suppose is "disabled"? That's the question Lars Gronvik asked in Sweden. He analyzed several databases that encompassed four different definitions or measures of disability in Swedish society. One study asked people if they had hearing, seeing, walking, or other functional problems. Two other measures were based on whether people received one of two forms of government disability support. Another study asked people whether they believed they were disabled.

The four measures indicated different population totals for those citizens defined as "disabled," and each measure produced different demographic profiles that included variables such as sex, age, education, living arrangement, and labor-force participation. As you can see, it is impossible to answer a descriptive question such as this without specifying the meaning of terms.

Ironically, definitions are less problematic in the case of explanatory research. Let's suppose we're interested in explaining political conservatism. Why are some people conservative and others not? More specifically, let's suppose we're interested in whether conservatism increases with age. What if you and I have 25 different operational definitions of *conservative,* and we can't agree on which definition is best? As we saw in the discussion of indicators, this is not necessarily an insurmountable obstacle to our research. Suppose we found old people to be more conservative than young people in terms of all 25 definitions. Clearly, the exact definition wouldn't matter much. We would conclude that old people are generally more conservative than young people—even though we couldn't agree about exactly what *conservative* means.

In practice, explanatory research seldom results in findings quite as unambiguous as this example suggests; nonetheless, the general pattern is quite common in actual research. There are consistent patterns of relationships in human social life that result in consistent research findings. However, such consistency does not appear in a descriptive situation. Changing definitions almost inevitably results in different descriptive conclusions.

Operationalization Choices

In discussing conceptualization, I frequently have referred to operationalization, for the two are intimately linked. To recap: Conceptualization is the refinement and specification of abstract concepts, and operationalization is the development of specific research procedures (operations) that will result in empirical observations representing those concepts in the real world.

As with the methods of data collection, social researchers have a variety of choices when operationalizing a concept. Although the several choices are intimately interconnected, I've separated them for the sake of discussion. Realize, though, that operationalization does not proceed through a systematic checklist.

Range of Variation

In operationalizing any concept, researchers must be clear about the range of variation that interests them. The question is, to what extent are they willing to combine attributes in fairly gross categories?

Let's suppose you want to measure people's incomes in a study by collecting the information from either records or interviews. The highest annual incomes people receive run into the millions of dollars, but not many people earn that much. Unless you're studying the very rich, it probably won't add much to your study to keep track of extremely high categories. Depending on whom you study, you'll probably want to establish a highest income category with a much lower floor—maybe $250,000 or more. Although this decision will lead you to throw together

people who earn a trillion dollars a year with paupers earning a mere $250,000, they'll survive it, and that mixing probably won't hurt your research any, either. The same decision faces you at the other end of the income spectrum. In studies of the general U.S. population, a bottom category of $10,000 or less usually works fine.

In studies of attitudes and orientations, the question of range of variation has another dimension. Unless you're careful, you may end up measuring only half an attitude without really meaning to. Here's an example of what I mean.

Suppose you're interested in people's attitudes toward expanding the use of nuclear power generators. You'd anticipate that some people consider nuclear power the greatest thing since the wheel, whereas other people have absolutely no interest in it. Given that anticipation, it would seem to make sense to ask people how much they favor expanding the use of nuclear energy and to give them answer categories ranging from "Favor it very much" to "Don't favor it at all."

This operationalization, however, conceals half the attitudinal spectrum regarding nuclear energy. Many people have feelings that go beyond simply not favoring it: They are, with greater or lesser degrees of intensity, actively opposed to it. In this instance, there is considerable variation on the left side of zero. Some oppose it a little, some quite a bit, and others a great deal. To measure the full range of variation, then, you'd want to operationalize attitudes toward nuclear energy with a range from favoring it very much, through no feelings one way or the other, to opposing it very much.

This consideration applies to many of the variables social scientists study. Virtually any public issue involves both support and opposition, each in varying degrees. In measuring religiosity, people are not just more or less religious; some are positively antireligious. Political orientations range from very liberal to very conservative, and depending on the people you're studying, you may want to allow for radicals on one or both ends.

The point is not that you must measure the full range of variation in every case. You should, however, consider whether you need to, given your particular research purpose. If the difference between not religious and antireligious isn't relevant to your research, forget it. Someone has defined pragmatism as "any difference that makes no difference is no difference." Be pragmatic.

Finally, decisions on the range of variation should be governed by the expected distribution of attributes among the subjects of the study. In a study of college professors' attitudes toward the value of higher education, you could probably stop at no value and not worry about those who might consider higher education dangerous to students' health. (If you were studying students, however . . .)

Variations between the Extremes

Degree of precision is a second consideration in operationalizing variables. What it boils down to is how fine you will make distinctions among the various possible attributes composing a given variable. Does it matter for your purposes whether a person is 17 or 18 years old, or could you conduct your inquiry by throwing them together in a group labeled 10 to 19 years old? Don't answer too quickly. If you wanted to study rates of voter registration and participation, you'd definitely want to know whether the people you studied were old enough to vote. In general, if you're going to measure age, you must look at the purpose and procedures of your study and decide whether fine or gross differences in age are important to you. In a survey, you'll need to make these decisions in order to design an appropriate questionnaire. In the case of in-depth interviews, these decisions will condition the extent to which you probe for details.

The same thing applies to other variables. If you measure *political affiliation,* will it matter to your inquiry whether a person is a conservative Democrat rather than a liberal Democrat, or will it be sufficient to know the party? In measuring *religious affiliation,* is it enough to know that a person is Protestant, or do you need to know the denomination? Do you simply need to know if a person is married, or will it make a difference to know if he or she has never married or is separated, widowed, or divorced?

There is, of course, no general answer to such questions. The answers come out of the purpose of a given study, or why we are making a particular measurement. I can give you a useful

guideline, though. Whenever you're not sure how much detail to pursue in a measurement, get too much rather than too little. When a subject in an in-depth interview volunteers that she is 37 years old, record "37" in your notes, not "in her thirties." When you're analyzing the data, you can always combine precise attributes into more-general categories, but you can never separate any variations you lumped together during observation and measurement.

A Note on Dimensions

We've already discussed dimensions as a characteristic of concepts. When researchers get down to the business of creating operational measures of variables, they often discover—or worse, never notice—that they're not exactly clear about which dimensions of a variable they're really interested in. Here's an example.

Let's suppose you're studying people's attitudes toward government, and you want to include an examination of how people feel about corruption. Here are just a few of the dimensions you might examine:

- Do people think there is corruption in government?
- How much corruption do they think there is?
- How certain are they in their judgment of how much corruption there is?
- How do they feel about corruption in government as a problem in society?
- What do they think causes it?
- Do they think it's inevitable?
- What do they feel should be done about it?
- What are they willing to do personally to eliminate corruption in government?
- How certain are they that they would be willing to do what they say they would do?

The list could go on and on—how people feel about corruption in government has many dimensions. It's essential to be clear about which ones are important in your inquiry; otherwise, you may measure how people feel about corruption when you really wanted to know how much they think there is, or vice versa.

Once you've determined how you're going to collect your data (for example, survey, field research) and have decided on the relevant range of variation, the degree of precision needed between the extremes of variation, and the specific dimensions of the variables that interest you, you may have another choice: a mathematical-logical one. That is, you may need to decide what level of measurement to use. To discuss this point, we need to take another look at attributes and their relationship to variables.

Defining Variables and Attributes

The conceptualization and operationalization processes can be seen as the specification of variables and the attributes composing them. Thus, in the context of a study of unemployment, *employment status* is a variable having the attributes *employed* and *unemployed*; the list of attributes could also be expanded to include the other possibilities discussed earlier, such as *homemaker*.

Every variable must have two important qualities. First, the attributes composing it should be exhaustive. For the variable to have any utility in research, we must be able to classify every observation in terms of one of the attributes composing the variable. We'll run into trouble if we conceptualize the variable *political party affiliation* in terms of the attributes *Republican* and *Democrat*, because some of the people we set out to study will identify with the Green Party, the Libertarian Party, or some other organization, and some (often a large percentage) will tell us they have no party affiliation. We could make the list of attributes exhaustive by adding "other" and "no affiliation." Whatever we do, we must be able to classify every observation.

At the same time, attributes composing a variable must be mutually exclusive. That is, we must be able to classify every observation in terms of one and only one attribute. For example, we need to define "employed" and "unemployed" in such a way that nobody can be both at the same time. That means being able to classify the person who is working at a job but is also looking for work. (We might run across a fully employed mud wrestler who is looking for the glamour and excitement of being a social researcher.) In this case, we might define the attributes so that *employed* takes precedence over *unemployed*, and anyone working at a job is *employed* regardless of whether he or she is looking for something better.

The process of conceptualizing variables is very situation dependent. What works in one situation won't necessarily work elsewhere. Malcom Williams and Kerryn Husk (2013) have examined in detail the many problems involved in measuring ethnicity. To begin, there are no absolute definitions of various ethnic groups; they are a matter of social conventions, which are understood differently by different people and which change over time. Although they focused on Cornwall County in Britain, the authors' analysis applies to the measurement of ethnicity more broadly and, indeed, applies to conceptualization in general.

Two general conclusions can be drawn. First, the conceptualization of variables depends, obviously perhaps, on the population being studied. A survey conducted in Cornwall might include the ethnic category of "Cornish," whereas you wouldn't have that category in a survey of Arkansas. Second, conceptualization should be tailored to the purposes of the study. In the case of ethnicity, four or five broad ethnic categories might suffice in one study, while the intentions of another might require much finer distinctions.

Levels of Measurement

All variables are composed of attributes, but as we are about to see, the attributes of a given variable can have a variety of different relationships to one another. In this section, we'll examine four levels of measurement: nominal, ordinal, interval, and ratio.

Nominal Measures

Variables whose attributes are simply different from one another are called *nominal measures.* Examples include *gender, religious affiliation, political party affiliation, birthplace, college major,* and *hair color.* Although the attributes composing each of these variables—as *male* and *female* compose the variable *gender*—are distinct from one another, they have no additional structures. **Nominal measures** merely offer names or labels for characteristics.

Imagine a group of people characterized in terms of one such nominal variable and physically grouped by the applicable attributes. For example, say we've asked a large gathering of people to stand together in groups according to the states in which they were born: all those born in Vermont in one group, those born in California in another, and so forth. The variable is *state of birth;* the attributes are *born in California, born in Vermont,* and so on. All the people standing in a given group have at least one thing in common and differ from the people in all other groups in that same regard. Where the individual groups form, how close they are to one another, or how the groups are arranged in the room is irrelevant. What matters is that all the members of a given group share the same state of birth and that each group has a different shared state of birth. All we can say about two people in terms of a nominal variable is that they are either the same or different.

Ordinal Measures

Variables with attributes we can logically rank-order are *ordinal measures.* The different attributes of ordinal variables represent relatively more or less of the variable. Variables of this type are *social class, conservatism, alienation, prejudice, intellectual sophistication,* and the like. In addition to saying whether two people are the same or different in terms of an ordinal variable, you can also say one is "more" than the other—that is, more conservative, more religious, older, and so forth.

In the physical sciences, *hardness* is the most frequently cited example of an ordinal measure. We may say that one material (for example, diamond) is harder than another (say, glass) if the former can scratch the latter and not vice versa. By attempting to scratch various materials with other materials, we might eventually be able to arrange several materials in a row, ranging from the softest to the hardest. We could never say how hard a given material was in absolute terms; we could only say how hard in relative terms—which materials it is harder than and which softer than.

Let's pursue the earlier example of grouping the people at a social gathering. This time imagine that we ask all the people who have

nominal measure A nominal variable has attributes that are merely different, as distinguished from ordinal, interval, or ratio measures. *Gender* is an example of a nominal measure. All a nominal variable can tell us about two people is if they are the same or different.

graduated from college to stand in one group, all those with only a high school diploma to stand in another group, and all those who have not graduated from high school to stand in a third group. This manner of grouping people satisfies the nominal-variable quality of being different, as discussed earlier. In addition, however, we might logically arrange the three groups in terms of the relative amount of formal education (the shared attribute) each had. We might arrange the three groups in a row, ranging from most- to least-formal education. This arrangement would provide a physical representation of an **ordinal measure**. If we knew which groups two individuals were in, we could determine that one had more, less, or the same formal education as the other.

In this example, it is irrelevant how close or far apart the educational groups are from one another. The college and high school groups might be 5 feet apart, and the less-than-high-school group 500 feet farther down the line. These actual distances don't have any meaning. The high school group, however, should be between the less-than-high-school group and the college group, or else the rank order will be incorrect.

Interval Measures

For the attributes composing some variables, the actual distance separating those attributes does have meaning. Such variables are **interval measures**. For these, the logical distance between attributes can be expressed in meaningful standard intervals.

For example, in the Fahrenheit temperature scale, the difference, or distance, between 80 degrees and 90 degrees is the same as that between 40 degrees and 50 degrees. However, 80 degrees Fahrenheit is not twice as hot as 40 degrees, because the zero point in the Fahrenheit scale is arbitrary; zero degrees does not really mean lack of heat. Similarly, minus 30 degrees on this scale doesn't represent 30 degrees less than no heat. (This is true for the Celsius scale as well. In contrast, the Kelvin scale is based on an absolute zero, which does mean a complete lack of heat.)

About the only interval measures commonly used in social science research are constructed measures such as standardized intelligence tests that have been more or less accepted. The interval separating IQ scores of 100 and 110 may be regarded as the same as the interval separating scores of 110 and 120 by virtue of the distribution of observed scores obtained by many thousands of people who have taken the tests over the years. But it would be incorrect to infer that someone with an IQ of 150 is 50 percent more intelligent than someone with an IQ of 100. (A person who received a score of 0 on a standard IQ test could not be regarded, strictly speaking, as having no intelligence, although we might feel he or she was unsuited to be a college professor or even a college student. But perhaps a dean . . . ?)

When comparing two people in terms of an interval variable, we can say they are different from each other (nominal), and that one is more than the other (ordinal). In addition, we can say "how much" more.

Ratio Measures

Most of the social science variables meeting the minimum requirements for interval measures also meet the requirements for ratio measures. In **ratio measures**, the attributes composing a variable, besides having all the structural characteristics mentioned previously, are based on a true zero point. The Kelvin temperature scale is one such measure. Examples from social science research include *age, length of residence in a given place, number of organizations belonged to, number of times attending religious services during a particular period of time, number of times married,* and *number of Arab friends.*

Returning to the illustration of methodological party games, we might ask a gathering of people to group themselves by age. All the one-year-olds would stand (or sit or lie) together, the

ordinal measure A level of measurement describing a variable with attributes we can rank-order along some dimension. An example is *socioeconomic status* as composed of the attributes *high, medium, low.*

interval measure A level of measurement describing a variable whose attributes are rank-ordered and have equal distances between adjacent attributes. The Fahrenheit temperature scale is an example of this, because the distance between 17 and 18 is the same as that between 89 and 90.

ratio measure A level of measurement describing a variable with attributes that have all the qualities of nominal, ordinal, and interval measures and in addition are based on a "true zero" point. *Age* is an example of a ratio measure.

two-year-olds together, the three-year-olds, and so forth. The fact that members of a single group share the same age and that each different group has a different shared age satisfies the minimum requirements for a nominal measure. Arranging the several groups in a line from youngest to oldest meets the additional requirements of an ordinal measure and lets us determine if one person is older than, younger than, or the same age as another. If we space the groups equally far apart, we satisfy the additional requirements of an interval measure and can say how much older one person is than another. Finally, because one of the attributes included in age represents a true zero (babies carried by women about to give birth), the phalanx of hapless partygoers also meets the requirements of a ratio measure, permitting us to say that one person is twice as old as another. (Remember this in case you're asked about it in a workbook assignment.) Another example of a ratio measure is *income*, which extends from an absolute zero to approximately infinity, if you happen to be the founder of Microsoft.

Comparing two people in terms of a ratio variable, then, allows us to conclude (1) whether they are different (or the same), (2) whether one is more than the other, (3) how much they differ, and (4) what the ratio of one to another is. Figure 5-1 summarizes this discussion by presenting a graphic illustration of the four levels of measurement.

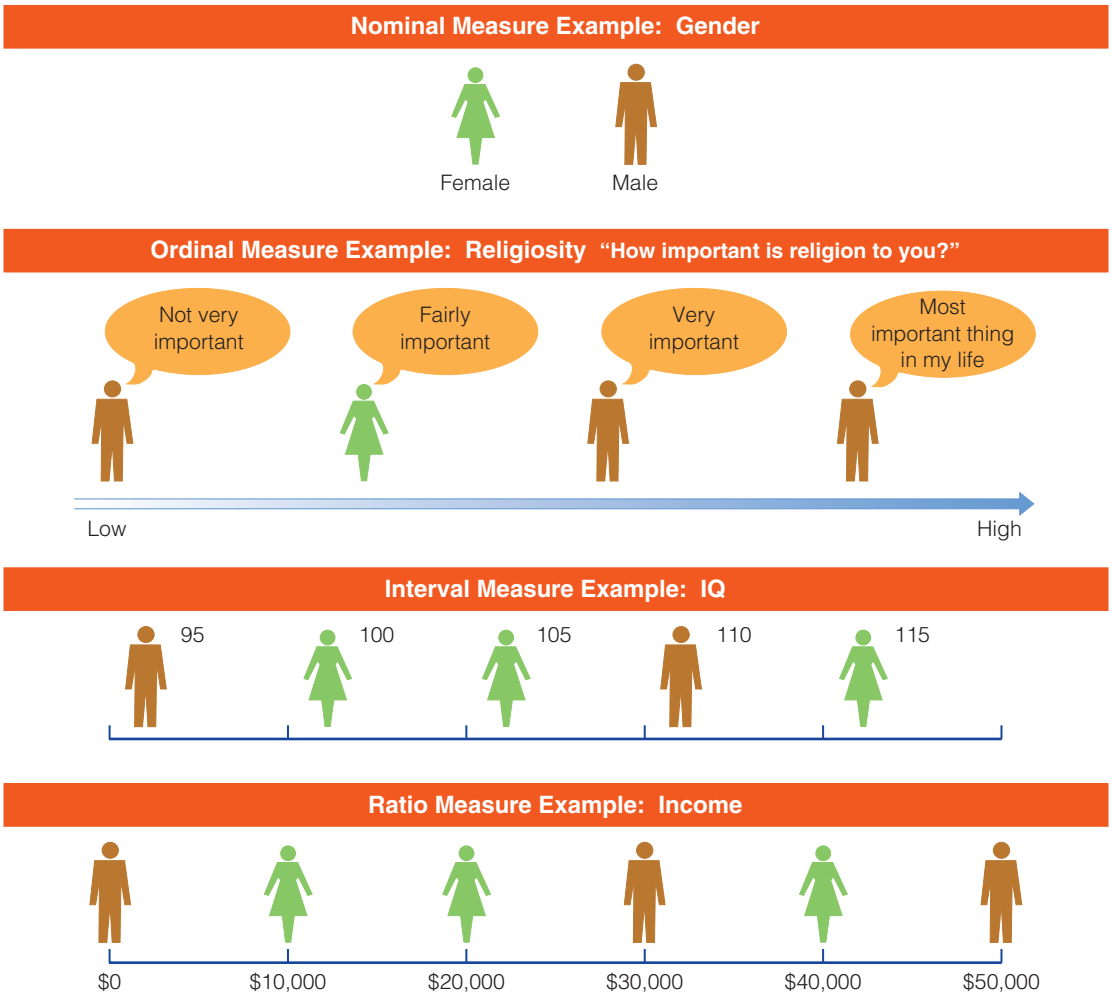

FIGURE 5-1

Levels of Measurement. Often you can choose among different levels of measurement—nominal, ordinal, interval, or ratio—carrying progressively more amounts of information.

Implications of Levels of Measurement

Because it's unlikely that you'll undertake the physical grouping of people just described (try it once, and you won't be invited to many parties), I should draw your attention to some of the practical implications of the differences that have been distinguished. These implications appear primarily in the analysis of data (discussed in Part 4), but you need to anticipate such implications when you're structuring any research project.

Certain quantitative analysis techniques require variables that meet certain minimum levels of measurement. To the extent that the variables to be examined in a research project are limited to a particular level of measurement—say, ordinal—you should plan your analytic techniques accordingly. More precisely, you should anticipate drawing research conclusions appropriate to the levels of measurement used in your variables. For example, you might reasonably plan to determine and report the mean age of a population under study (add up all the individual ages and divide by the number of people), but you should not plan to report the mean religious affiliation, because that is a nominal variable, and the mean requires ratio-level data. (You could report the modal—the most common—religious affiliation.)

At the same time, you can treat some variables as representing different levels of measurement. Ratio measures are the highest level, descending through interval and ordinal to nominal, the lowest level of measurement. A variable representing a higher level of measurement—say, ratio—can also be treated as representing a lower level of measurement—say, ordinal. Recall, for example, that age is a ratio measure. If you wished to examine only the relationship between age and some ordinal-level variable—say, *self-perceived religiosity*: high, medium, and low—you might choose to treat *age* as an ordinal-level variable as well. You might characterize the subjects of your study as being young, middle-aged, and old, specifying what age range composed each of these groupings. Finally, age might be used as a nominal-level variable for certain research purposes. People might be grouped as being born during the Iraq War or not. Another nominal measurement, based on birth date rather than just age, would be the grouping of people by astrological signs.

The level of measurement you'll seek, then, is determined by the analytic uses you've planned for a given variable, keeping in mind that some variables are inherently limited to a certain level. If a variable is to be used in a variety of ways, requiring different levels of measurement, the study should be designed to achieve the highest level required. For example, if the subjects in a study are asked their exact ages, they can later be organized into ordinal or nominal groupings.

Again, you need not necessarily measure variables at their highest level of measurement. If you're sure to have no need for ages of people at higher than the ordinal level of measurement, you may simply ask people to indicate their age range, such as 20 to 29, 30 to 39, and so forth. In a study of the wealth of corporations, rather than seek more-precise information, you may use Dun & Bradstreet ratings to rank corporations. Whenever your research purposes are not altogether clear, however, seek the highest level of measurement possible. As we've discussed, although ratio measures can later be reduced to ordinal ones, you cannot convert an ordinal measure to a ratio one. More generally, you cannot convert a lower-level measure to a higher-level one. That is a one-way street worth remembering.

The level of measurement is significant in terms of the arithmetic operations that can be applied to a variable and the statistical techniques using those operations. The accompanying table summarizes some of the implications, including ways of stating the comparison of two incomes.

Level of Measurement	*Arithmetic Operations*	*How to Express the Fact That Jan Earns $80,000 a Year and Andy Earns $40,000*
Nominal	$= \neq$	Jan and Andy earn *different* amounts.
Ordinal	$> <$	Jan earns *more* than Andy.
Interval	$+ -$	Jan earns *$40,000 more* than Andy.
Ratio	$\div \times$	Jan earns *twice* as much as Andy.

Typically a research project will tap variables at different levels of measurement. For example, William Bielby and Denise Bielby (1999) set out to examine the world of film and television,

using a nomothetic, longitudinal approach (take a moment to remind yourself what that means). In what they referred to as the "culture industry," the authors found that *reputation* (an ordinal variable) is the best predictor of screenwriters' future productivity. More interestingly, they found that screenwriters who were represented by "core" (or elite) agencies were not only far more likely to find jobs (a nominal variable), but also jobs that paid more (a ratio variable). In other words, the researchers found that agencies' reputations (ordinal) were a key independent variable for predicting a screenwriter's career success. The researchers also found that being older (ratio), female (nominal), an ethnic minority (nominal), and having more years of experience (ratio) were disadvantageous for a writer's career. On the other hand, higher earnings from previous years (measured in ordinal categories) led to more success in the future. In Bielby and Bielby's terms, "success breeds success" (1999: 80).

Single or Multiple Indicators

With so many alternatives for operationalizing social science variables, you may find yourself worrying about making the right choices. To counter this feeling, let me add a momentary dash of certainty and stability.

Many social research variables have fairly obvious, straightforward measures. No matter how you cut it, gender usually turns out to be a matter of male or female: a nominal-level variable that can be measured by a single observation—either by looking (well, not always) or by asking a question (usually). In a study involving the size of families, you'll want to think about adopted and foster children, as well as blended families, but it's usually pretty easy to find out how many children a family has. For most research purposes, the resident population of a country is the resident population of that country—you can look it up on the web and know the answer. A great many variables, then, have obvious single indicators. If you can get one piece of information, you have what you need.

Sometimes, however, there is no single indicator that will give you the measure of a variable you really want. As discussed earlier in this chapter, many concepts are subject to varying interpretations—each with several possible indicators. In these cases, you'll want to make several observations for a given variable. You can then combine the several pieces of information you've collected, creating a composite measurement of the variable in question. Chapter 6 is devoted to ways of doing that, so here let's just discuss one simple illustration.

Consider the concept "college performance." All of us have noticed that some students perform well in college courses and others don't. In studying these differences, we might ask what characteristics and experiences are related to high levels of performance (many researchers have done just that). How should we measure overall performance? Each grade in any single course is a potential indicator of college performance, but it also may not typify the student's general performance. The solution to this problem is so firmly established that it is, of course, obvious: the grade point average (GPA). We assign numerical scores to each letter grade, total the points earned by a given student, and divide by the number of courses taken, thus obtaining a composite measure. (If the courses vary in number of credits, we adjust the point values accordingly.) Creating such composite measures in social research is often appropriate.

Some Illustrations of Operationalization Choices

To bring together all the operationalization choices available to the social researcher and to show the potential in those possibilities, let's look at some of the distinct ways you might address various research problems. The alternative ways of operationalizing the variables in each case should demonstrate the opportunities that social research can present to our ingenuity and imaginations. To simplify matters, I have not attempted to describe all the research conditions that would make one alternative superior to the others, though in a given situation they would not all be equally appropriate.

Here are specific research questions, then, and some of the ways you could address them. We'll begin with an example discussed earlier in the chapter. It has the added advantage that one of the variables is straightforward to operationalize.

1. Are women more compassionate than men?
   a. Select a group of subjects for study, with equal numbers of men and women. Present them with hypothetical situations that involve someone's being in trouble. Ask them what they would do if they were confronted with that situation. What would they do, for example, if they came across a small child who was lost and crying for his or her parents? Consider any answer that involves helping or comforting the child as an indicator of compassion. See whether men or women are more likely to indicate they would be compassionate.
   b. Set up an experiment in which you pay a small child to pretend that he or she is lost. Put the child to work on a busy sidewalk and observe whether men or women are more likely to offer assistance. Also be sure to count the total number of men and women who walk by, because there may be more of one than the other. If that's the case, simply calculate the percentage of men and the percentage of women who help.
   c. Select a sample of people and do a survey in which you ask them what organizations they belong to. Calculate whether women or men are more likely to belong to those that seem to reflect compassionate feelings. To account for the case in which one group belongs to more organizations than the other does, do this: For each person you study, calculate the percentage of his or her organizational memberships that reflect compassion. See if men or women have a higher average percentage.
2. Are sociology students or accounting students better informed about world affairs?
   a. Prepare a short quiz on world affairs and arrange to administer it to the students in a sociology class and in an accounting class at a comparable level. If you want to compare sociology and accounting majors, be sure to ask students what they are majoring in.
   b. Get the instructor of a course in world affairs to give you the average grades of sociology and accounting students in the course.
   c. Take a petition to sociology and accounting classes that urges that "the United Nations headquarters be moved to New York City." Keep a count of how many in each class sign the petition and how many inform you that the UN headquarters is already located in New York City.
3. Who are the most popular instructors on your campus—those in the social sciences, the natural sciences, or the humanities?
   a. If your school has a provision for student evaluation of instructors, review some recent results and compute the average rating of each of the three groups.
   b. Begin visiting the introductory courses given in each group of disciplines and measure the attendance rate of each class.
   c. In December, select a group of faculty in each of the three divisions and ask them to keep a record of the numbers of holiday greeting cards and presents they receive from admiring students. See who wins.

The point of these examples is not necessarily to suggest respectable research projects but to illustrate the many ways variables can be operationalized. The Research in Real Life box, "Measuring College Satisfaction," briefly overviews the preceding steps in terms of a concept mentioned at the outset of this chapter.

Operationalization Goes On and On

Although I've discussed conceptualization and operationalization as activities that precede data collection and analysis—for example, you must design questionnaire items before you send out a questionnaire—these two processes continue throughout any research project, even if the data have been collected in a structured mass survey. As we've seen, in less-structured methods such as field research, the identification and specification of relevant concepts is inseparable from the ongoing process of observation.

Imagine, for example, that you're doing a qualitative, observational study of members of a new religious cult, and, in part, you want to identify those members who are more religious and those who are less religious. You may begin with a focus on certain kinds of ritual behavior, only to eventually discover that the members of the group place a higher premium on religious experience or steadfast beliefs.

Research in Real Life

Measuring College Satisfaction

Early in this chapter, we considered "college satisfaction" as an example of a concept people often talk about casually. To study such a concept, however, we need to engage in the processes of conceptualization and operationalization. I'll sketch out the process briefly, then you might try your hand at expanding on my comments.

What are some of the dimensions of college satisfaction? Here are a few to get you started, but feel free to add your own:

Academic quality: faculty, courses, majors
Physical facilities: classrooms, dorms, cafeteria, grounds
Athletics and extracurricular activities
Costs and availability of financial aid
Sociability of students, faculty, staff
Security, crime on campus

How would you measure each of these dimensions? One method would be to ask a sample of students, "How would you rate your level of satisfaction with each of the following?," to give them a list of items similar to those listed here, and to provide a set of categories for them to use (such as very satisfied, satisfied, dissatisfied, very dissatisfied).

But suppose you didn't have the time and/or money to conduct a survey and were interested in comparing overall levels of satisfaction at several schools. What data about schools (the unit of analysis) might give you the answer you were interested in? Retention rates might be one general indicator. Can you think of others?

Notice that you can measure college quality both positively and negatively. Modern classrooms with Wi-Fi access would count positively, whereas the number of crimes on campus would count negatively. But the latter could be used as a measure of college quality: with low crime rates counting as high quality.

The open-endedness of conceptualization and operationalization is perhaps more obvious in qualitative than in quantitative research, since changes can be made at any point during data collection and analysis. In quantitative methods such as survey research or experiments, you will be required to commit yourself to particular measurement structures. Once a questionnaire has been printed and administered, for example, altering it would be impractical if not impossible, even when the unfolding of the research might suggest changes. Even in the case of a survey questionnaire, however, you may have some flexibility in how you measure variables during the analysis phase, as we'll see in the following chapter.

As I mentioned, however, the qualitative researcher has a greater flexibility in this regard. Things you notice during in-depth interviews, for example, may suggest a different set of questions than you initially planned, allowing you to pursue unanticipated avenues. Then later, as you review and organize your notes for analysis, you may again see unanticipated patterns and redirect your analysis.

Regardless of whether you are using qualitative or quantitative methods, you should always be open to reexamining your concepts and definitions. The ultimate purpose of social research is to clarify the nature of social life. The validity and utility of what you learn in this regard doesn't depend on when you first figured out how to look at things any more than it matters whether you got the idea from a learned textbook, a dream, or your brother-in-law.

Criteria of Measurement Quality

This chapter has come some distance. It began with the bald assertion that social scientists can measure anything that exists. Then we discovered that most of the things we might want to measure and study don't really exist. Next we learned that it's possible to measure them anyway. Now we'll discuss of some of the yardsticks against which we judge our relative success or failure in measuring things—even things that don't exist.

Precision and Accuracy

To begin, measurements can be made with varying degrees of precision. As we saw in the discussion of operationalization, precision concerns the fineness of distinctions made between the attributes that compose a variable. The description of a woman as "43 years old" is more precise than "in her forties." Saying a street-corner gang was formed "in the summer of 1996" is more precise than saying "during the 1990s."

As a general rule, precise measurements are superior to imprecise ones, as common sense dictates. There are no conditions under which imprecise measurements are intrinsically superior to precise ones. Even so, exact precision is not always necessary or desirable. If knowing that a woman is in her forties satisfies your research requirements, then any additional effort invested in learning her precise age is wasted. The operationalization of concepts, then, must be guided partly by an understanding of the degree of precision required. If your needs are not clear, be more precise rather than less.

Don't confuse precision or specificity with accuracy, however. Describing someone as "born in New England" is less specific than "born in Stowe, Vermont"—but suppose the person in question was actually born in Boston. The less-specific description, in this instance, is more accurate, a better reflection of the real world.

Precision and accuracy are obviously important qualities in research measurement, and they probably need no further explanation. When social scientists construct and evaluate measurements, however, they pay special attention to two technical considerations: reliability and validity.

Reliability

In the abstract, **reliability** is a matter of whether a particular technique, applied repeatedly to the same object, yields the same result each time. Let's say you want to know how much I weigh. (No, I don't know why.) As one technique, say you ask two different people to estimate my weight. If the first person estimates 150 pounds and the other estimates 300, we have to conclude the technique of having people estimate my weight isn't very reliable.

Suppose, as an alternative, that you use a bathroom scale as your measurement technique. I step on the scale twice, and you note the same result each time. The scale has presumably reported the same weight for me both times, indicating that the scale provides a more reliable technique for measuring a person's weight than asking people to estimate it does.

Reliability, however, does not ensure accuracy any more than precision does. Suppose I've set my bathroom scale to shave five pounds off my weight just to make me feel better. Although you would (reliably) report the same weight for me each time, you would always be wrong. This new element, called *bias,* is discussed in Chapter 8. For now, just be warned that reliability does not ensure accuracy.

Let's suppose we're interested in studying morale among factory workers in two different kinds of factories. In one set of factories, workers have specialized jobs, reflecting an extreme division of labor. Each worker contributes a tiny part to the overall process performed on a long assembly line. In the other set of factories, each worker performs many tasks, and small teams of workers complete the whole process.

How should we measure morale? Following one strategy, we could observe the workers in each factory, noticing such things as whether they joke with one another, whether they smile and laugh a lot, and so forth. We could ask them how they like their work and even ask them whether they think they would prefer their current arrangement or the other one being studied. By comparing what we observed in the different factories, we might reach a conclusion about which assembly process produces the higher morale. Notice that I've just described a qualitative measurement procedure.

Now let's look at some reliability problems inherent in this method. First, how you and I are feeling when we do the observing will likely color what we see. We may misinterpret what we see. We may see workers kidding each other but think they're having an argument. We may catch them on an off day. If we were to observe the same group of workers several days in a row, we might arrive at different evaluations on each day. Further, even if several observers evaluated the same behavior, they might arrive at different conclusions about the workers' morale.

Here's another strategy for assessing morale, a quantitative approach. Suppose we check the company records to see how many grievances have been filed with the union during some

reliability That quality of measurement method that suggests that the same data would have been collected each time in repeated observations of the same phenomenon. In the context of a survey, we would expect that the question "Did you attend religious services last week?" would have higher reliability than the question "About how many times have you attended religious services in your life?" This is not to be confused with validity.

fixed period. Presumably this would be an indicator of morale: the more grievances, the lower the morale. This measurement strategy would appear to be more reliable: Counting up the grievances over and over, we should keep arriving at the same number.

If you find yourself thinking that the number of grievances doesn't necessarily measure morale, you're worrying about validity, not reliability. We'll discuss validity in a moment. The point for now is that the last method is more like my bathroom scale—it gives consistent results.

In social research, reliability problems crop up in many forms. Reliability is a concern every time a single observer is the source of data, because we have no certain guard against the impact of that observer's subjectivity. We can't tell for sure how much of what's reported originated in the situation observed and how much in the observer.

Subjectivity is not only a problem with single observers, however. Survey researchers have known for a long time that different interviewers, because of their own attitudes and demeanors, get different answers from respondents. Or, if we were to conduct a study of newspapers' editorial positions on some public issue, we might create a team of coders to take on the job of reading hundreds of editorials and classifying them in terms of their position on the issue. Unfortunately, different coders will code the same editorial differently. Or we might want to classify a few hundred specific occupations in terms of some standard coding scheme, say a set of categories created by the Department of Labor or by the Census Bureau. You and I would not place all those occupations in the same categories.

Each of these examples illustrates problems of reliability. Similar problems arise whenever we ask people to give us information about themselves. Sometimes we ask questions that people don't know the answers to: "How many times have you been to religious services?" Sometimes we ask people about things they consider totally irrelevant: "Are you satisfied with China's current relationship with Albania?" In such cases, people will answer differently at different times because they're making up answers as they go. Sometimes we explore issues so complicated that a person who had a clear opinion in the matter might arrive at a different interpretation of the question when asked a second time.

So how do you create reliable measures? If your research design calls for asking people for information, you can be careful to ask only about things the respondents are likely to know the answer to. Ask about things relevant to them, and be clear in what you're asking. Of course, these techniques don't solve every possible reliability problem. Fortunately, social researchers have developed several techniques for cross-checking the reliability of the measures they devise.

Test-Retest Method

Sometimes it's appropriate to make the same measurement more than once, a technique called the *test-retest method*. If you don't expect the sought-after information to change, then you should expect the same response both times. If answers vary, the measurement method may, to the extent of that variation, be unreliable. Here's an illustration.

In their classic research on Health Hazard Appraisal (HHA), a part of preventive medicine, Jeffrey Sacks, W. Mark Krushat, and Jeffrey Newman (1980) wanted to determine the risks associated with various background and lifestyle factors, making it possible for physicians to counsel their patients appropriately. By knowing patients' life situations, physicians could advise them on their potential for survival and on how to improve it. This purpose, of course, depended heavily on the accuracy of the information gathered about each subject in the study.

To test the reliability of their information, Sacks and his colleagues had all 207 subjects complete a baseline questionnaire that asked about their characteristics and behavior. Three months later, a follow-up questionnaire asked the same subjects for the same information, and the results of the two surveys were compared. Overall, only 15 percent of the subjects reported the same information in both studies.

Sacks and his colleagues report the following:

> Almost 10 percent of subjects reported a different height at follow-up examination. Parental age was changed by over one in three subjects. One parent reportedly aged 20 chronologic years in three months. One in five ex-smokers and ex-drinkers have apparent difficulty in reliably recalling their previous consumption pattern.
>
> *(1980: 730)*

Some subjects erased all trace of previously reported heart murmur, diabetes, emphysema, arrest record, and thoughts of suicide. One subject's mother, deceased in the first questionnaire, was apparently alive and well in time for the second. One subject had one ovary missing in the first study but present in the second. In another case, an ovary present in the first study was missing in the second study—and had been for ten years! One subject was reportedly 55 years old in the first study and 50 years old three months later. (You have to wonder whether the physician-counselors could ever have nearly the impact on their patients that their patients' memories did.) Thus, test-retest revealed that this data-collection method was not especially reliable.

Split-Half Method

As a general rule, it's always good to make more than one measurement of any subtle or complex social concept, such as prejudice, alienation, or social class. This procedure lays the groundwork for another check on reliability. Let's say you've created a questionnaire that contains ten items you believe measure prejudice against women. Using the split-half technique, you would randomly assign those ten items to two sets of five. Each set should provide a good measure of prejudice against women, and the two sets should classify respondents the same way. If the two sets of items classify people differently, you most likely have a problem of reliability in your measure of the variable.

Using Established Measures

Another way to help ensure reliability in getting information from people is to use measures that have proved their reliability in previous research. If you want to measure anomia, for example, you might want to follow Srole's lead.

The heavy use of measures, though, does not guarantee their reliability. For example, the Scholastic Assessment Tests (SATs) and the Minnesota Multiphasic Personality Inventory (MMPI) have been accepted as established standards in their respective domains for decades. In recent years, though, they've needed fundamental overhauling to reflect changes in society, eliminating outdated topics and gender bias in wording.

Reliability of Research Workers

As we've seen, it's also possible for measurement unreliability to be generated by research workers: interviewers and coders, for example. There are several ways to check on reliability in such cases. To guard against interviewer unreliability in surveys, for example, a supervisor will call a subsample of the respondents on the telephone and verify selected pieces of information.

Replication works in other situations also. If you're worried that newspaper editorials or occupations may not be classified reliably, you could have each independently coded by several coders. Those cases that are classified inconsistently can then be evaluated more carefully and resolved.

Finally, clarity, specificity, training, and practice can prevent a great deal of unreliability and grief. If you and I spent some time reaching a clear agreement on how to evaluate editorial positions on an issue—discussing various positions and reading through several together—we could probably do a good job of classifying them in the same way independently.

The reliability of measurements is a fundamental issue in social research, and we'll return to it more than once in the chapters ahead. For now, however, let's recall that even total reliability doesn't ensure that our measures actually measure what we think they measure. Now let's plunge into the question of validity.

validity A term describing a measure that accurately reflects the concept it is intended to measure. For example, your IQ would seem a more valid measure of your intelligence than the number of hours you spend in the library would. Though the ultimate validity of a measure can never be proved, we may agree to its relative validity on the basis of face validity, criterion-related validity, construct validity, content validity, internal validation, and external validation (see Chapter 6). This must not be confused with reliability.

Validity

In conventional usage, **validity** refers to the extent to which an empirical measure adequately reflects the real meaning of the concept under consideration. A measure of social class should

measure social class, not political orientations. A measure of political orientations should measure political orientations, not sexual permissiveness. Validity means that we are actually measuring what we say we are measuring.

Whoops! I've already committed us to the view that concepts don't have real meanings. How can we ever say whether a particular measure adequately reflects the concept's meaning, then? Ultimately, of course, we can't. At the same time, as we've already seen, all of social life, including social research, operates on agreements about the terms we use and the concepts they represent. There are several criteria of success in making measurements that are appropriate to these agreed-on meanings of concepts.

First, there's something called **face validity**. Particular empirical measures may or may not jibe with our common agreements and our individual mental images concerning a particular concept. For example, you and I might quarrel about whether counting the number of grievances filed with the union will adequately measure morale. Still, we'd surely agree that the number of grievances has *something* to do with morale. That is, the measure is valid "on its face," whether or not it's adequate. If I were to suggest that we measure morale by finding out how many books the workers took out of the library during their off-duty hours, you'd undoubtedly raise a more serious objection: That measure wouldn't have much face validity.

Second, I've already pointed to many of the more formally established agreements that define some concepts. The Census Bureau, for example, has created operational definitions of such concepts as family, household, and employment status that seem to have a workable validity in most studies using these concepts.

Three additional types of validity also specify particular ways of testing the validity of measures. The first, **criterion-related validity**, sometimes called *predictive validity,* is based on some external criterion. For example, the validity of College Board exams is shown in their ability to predict students' success in college. The validity of a written driver's test is determined, in this sense, by the relationship between the scores people get on the test and their subsequent driving records. In these examples, college success and driving ability are the criteria.

To test your understanding of criterion-related validity, see whether you can think of behaviors that might be used to validate each of the following attitudes:

Is very religious

Supports equality of men and women

Supports far-right militia groups

Is concerned about the environment

Some possible validators would be, respectively, attends religious services, votes for women candidates, belongs to the NRA, and belongs to the Sierra Club.

Sometimes it's difficult to find behavioral criteria that can be taken to validate measures as directly as in such examples. In those instances, however, we can often approximate such criteria by applying a different test. We can consider how the variable in question ought, theoretically, to relate to other variables. **Construct validity** is based on the logical relationships among variables.

Suppose, for example, that you want to study the sources and consequences of marital satisfaction. As part of your research, you develop a measure of marital satisfaction, and you want to assess its validity.

In addition to developing your measure, you'll have developed certain theoretical expectations about the way the variable *marital satisfaction* relates to other variables. For example, you might reasonably conclude that satisfied husbands and wives will be less likely than dissatisfied ones to cheat on their spouses. If your measure relates to marital fidelity in the expected fashion, that constitutes evidence of

face validity That quality of an indicator that makes it seem a reasonable measure of some variable. That the frequency of attendance at religious services is some indication of a person's religiosity seems to make sense without a lot of explanation. It has face validity.

criterion-related validity The degree to which a measure relates to some external criterion. For example, the validity of College Board tests is shown in their ability to predict the college success of students. Also called predictive validity.

construct validity The degree to which a measure relates to other variables as expected within a system of theoretical relationships.

Tips and Tools

Validity and Social Desirability

A particular challenge in measurement occurs when the attitudes or behaviors being asked about are generally considered socially undesirable—regardless of whether the report involves an actual crime or something more harmless, like not voting. (A case in point, postelection surveys show a higher percentage of respondents *reporting* they voted than actually was tallied on election day.)

One technique has been to downplay the deviance involved, but this in itself has unexpected consequences. For example, the format of asking "Do you feel the president should do *X* or do you feel the president should do *Y*?" may seem a little too blunt, and researchers have sought to soften it by prefacing the question with "Some people feel the president should do *X*, while others feel the president should do *Y*." Experiments by David Scott Yeager and Jon Krosnick (2011, 2012) utilizing both face-to-face and Internet surveys suggest this is not a useful variation. It may affect respondents' assumptions about how others feel, but it does not seem to improve reports of respondents' own opinions. In fact, where independent checks on attitudes and behaviors were possible, the "some"/"other" format *reduced* the validity of reports. Adolescents, for example, tended to report more deviant behavior than they had actually done. As a bottom line, the "softened" format requires more words (and time) and makes the questions more complicated without adding to the validity of responses.

Sources: David Scott Yeager and Jon A. Krosnick. (2012). "Does Mentioning 'Some People' and 'Other People' in an Opinion Question Improve Measurement Quality?" *Public Opinion Quarterly* 76 (1): 131–41; David Scott Yeager and Jon A. Krosnick. (2011). "Does Mentioning 'Some People' and 'Other People' in a Survey Question Increase the Accuracy of Adolescents' Self-Reports?" *Developmental Psychology* 47 (6): 1674–9. doi:10.1037/a0025440.

your measure's construct validity. If satisfied marriage partners are as likely to cheat on their spouses as the dissatisfied ones are, however, that would challenge the validity of your measure.

Tests of construct validity, then, can offer a weight of evidence that your measure either does or doesn't tap the quality you want it to measure, without providing definitive proof. Although I have suggested that tests of construct validity are less compelling than those of criterion validity, there is room for disagreement about which kind of test a particular comparison variable (*driving record, marital fidelity*) represents in a given situation. It's less important to distinguish the two types of validity tests than to understand the logic of validation that they have in common: If we've succeeded in measuring some variable, then our measures should relate in some logical way to other measures.

Finally, **content validity** refers to how much a measure covers the range of meanings included within a concept. For example, a test of mathematical ability cannot be limited to addition but also needs to cover subtraction, multiplication, division, and so forth. Or, if we're measuring prejudice, do our measurements reflect all types of prejudice, including prejudice against racial and ethnic groups, religious minorities, women, the elderly, and so on?

content validity The degree to which a measure covers the range of meanings included within a concept.

The Tips and Tools box, "Validity and Social Desirability," examines the special challenges involved in asking people to report deviant attitudes or behaviors.

Figure 5-2 presents a graphic portrayal of the difference between validity and reliability. If you think of measurement as analogous to repeatedly shooting at the bull's-eye on a target, you'll see that reliability looks like a "tight pattern," regardless of where the shots hit, because reliability is a function of consistency. Validity, on the other hand, is a function of shots being arranged around the bull's-eye. The failure of reliability in the figure is randomly distributed around the target; the failure of validity is systematically off the mark. Notice that neither an unreliable nor an invalid measure is likely to be very useful.

Who Decides What's Valid?

Our discussion of validity began with a reminder that we depend on agreements to determine what's real, and we've just seen some of the ways social scientists can agree among themselves that they have made valid measurements. There is yet another way of looking at validity.

Social researchers sometimes criticize themselves and one another for implicitly assuming

FIGURE 5-2

An Analogy to Validity and Reliability. A good measurement technique should be both valid (measuring what it is intended to measure) and reliable (yielding a given measurement dependably).

they are somewhat superior to those they study. For example, researchers often seek to uncover motivations that the social actors themselves are unaware of. You think you bought that new Turbo Tiger because of its high performance and good looks, but *we* know you're really trying to achieve a higher social status.

This implicit sense of superiority would fit comfortably with a totally positivistic approach (the biologist feels superior to the frog on the lab table), but it clashes with the more humanistic and typically qualitative approach taken by many social scientists. We'll explore this issue more deeply in Chapter 10. In seeking to understand the way ordinary people make sense of their worlds, ethnomethodologists have urged all social scientists to pay more respect to the natural social processes of conceptualization and shared meaning. At the very least, behavior that may seem irrational from the scientist's paradigm may make logical sense when viewed through the actor's paradigm.

Clifford Geertz (1973) applies the term *thick description* in reference to the goal of understanding, as deeply as possible, the meanings that elements of a culture have for those who live within that culture. He recognizes that the outside observer will never grasp those meanings fully, however, and warns, "Cultural analysis is intrinsically incomplete." He then elaborates:

> There are a number of ways to escape this—turning culture into folklore and collecting it, turning it into traits and counting it, turning it into institutions and classifying it, turning it into structures and toying with it. But they are escapes. The fact is that to commit oneself to a semiotic concept of culture and an interpretive approach to the study of it is to commit oneself to a view of ethnographic assertion as, to borrow W. B. Gallie's by now famous phrase, "essentially contestable." Anthropology, or at least interpretive anthropology, is a science whose progress is marked less by a perfection of consensus than by a refinement of debate. What gets better is the precision with which we vex each other.
>
> *(1973: 29)*

Ultimately, social researchers should look both to their colleagues and to their subjects as sources of agreement on the most useful meanings and measurements of the concepts they study. Sometimes one source will be more useful, sometimes the other. But neither one should be dismissed.

Tension between Reliability and Validity

Clearly, we want our measures to be both reliable and valid. However, a tension often arises between the criteria of reliability and validity, forcing a trade-off between the two.

Recall the example of measuring morale in different factories. The strategy of immersing yourself in the day-to-day routine of the assembly line, observing what goes on, and talking to the workers would seem to provide a more valid measure of morale than counting grievances would. It just seems obvious that we'd get a clearer sense of whether the morale was high or low using this first method.

As I pointed out earlier, however, the counting strategy would be more reliable. This situation reflects a more general strain in research measurement. Most of the really interesting concepts we want to study have many subtle nuances, so specifying precisely what we mean by them is hard. Researchers sometimes speak of such concepts as having a "richness of meaning." Although scores of books and articles have been written on the topic of anomie/anomia, for example, they still haven't exhausted its meaning.

Very often, then, specifying reliable operational definitions and measurements seems to rob concepts of their richness of meaning. Positive morale is much more than a lack of grievances filed with the union; anomia is much more than what is measured by the five items created by Leo Srole. Yet, the more variation and richness we allow for a concept, the more opportunity there is for disagreement on how it applies to a particular situation, thus reducing reliability.

To some extent, this dilemma explains the persistence of two quite different approaches to social research: quantitative, nomothetic, structured techniques such as surveys and experiments on the one hand, and qualitative, idiographic methods such as field research and historical studies on the other. In the simplest generalization, the former methods tend to be more reliable, the latter more valid.

By being forewarned, you'll be effectively forearmed against this persistent and inevitable dilemma. If there is no clear agreement on how to measure a concept, measure it several different ways. If the concept has several dimensions, measure them all. Above all, know that the concept does not have any meaning other than what you and I give it. The only justification for giving any concept a particular meaning is utility. Measure concepts in ways that help us understand the world around us.

The Ethics of Measurement

Measurement decisions can sometimes be judged by ethical standards. We have seen that most of the concepts of interest to social researchers are open to varied meanings. Suppose, for example, that you are interested in sampling public opinion on the abortion issue in the United States. Notice the difference it would make if you conceptualized one side of the debate as "pro-choice" or as "pro-abortion." If your personal bias made you want to minimize support for having an abortion, you might be tempted to frame the concept and the measurements based on it in terms of people being "pro-abortion," thereby eliminating all those who were not especially fond of abortion per se but felt a woman should have the right to make that choice for herself. To pursue this strategy, however, would violate accepted research ethics.

Consider the choices available to you in conceptualizing attitudes toward the U.S. invasion of Iraq in 2003. Imagine the different levels of support you would "discover" if you framed the position as an unprovoked invasion of a sovereign nation, as a retaliation for the September 11, 2001, attack on the World Trade Towers (many Americans still believe Saddam Hussein masterminded that attack), as a defensive act against a perceived threat, as part of a global war on terrorism, or in any of the other ways this event has been portrayed. There is no one, correct way to conceptualize this issue, but it would be unethical to seek to slant the results through a biased definition of the issue.

MAIN POINTS

Introduction

- The interrelated processes of conceptualization, operationalization, and measurement allow researchers to move from a general idea about what they want to study to effective and well-defined measurements in the real world.

Measuring Anything That Exists

- Conceptions are mental images we use as summary devices for bringing together observations and experiences that seem to have something in common. We use terms or labels to reference these conceptions.
- Concepts are constructs; they represent the agreed-on meanings we assign to terms. Our concepts don't exist in the real world, so they can't be measured directly, but we can measure the things that our concepts summarize.

Conceptualization

- Conceptualization is the process of specifying observations and measurements that give

concepts definite meaning for the purposes of a research study.

- Conceptualization includes specifying the indicators of a concept and describing its dimensions. Operational definitions specify how variables relevant to a concept will be measured.

Definitions in Descriptive and Explanatory Studies

- Precise definitions are even more important in descriptive than in explanatory studies. The degree of precision needed varies with the type and purpose of a study.

Operationalization Choices

- Operationalization is an extension of conceptualization that specifies the exact procedures that will be used to measure the attributes of variables.
- Operationalization involves a series of interrelated choices: specifying the range of variation that is appropriate for the purposes of a study, determining how precisely to measure variables, accounting for relevant dimensions of variables, clearly defining the attributes of variables and their relationships, and deciding on an appropriate level of measurement.
- Researchers must choose from four levels of measurement, which capture increasing amounts of information: nominal, ordinal, interval, and ratio. The most appropriate level depends on the purpose of the measurement.
- A given variable can sometimes be measured at different levels. When in doubt, researchers should use the highest level of measurement appropriate to that variable so they can capture the greatest amount of information.
- Operationalization begins in the design phase of a study and continues through all phases of the research project, including the analysis of data.

Criteria of Measurement Quality

- Criteria of the quality of measures include precision, accuracy, reliability, and validity.
- Whereas reliability means getting consistent results from the same measure, validity refers to getting results that accurately reflect the concept being measured.
- Researchers can test or improve the reliability of measures through the test-retest method, the split-half method, the use of established measures, and the examination of work performed by research workers.
- The yardsticks for assessing a measure's validity include face validity, criterion-related validity, construct validity, and content validity.
- Creating specific, reliable measures often seems to diminish the richness of meaning our general concepts have. This problem is inevitable. The best solution is to use several different measures, tapping the different aspects of a concept.

The Ethics of Measurement

- Conceptualization and measurement must never be guided by bias or preferences for particular research outcomes.

KEY TERMS

The following terms are defined in context in the chapter and at the bottom of the page where the term is introduced, as well as in the comprehensive glossary at the back of the book.

conceptualization
construct validity
content validity
criterion-related validity
dimension
face validity
indicator
interval measure
nominal measure
ordinal measure
ratio measure
reliability
specification
validity

PROPOSING SOCIAL RESEARCH: MEASUREMENT

This chapter has taken us deeper into the matter of measurement. In previous exercises, you've identified the concepts and variables you want to address in your research project. Now you'll need to get more specific in terms of conceptualization and operationalization. You should conclude this portion of the proposal with a description of how, precisely, you will make distinctions regarding your variables. If you want to compare liberals and conservatives, for example, how exactly will you identify subjects' political orientations?

The ease or difficulty of this exercise may vary with the type of data collection you're planning. It will probably be easier to accomplish in the case of quantitative studies, such as surveys, where you can report the questionnaire items you'll use for measurements. In qualitative research, however, you'll have more opportunities to modify the ways variables are measured as the study unfolds, taking advantage of insights gained "in the trenches." Even so, you'll still need to begin with some clear ideas about how you'll begin your measurements.

Criteria such as precision, accuracy, validity, and reliability matter greatly in all kinds of social research projects.

REVIEW QUESTIONS AND EXERCISES

1. Pick a social science concept such as liberalism or alienation, then specify that concept so that it could be studied in a research project. Be sure to specify the indicators you'll use as well as the dimensions you wish to include in and exclude from your conceptualization.
2. What level of measurement—nominal, ordinal, interval, or ratio—describes each of the following variables?
   a. Race (white, African American, Asian, and so on)
   b. Order of finish in a race (first, second, third, and so on)
   c. Number of children in families
   d. Populations of nations
   e. Attitudes toward nuclear energy (strongly approve, approve, disapprove, strongly disapprove)
   f. Region of birth (Northeast, Midwest, and so on)
   g. Political orientation (very liberal, somewhat liberal, somewhat conservative, very conservative)
3. To conceptualize the variable *prejudice,* use your favorite web browser to search for this term. After reviewing several of the websites resulting from your search, make a list of some different forms of prejudice that might be studied in an omnibus project dealing with that topic.
4. In a dictionary, look up *truth* and *true,* then copy out the definitions. Note the key terms used in those definitions (such as *reality*), look up the definitions of those terms, and copy out these definitions as well. Continue this process until no new terms appear. Comment on what you've learned from this exercise. Did you discover "truth"?