5e

The ESSENTIALS of Political Analysis

PHILIP H. POLLOCK III

1

The Definition and Measurement of Concepts

LEARNING OBJECTIVES

In this chapter you will learn:

- How to clarify the meaning of concepts
- How to identify multidimensional concepts
- How to write a definition for a concept
- How systematic error affects the measurement of a concept
- How random error affects the measurement of a concept
- How to recognize problems of reliability and validity

Think for a moment about all the political variety in the world. People vary in their party affiliation: Some are Democrats, some Republicans, and many (self-described Independents) profess no affiliation at all. Some nations are democracies, whereas others are not. Even among democracies there is variety: parliamentary systems, presidential systems, or a combination. Would-be presidential nominees run the ideological gamut from conservatism to liberalism. Each of the terms just mentioned—*party affiliation, democracy, conservatism, liberalism*—refers to an idea that helps us discuss and describe the world. It is virtually impossible to converse about politics without using ideas such as these. Ideas, of course, are not concrete. You cannot see, taste, hear, touch, or smell "partisanship," "democracy," or "liberalism." Each of these is a **concept**, an idea or mental construct that represents phenomena in the real world. Some concepts are quite complicated: "globalization," "power," "democratization." Others, such as "political participation" or "social status," are somewhat simpler.

Simple or complicated, concepts are everywhere in political debate, in journalistic analysis, in ordinary discussion, and, of course, in political research. How are concepts used? In partisan or ideological debate—debates about values—concepts can evoke powerful symbols with which people easily identify. A political candidate, for example, might claim

that his or her agenda will ensure "freedom," create "equality," or foster "self-determination" around the globe. These are evocative ideas, and they are meant to be. In political research, concepts are not used to stir up value-laden symbols. Quite the opposite. In empirical political science, concepts refer to facts, not values. So when political researchers discuss ideas like "freedom," "equality," or "self-determination," they are using these ideas to summarize and label observable phenomena, characteristics in the real world.

The primary goals of political research are to describe concepts and to analyze the relationships between them. A researcher may want to know, for example, if social trust is declining or increasing in the United States, whether political elites are more tolerant of dissent than are ordinary citizens, or whether economic development causes democracy. The tasks of describing and analyzing concepts—social trust, political elites, tolerance of dissent, economic development, democracy, and any other concepts that interest us—present formidable obstacles. A **conceptual question**, a question expressed using ideas, is frequently unclear and thus is difficult to answer empirically. A **concrete question**, a question expressed using tangible properties, can be answered empirically. In her path-breaking book, *The Concept of Representation,* Hanna Pitkin describes the challenge of defining concepts such as "representation," "power," or "interest." She writes that instances "of representation (or of power, or of interest) . . . can be observed, but the observation always presupposes at least a rudimentary conception of what representation (or power, or interest) *is,* what *counts as* representation, where it leaves off and some other phenomenon begins."[1] We need to somehow transform concepts into concrete terms, to express vague ideas in such a way that they can be described and analyzed.

The same concept can, and often does, refer to a variety of different concrete terms. "Are women more liberal than men?" What is the answer: yes or no? "It depends," you might say, "on what you mean by *liberal.* Do you mean to ask if women are more likely than men to support abortion rights, gun control, government support of education, spending to assist poor people, environmental protection, affirmative action, gay and lesbian rights, funding for drug rehabilitation, or what? Do you mean all these things, some of these things, none of these things, or completely different things?" "Liberal," for some, may mean support for gun control. For others, the concept might refer to support for environmental protection. Still others might think the real meaning of liberalism is support for government spending to assist the poor.

A **conceptual definition** clearly describes the concept's measurable properties and specifies the units of analysis (people, nations, states, and so on) to which the concept applies. For example, consider the following conceptual definition of liberalism: Liberalism is the extent to which individuals support increased government spending for social programs. This statement clarifies a vague idea, liberalism, by making reference to a measurable attribute—support for government spending. Notice the words, "the extent to which." This phrase suggests that the concept's measurable attribute—support for government spending—varies across people. Someone who supports government spending has "more liberalism" than someone who does not support government spending. It is clear, as well, that this particular definition is meant to apply to individuals.[2] As you can see, in thinking about concepts and

defining them, we keep an eye trained on the empirical world: What are the concrete, measurable characteristics of this concept? Conceptual definitions are covered in depth in the first part of this chapter.

Having clarified and defined a concept, we must then describe an instrument for measuring the concept in the real world. An **operational definition** describes the instrument to be used in measuring the concept and putting a conceptual definition "into operation." How might we go about implementing the conceptual definition of liberalism? Imagine crafting a series of ten or twelve survey questions and administering them to a large number of individuals. Each question would name a specific social program: funding for education, assistance to the poor, spending on medical care, support for childcare subsidies, and so on. For each program, individuals would be asked whether government spending should be decreased, kept the same, or increased. Liberalism could then be operationally defined as the number of times a respondent said "increased." Higher scores would denote more liberal attitudes and lower scores would denote less liberal attitudes. As this example suggests, an operational definition provides a procedural blueprint, a measurement strategy. Yet, in describing a measurement strategy, we keep an eye trained on the conceptual world: Does this operational definition accurately reflect the meaning of the concept? In this chapter we consider problems that can emerge when researchers decide on an operational definition. In Chapter 2 we take a closer look at variables, the concrete measurements of concepts.

CONCEPTUAL DEFINITIONS

The first step in defining a concept is to clarify its empirical meaning. To clarify a concept, we begin by making an inventory of the concept's concrete properties. Three problems often arise during the inventory-building process. First, we might think of empirical attributes that refer to a completely different concept. Second, the inventory may include conceptual terms, with attributes that are not measurable. Third, the empirical properties may represent different dimensions of the concept. After settling on a set of properties that best represent the concept, we write down a definition of the concept. This written definition communicates the subjects to which the concept applies and suggests a measurement strategy. Let's illustrate these steps by working through the example introduced earlier: liberalism.

Clarifying a Concept

The properties of a concept must have two characteristics. They must be concrete, and they must vary. Return to the question posed earlier: "Are women more liberal than men?" This is a conceptual question because it uses the intangible term *liberal* and, thus, does not readily admit to an empirical answer. But notice two things. First, the conceptual term *liberal* certainly represents measurable characteristics of people. After all, when we say that a person or group of people is "liberal," we must have some attributes or characteristics in mind. Second, the question asks whether liberalism varies between people. That is, it asks whether some people have more or less of these attributes or characteristics than other people. In clarifying a concept, then, we want to describe characteristics that are concrete and that vary. What, exactly, are these characteristics?

A mental exercise can help you to identify characteristics that are concrete and that vary. Think of two subjects who are polar opposites. In this example, we are interested in defining liberalism among individuals, so we would think of two polar-opposite people. At one pole is

a person who has a great deal of the concept's characteristics. At the other pole is a person who has the perfect opposite of the characteristics. What images of a perfectly liberal person do you see in your mind's eye? What images of a perfect opposite, an antiliberal or conservative, do you see?[3] In constructing these images, be open and inclusive. Here is an example of what you might come up with:

A liberal:	A conservative:
Has low income	Has high income
Is younger	Is older
Supports social justice	Opposes social justice
Opposes the free market	Supports the free market
Supports government-funded health care	Opposes government-funded health care
Opposes tax cuts	Supports tax cuts
Opposes restrictions on abortion	Supports restrictions on abortion
Supports same-sex marriage	Opposes same-sex marriage

Brainstorming polar opposites is an open-ended process, and it always produces the raw materials from which a conceptual definition can be built. Once the inventory is made, however, we need to become more critical and discerning. Consider the first two characteristics. According to the list, a liberal "has low income" and "is younger," whereas a conservative "has high income" and "is older." Think about this for a moment. Are people's incomes and ages really a part of the concept of liberalism? Put another way: Can we think about what it means to be liberal or conservative without thinking about income and age? You would probably agree that we could. To be sure, liberalism may be related to demographic factors, such as income and age, but the concept is itself distinct from these characteristics. This is the first problem to look for when clarifying a concept. Some traits seem to fit with the portraits of the polar-opposite subjects, but they are not essential parts of the concept. Let's drop the nonessential traits and reconsider our newly abbreviated inventory:

A liberal:	A conservative:
Supports social justice	Opposes social justice
Opposes the free market	Supports the free market
Supports government-funded health care	Opposes government-funded health care
Opposes tax cuts	Supports tax cuts
Opposes restrictions on abortion	Supports restrictions on abortion
Supports same-sex marriage	Opposes same-sex marriage

According to the list, a liberal "supports social justice" and "opposes the free market." A conservative "opposes social justice" and "supports the free market." Neither of these items should be on the list. Why not? Because neither one is measurable. Both terms are themselves concepts, and we cannot use one concept to define another. When constructing an inventory, imagine that a skeptical observer is looking over your shoulder, pressing you to specify concrete, measurable traits. How, exactly, would you determine whether someone supports free markets? How would you define social justice? If your initial response is, "I can't define it, but I know it when I see it"—to paraphrase an infamous remark about pornography—then

you need to dig deeper for concrete elements.[4] This is the second problem to look for when clarifying a concept. Some descriptions seem to fit the portraits of the polar-opposite subjects, but these descriptions are themselves vague, conceptual terms. Let's drop the conceptual terms from the inventory.

A liberal:	A conservative:
Supports government-funded health care	Opposes government-funded health care
Opposes tax cuts	Supports tax cuts
Opposes restrictions on abortion	Supports restrictions on abortion
Supports same-sex marriage	Opposes same-sex marriage

One could reasonably argue that all these traits belong on an empirical inventory of liberalism. One can think of observable phenomena that would offer tangible measurements, including checkmarks on a questionnaire gauging opinion on different government policies, the display of bumper stickers or yard signs, monetary contributions to issue groups, or a number of other overt behaviors. But examine the list carefully. Can the attributes be grouped into different types? Are some items similar to each other and, as a group, different from other items? You may have already noticed that supports/opposes government-funded health care and opposes/supports tax cuts refer to traditional differences between those who favor a larger public sector and more social services (liberals) and those who favor a more limited governmental role (conservatives). The other items, opposes/supports abortion restrictions and supports/opposes same-sex marriage, refer to more recent disputes between those who favor personal freedoms (liberals) and those who support proscriptions on these behaviors (conservatives). This example illustrates the third problem to look for when clarifying a concept. All the traits fit with the portraits of the polar-opposite subjects, but they may describe different dimensions of the concept.

A **conceptual dimension** is defined by a set of concrete traits of similar type. Some concepts, such as liberalism, are multidimensional. A **multidimensional concept** has two or more distinct groups of empirical characteristics. In a multidimensional concept, each group contains empirical traits that are similar to each other. Furthermore, each group of traits is qualitatively distinct from other groups of traits. To avoid confusion, the different dimensions need to be identified, labeled, and measured separately. Thus the traditional dimension of liberalism, often labeled *economic liberalism,* subsumes an array of similar attributes: support for government-funded health care, aid to poor people, funding for education, spending for infrastructure, and so on. The moral dimension, often labeled *social liberalism,* includes policies dealing with gay and lesbian rights, abortion, the legalization of marijuana, the teaching of evolution, and prayer in schools. By grouping similar properties together, the two dimensions can be labeled separately—economic liberalism and social liberalism—and measured separately.[5]

Many ideas in political science are multidimensional concepts. For example, in his seminal work, *Polyarchy,* Robert A. Dahl points to two dimensions of democracy: contestation and inclusiveness.[6] Contestation refers to attributes that describe the competitiveness of political systems—for example, the presence or absence of frequent elections or whether a country has legal guarantees of free speech. Inclusiveness refers to characteristics that measure how many people are allowed to participate, such as the presence or absence of restrictions on the right to vote or conditions on eligibility for public

office. Dahl's conceptual analysis has proven to be an influential guide for the empirical study of democracy.[7]

Many political concepts have a single dimension. The venerable social science concept of social status or socioeconomic status (SES), for example, has three concrete attributes that vary across people: income, occupation, and education. Yet it seems reasonable to say that all three are empirical manifestations of one dimension of SES.[8] Similarly, if you sought to clarify the concept of cultural fragmentation, you might end up with a polar-opposite list of varied but dimensionally similar characteristics of polities: many/few major religions practiced, one/several languages spoken, one/many racial groups, and so on. For each of these concepts, SES and cultural fragmentation, you can arrive at a single measure by determining whether people or polities have a great deal of the concept's characteristics.

A Template for Writing a Conceptual Definition

A conceptual definition must communicate three things:

1. The variation within a measurable characteristic or set of characteristics
2. The subjects or groups to which the concept applies
3. How the characteristic is to be measured

The following is a workable template for stating a conceptual definition that meets all three requirements:

> The concept of _____ is defined as the extent to which _____ exhibit the characteristic of _____.

For a conceptual definition of economic liberalism, we would write the following:

> The concept of economic liberalism is defined as the extent to which individuals exhibit the characteristic of supporting government spending for social programs.

The first term, *economic liberalism,* when combined with the words "the extent to which," restates the concept's label and communicates the polar-opposite variation at the heart of the concept. The second term, *individuals,* states the subjects to whom the concept applies. The third term, *supporting government spending for social programs,* suggests the concept's measurement. Let's consider the template in more detail.

By referring to a subject or group of subjects, a conceptual definition conveys the units of analysis. A **unit of analysis** is the entity (person, city, country, county, university, state, bureaucratic agency, etc.) we want to describe and analyze; it is the entity to which the concept applies. Units of analysis can be either individual level or aggregate level. When a concept describes a phenomenon at its lowest possible level, it is using an **individual-level unit of analysis**. Most polling or survey research deals with concepts that apply to individual persons, which are the most common individual-level units of analysis you will encounter. Individual-level units are not always persons, however. If you were conducting research on the political themes contained in the Democratic and Republican Party platforms over the past several elections, the units of analysis would be the individual platforms from each year. Similarly, if you were interested in finding out whether

environmental legislation was a high priority in Congress, you might examine each bill that is introduced as an individual unit of analysis.

Much political science research deals with the **aggregate-level unit of analysis**, which is a collection of individual entities. Neighborhoods or census tracts are aggregate-level units, as are congressional districts, states, and countries. A university administrator who wondered if student satisfaction is affected by class size would gather information on each class, an aggregation of individual students. Someone wanting to know whether states with lenient voter registration laws had higher turnout than states with stricter laws could use legal statistics and voting data from fifty aggregate-level units of analysis, the states. Notice that collections of individual entities, and thus overall aggregate levels, can vary in size. For example, both congressional districts and states are aggregate-level units of analysis—both are collections of individuals within politically defined geographic areas—but states usually represent a higher level of aggregation because they are composed of more individual entities.

Notice, too, that the same concept often can be defined at both the individual and aggregate levels. Dwell on this point for a moment. Just as economic liberalism can be defined for individual persons, economic liberalism can be defined for states by aggregating the numbers of state residents who support or oppose government spending: The concept of economic liberalism is defined as the extent to which states exhibit the characteristic of having residents who support government spending for social programs. This conceptual definition makes perfect sense. One can imagine comparing states that have a large percentage of pro-spending residents with states having a lower percentage of pro-spending residents. For statistical reasons, however, the relationship between two aggregate-level concepts usually cannot be used to make inferences about the relationship at the individual level. Suppose we find that states with larger percentages of college-educated people have higher levels of economic liberalism than states with fewer college graduates. Based on this finding, we could not conclude that college-educated individuals are more likely to be economic liberals than are individuals without a college degree.

A classic problem, known as the **ecological fallacy**, arises when an aggregate-level phenomenon is used to make inferences at the individual level. W. S. Robinson, who coined the term more than sixty years ago, illustrated the ecological fallacy by pointing to a counterintuitive fact: States with higher percentages of foreign-born residents had higher rates of English-language literacy than states with lower percentages of foreign-born residents. At the individual level, Robinson found the opposite pattern, with foreign-born individuals having lower English literacy than native-born individuals. What accounted for these paradoxical findings? The aggregate-level pattern was produced by the tendency for immigrants to settle in states whose native-born residents had comparatively high levels of language proficiency.[9] The ecological fallacy is not new. Indeed, Emile Durkheim's towering study of religion and suicide, published in 1897, may have suffered from it.[10] The main point here is that a proper conceptual definition needs to specify the units of analysis. Researchers must be careful when drawing conclusions based on the study of aggregate-level units of analysis.

OPERATIONAL DEFINITIONS

In suggesting how the concept is to be measured, a conceptual definition points the way to a clear operational definition.[11] An operational definition describes explicitly how the concept is to be measured empirically. Just how would we determine the extent to which people hold opinions that

are consistent with economic liberalism? What procedure would produce the truest measure of social liberalism? Suppose we wanted to quantify Dahl's inclusiveness dimension of democracy. We would need to devise a metric that combines the different concrete attributes of inclusiveness. Exactly what form would this metric take? Would it faithfully reflect the conceptual dimension of inclusiveness, or might our measure be flawed in some way? This phase of the measurement process, the step between conceptual definition and operational definition, is often the most difficult to traverse. To introduce some of these difficulties, we describe an example from public opinion research, the study of the concept of political tolerance.

Political tolerance is important to many students of democracy because, arguably, democratic health can be maintained only if people remain open to different ways of thinking and solving problems. If tolerance is low, then democratic procedures will be weakly supported, and the free exchange of ideas might be threatened. Political tolerance is a rather complex concept, and a large body of research and commentary is devoted to it.[12] For our more limited purpose here, consider the following conceptual definition:

> The concept of political tolerance is defined as the extent to which individuals exhibit the characteristic of expressing a willingness to allow basic political freedoms for unpopular groups.

Awkward syntax aside, this is a serviceable definition, and it has been the starting point for a generation of scholars interested in studying the concept. Beginning in the 1950s, the earliest research "operationalized" political tolerance by asking large numbers of individuals if certain procedural freedoms (for example, giving a speech or publishing a book) should be extended to members of specific groups: atheists, communists, and socialists. This seemed like a reasonable operational definition because, at the time at least, these groups represented ideas outside the conformist mainstream and were generally considered unpopular. The main finding was somewhat unsettling: Whereas those in positions of political leadership expressed high levels of tolerance, the public-at-large appeared much less willing to allow basic freedoms for these groups.

Later research, however, pointed to important slippage between the conceptual definition, which clarified and defined the important properties of political tolerance, and the operational definition, the procedure used to measure political tolerance. The original investigators had themselves chosen which unpopular groups were outside the mainstream, and these groups tended to have a left-wing or left-leaning ideological bent. The researchers were therefore gauging tolerance only toward leftist groups. Think about this measurement problem. Consider a scenario in which a large number of people are asked to "suppose that an admitted communist wanted to make a speech in your community. Should he be allowed to speak, or not?" For the question's designers, the key words are "wanted to make a speech." Thus people who respond "allowed to speak" are measured as having a larger amount of political tolerance than are those who say "not allowed to speak." But it could be that for some respondents—it is impossible to know how many—the key word is "communist." These respondents might base their answers on how they feel about communists, not on how willing they are to apply the principle of free speech. Ideological liberals, who may regard communists as less threatening than other groups, would be measured as more tolerant than ideological conservatives, who regard communists as more threatening than other groups. In sum, although the operational goal was to gauge tolerance, this measurement strategy also measured respondents' ideological sympathies.

A better measurement strategy, one more faithful to the concept, would allow respondents *themselves* to name the groups they most strongly oppose—that is, the groups most unpopular with or disliked by each person being surveyed. Individuals would then be asked about extending civil liberties to the groups they had identified, not those picked beforehand by the researchers. Think about why this is a superior approach. Consider a scenario in which a large number of people are presented with a list of groups: racists, communists, socialists, homosexuals, white separatists, and so on. Respondents are asked to name the group they "like the least." Now recast the earlier survey instrument: "Suppose that [a member of the least-liked group] wanted to make a speech in your community. Should he be allowed to speak, or not?" Because the respondents themselves have selected the least-liked group, the investigators can be confident that those who say "allowed to speak" have a larger amount of tolerance than those who say "not allowed to speak." Interestingly, this superior measurement strategy led to equally unsettling findings: Just about everyone, elites and nonelites alike, expressed rather anemic levels of political tolerance toward the groups they liked the least.[13]

Measurement Error

As the tolerance example suggests, we want to devise an operational instrument that maximizes the congruence or fit between the definition of the concept and the empirical measure of the concept. Let's use the term *intended characteristic* to refer to the conceptual property we want to measure. The term *unintended characteristic* will refer to any other property or attribute that we do not want our instrument to measure. The researcher asks, "Does this operational instrument measure the intended characteristic? If so, does it measure *only* that characteristic? Or might it also be gauging an unintended characteristic?" Students of political tolerance are interested in asking individuals a set of questions that accurately gauge their willingness to extend freedoms to unpopular groups. The first measurement of tolerance did not accurately measure this intended characteristic. Why not? Because it was measuring a characteristic that it was not supposed to measure: individuals' attitudes toward left-wing groups. To be sure, the original measurement procedure was tapping an intended characteristic of tolerance. After all, a thoroughly tolerant person would not be willing to restrict the freedoms of any unpopular group, regardless of the group's ideological leanings, whereas a completely intolerant person would express a willingness to do so. When the conceptual definition was operationalized, however, an unintended characteristic, individuals' feelings toward leftist groups, also was being measured. Thus the measurement strategy created a poor fit, an inaccurate link, between the concept of tolerance and the empirical measurement of the concept.

Two sorts of error can distort the linkage between a concept and its empirical measure. Serious problems arise when **systematic measurement error** is at work. Systematic error introduces consistent, chronic distortion into an empirical measurement. Often called measurement bias, systematic error produces operational readings that consistently mismeasure the characteristic the researcher is after. The original tolerance measure suffered from systematic measurement error, because subjects with liberal ideological leanings were consistently (and incorrectly) measured as more tolerant than were ideologically conservative subjects. Less serious, but still troublesome, problems occur when **random measurement error** is present. Random error introduces haphazard, chaotic distortion into the measurement process, producing inconsistent operational readings of a concept. To appreciate the difference between these two kinds of error, and to see how each affects measurement, consider an example.

Suppose that a math instructor wishes to test the math ability of a group of students. This measurement is operationalized by ten word problems covering basic features of math. First let's ask, "Does this operational instrument measure the intended characteristic, math ability?" It seems clear that *some* part of the operational measure will capture the intended characteristic, students' actual knowledge of math. But let's press the measurement question a bit further: "Does the instructor's operational instrument measure *only* the intended characteristic, math ability? Or might it also be gauging a characteristic that the instructor did not intend for it to measure?" We know that, quite apart from mathematical competence, students vary in their verbal skills. Some students can read and understand the math problems more quickly than others. Thus the exam is picking up an unintended characteristic, an attribute it is not supposed to measure—verbal ability.

You can probably think of other characteristics that would "hitch a ride" on the instructor's test measure. In fact, a large class of unintended characteristics is often at work when human subjects are the units of analysis. This phenomenon, dubbed the **Hawthorne effect**, inadvertently measures a subject's response to the knowledge that he or she is being studied. Test anxiety is a well-known example of the Hawthorne effect. Despite their actual grasp of a subject, some students become overly nervous simply by being tested, and their exam scores will be systematically depressed by the presence of test anxiety.[14]

The unintended characteristics we have been discussing, verbal ability and test anxiety, are sources of systematic measurement error. Systematic measurement error refers to factors that produce consistently inaccurate measures of a concept. Notice two aspects of systematic measurement error. First, unintended characteristics such as verbal ability and test anxiety are durable, not likely to change very much over time. If the tests were administered again the next day or the following week, the test scores of the same students—those with fewer verbal skills or more test anxiety—would yield consistently poor measures of their true math ability. Think of two students, both having the same levels of mathematical competence but one having less verbal ability than the other. The instructor's operational instrument will report a persistent difference in math ability between these students when, in fact, no difference exists. Second, this consistent bias is inherent in the measurement instrument. When the instructor constructed a test using word problems, a measure of the unintended characteristic, verbal ability, was built directly into the operational definition. The source of systematic error resides—often unseen by the researcher—in the measurement strategy itself.

Now consider some temporary or haphazard factors that might come into play during the instructor's math exam. Some students may be ill or tired; others may be well rested. Students sitting near the door may be distracted by commotion outside the classroom, whereas those sitting farther away may be unaffected. Commuting students may have been delayed by traffic congestion caused by a fender bender near campus, and so, arriving late, they may be pressed for time. The instructor may make errors in grading the tests, accidentally increasing the scores of some students and decreasing the scores of others.

These sorts of factors—fatigue, commotion, unavoidable distractions—are sources of random measurement error. Random measurement error refers to factors that produce inconsistently inaccurate measures of a concept. Notice two aspects of random measurement error. First, unintended characteristics such as commotion and grading errors are not durable, and they are not consistent across students. They may or may not be present in the same student if the test were administered again the next day or the following week. A student may be ill or delayed by traffic one week, well and on time the next. Second, chance events certainly can affect the operational readings of a concept, but they are not built into the

operational definition itself. When the instructor constructed the exam, he did not build traffic accidents into the measure. Rather, these factors intrude from outside the instrument. Chance occurrences introduce haphazard, external "noise" that may temporarily and inconsistently affect the measurement of a concept.

Reliability and Validity

We can effectively use the language of measurement error to evaluate the pros and cons of a particular measurement strategy. For example, we could say that the earliest measure of political tolerance, though perhaps having a small amount of random error, contained a large amount of systematic error. The hypothetical math instructor's measurement sounds like it had a dose of both kinds of error—systematic error introduced by durable differences between students in verbal ability and test anxiety, and random error that intruded via an array of haphazard occurrences. Typically, researchers do not evaluate a measure by making direct reference to the amount of systematic error or random error it may contain. Instead they discuss two criteria of measurement: reliability and validity. However, reliability and validity can be understood in terms of measurement error.

The **reliability** of a measurement is the extent to which it is a consistent measure of a concept. Assuming that the property being measured does not change between measurements, a reliable measure gives the same reading every time it is taken. Applying the ideas we just discussed, we see that a completely reliable measure is one that contains no random error. As random measurement noise increases—repeated measurements jump around haphazardly—a measure becomes less reliable. A measure need not be free of systematic error to be reliable. It just needs to be consistent. Consider a nonsensical example that nonetheless illustrates the point. Suppose a researcher gauges the degree to which the public approves of government spending by using a laser measuring device to precisely record respondents' heights in centimeters, with higher numbers of centimeters denoting stronger approval for spending. This researcher's measure would be quite reliable because it would contain very little random error and would therefore be consistent. But it would clearly be gauging a concept completely different from opinions about government spending. In a more realistic vein, suppose the math instructor recognized the problems caused by random occurrences and took steps to greatly reduce these sources of random error. Certainly his measurement of math ability would now be more consistent, more reliable. However, it would not reflect the true math ability of students because it would still contain systematic error. More generally, although reliability is a desirable criterion of measurement—any successful effort to purge a measure of random error is a good thing—it is a weaker criterion than validity.

The **validity** of a measurement is the extent to which it records the true value of the intended characteristic and does not measure any unintended characteristics. A valid measure provides a clear, unobstructed link between a concept and the empirical reading of the concept. Framed in terms of measurement error, the defining feature of a valid measure is that it contains no systematic error, no bias that consistently pulls the measurement off the true value. Suppose a researcher gauges opinions toward government spending by asking each respondent to indicate his or her position on a 7-point scale, from "spending should be increased" on the left to "spending should be decreased" on the right. Is this a valid measure? A measure's validity is harder to establish than is its reliability. But it seems reasonable to say that this measurement instrument is free from systematic error and thus would closely reflect respondents' true opinions on the issue. Or suppose the math instructor tries to alleviate the

sources of systematic error inherent in his test instrument—switching from word problems to a format based on mathematical symbols, and perhaps easing anxiety by shortening the exam or lengthening the allotted time. These reforms would reduce systematic error, strengthen the connection between true math ability and the measurement of math ability, and thus enhance the validity of the test.

Suppose we have a measurement that contains no systematic error but contains some random error. Would this be a valid measure? Can a measurement be valid but not reliable? Although we find conflicting scholarly answers to this question, let's settle on a qualified yes.[15] Instead of considering a measurement as either not valid or valid, think of validity as a continuum, with "not valid" at one end and "valid" at the other. An operational instrument that has serious measurement bias, lots of systematic error, would reside at the "not valid" pole, regardless of the amount of random error it contains. The early measure of political tolerance is an example. An instrument with no systematic error and no random error would be at the "valid" end. Such a measure would return an accurate reading of the characteristic that the researcher intends to measure, and it would do so with perfect consistency. The math instructor's reformed measurement process—changing the instrument to remove systematic error, taking pains to reduce random error—would be close to this pole. Now consider two measures of the same concept, neither of which contains systematic error, but one of which contains less random error. Because both measures vanquish measurement bias, both would fall on the "valid" side of the continuum. But the more consistent measure would be closer to the "valid" pole.

Evaluating Reliability

Methods for evaluating reliability are designed around this assumption: If a measurement is reliable, it will yield consistent results. In everyday language, "consistent" generally means, "stays the same over time." Accordingly, some approaches to reliability apply this measure-now-measure-again-later intuition. Other methods assess the internal consistency of an instrument and thus do not require readings taken at two time points. First we describe methods based on over-time consistency. Then we turn to approaches based on internal consistency. In the **test-retest method** the investigator applies the measure once and then applies it again to the same units of analysis. If the measurement is reliable, then the two results should be the same or very similar. If a great deal of random measurement error is present, then the two results will be very different. For example, suppose we construct a 10-item instrument to measure individuals' levels of economic liberalism. We create the scale by asking each respondent whether spending should or should not be increased on ten government programs. We then add up the number of programs on which the respondent says "increase spending." We administer the questionnaire and then readminister it at a later time to the same people. If the scale is reliable, then each person's score should change very little over time. The alternative-form method is similar to the test-retest approach. In the **alternative-form method** the investigator administers two different but equivalent versions of the instrument—one form at time point 1 and the equivalent form at time point 2. For our economic liberalism example, we would construct two 10-item scales, each of which elicits respondents' opinions on ten government programs. Why go to the trouble of devising two different scales? The alternative-form method remedies a key weakness of the test-retest method: In the second administration of the same questionnaire, respondents may remember their earlier responses and make sure that they give the same opinions again. Obviously, we want to measure economic liberalism, not memory retention.

Methods based on over-time consistency have two main drawbacks. First, these approaches make it hard to distinguish random error from true change. Suppose that, between the first and second administrations of the survey, a respondent becomes more economically liberal, perhaps scoring a 4 the first time and a 7 the second time. Over-time methods of evaluating reliability assume that the attribute of interest—in this case, economic liberalism—is unchanging. Thus the observed change, from 4 to 7, is assumed to be random error. The longer the time period between questionnaires, the bigger this problem becomes.[16] A second drawback is more practical: Surveys are expensive projects, especially when the researcher wants to administer an instrument to a large number of people. The test-retest and alternative-form approaches require data obtained from panel studies. A **panel study** contains information on the same units of analysis measured at two or more points in time. Respondents a, b, and c are interviewed at time 1; respondents a, b, and c are interviewed again at time 2. Data from cross-sectional studies are more the norm in social research. A **cross-sectional study** contains information on units of analysis measured at one point in time. Respondents a, b, and c are interviewed—that's it. Though far from inexpensive, cross-sectional measurements are obtained more easily than panel measures. As a practical matter, then, most political researchers face the challenge of evaluating the reliability of a measurement that was made using cross-sectional data.[17] Internal consistency methods are designed for these situations.

One internal consistency approach, the **split-half method**, is based on the idea that an operational measurement obtained from half of a scale's items should be the same as the measurement obtained from the other half. In the split-half method the investigator divides the scale items into two groups, calculates separate scores, and then compares the measurements. If the items are reliably measuring the same concept, then the two sets of scores should be the same. Following this technique, we would break our ten government spending questions into two groups of five items each, calculate two scores for each respondent, and then compare the scores. Plainly enough, if we have devised a reliable instrument, then the respondents' scores on one 5-item scale should match closely their scores on the other 5-item scale. A more sophisticated internal consistency approach, **Cronbach's alpha**, is a natural methodological extension of the split-half technique. Instead of evaluating consistency between separate halves of a scale, Cronbach's alpha compares consistency between pairs of individual items and provides an overall reading of a measure's reliability.[18] Imagine a perfectly consistent measure of economic liberalism. Every respondent who says "increase spending" on one item also says "increase spending" on all the other items, and every respondent who says "do not increase spending" on one item also says "do not increase spending" on every other item. In this scenario, Cronbach's alpha would report a value of 1, denoting perfect reliability. If responses to the items betray no consistency at all—opinions about one government program are not related to opinions about other programs—then Cronbach's alpha would be 0, telling us that the scale is completely unreliable. Of course, most measurements' reliability readings fall between these extremes.

It is easy to see how the methods of evaluating reliability help us to develop and improve our measures of concepts. To illustrate, let's say we wish to measure the concept of social liberalism, the extent to which individuals accept new moral values and personal freedoms. After building an empirical inventory, we construct a scale based on support for five policies: same-sex marriage, marijuana legalization, abortion rights, stem cell research, and physician-assisted suicide. Our hope is that by summing respondents' five positions, we can arrive at a reliable operational reading of social liberalism. With all five items included, the scale has a

Cronbach's alpha equal to .6. Some tinkering reveals that, by dropping the physician-assisted suicide item, we can increase alpha to .7, an encouraging improvement that puts the reliability of our measure near the threshold of acceptability.[19] The larger point to remember is that the work you do at the operational definition stage often helps you to refine the work you did at the concept clarification stage.

Evaluating Validity

Reliability is an important and sought-after criterion of measurement. Most standardized tests are known for their reliability. The SAT, the Law School Admission Test (LSAT), and the Graduate Record Examination (GRE), among others, all return consistent measurements. But the debate about such tests does not center on their reliability. It centers, instead, on their validity: Do these exams measure what they are supposed to measure and only what they are supposed to measure? Critics argue that because many of these tests' questions assume a familiarity with white, middle-class culture, they do not produce valid measurements of aptitudes and skills. Recall again the earliest measurements of political tolerance, which gauged the concept by asking respondents whether basic freedoms should be extended to specific groups: atheists, communists, and socialists. Because several different studies used this operationalization and produced similar findings, the measure was a reliable one. The problem was that a durable unintended characteristic, the respondents' attitudes toward left-wing groups, was "on board" as well, giving a consistent if inaccurate measurement of the concept.

The challenge of assessing validity is to identify durable unintended characteristics that are being gauged by an operational measure, that is, to identify the sources of systematic measurement error. To be sure, some sources of systematic error, such as verbal skills or test anxiety, are widely recognized, and steps can be taken to ameliorate their effects. In most situations, however, less well-known factors might be affecting validity. How can these problems be identified? There are two general ways to evaluate validity.

In the **face validity** approach, the investigator uses informed judgment to determine whether an operational procedure is measuring what it is supposed to measure. "On the face of it," the researcher asks, "are there good reasons to think that this measure is not an accurate gauge of the intended characteristic?" In the **construct validity** approach, the researcher examines the empirical relationships between a measurement and other concepts to which it should be related. Here the researcher asks, "Does this measurement have relationships with other concepts that one would expect it to have?" Let's look at an example of each approach.

Responses to the following agree-disagree question have been used by survey researchers to measure the concept of political efficacy, the extent to which individuals believe that they can have an effect on government: "Voting is the only way that people like me can have any say about how the government runs things." According to the question's operational design, a person with a low level of political efficacy would see few opportunities for influencing government beyond voting and thus would give an "agree" response. A more efficacious person would feel that other avenues exist for "people like me" and so would tend to "disagree." But examine the survey instrument closely. Using informed judgment, address the face validity question: Are there good reasons to think that this instrument would not produce an accurate measurement of the intended characteristic, political efficacy? Think of an individual or group of individuals whose sense of efficacy is so weak that they view any act of political participation, including voting, as an exercise in political futility. At the conceptual

level, one would certainly consider such people to have a low amount of the intended characteristic. But how might they respond to the survey question? Quite reasonably, they could say "disagree," a response that would measure them as having a large amount of the intended characteristic. Taken at face value, then, this survey question is not a valid measure.[20] This example underscores a general problem posed by factors that affect validity. We sometimes can identify potential sources of systematic error and suggest how this error is affecting the operational measure. Thus people with low and durable levels of efficacy might be measured, instead, as being politically efficacious. However, it is difficult to know the size of this effect. How many people are being measured inaccurately? A few? Many? It is impossible to know.

On a more hopeful note, survey methodologists have developed effective ways of weakening the chronic distortion of measurement bias, even when the reasons for the bias, or its precise size, remain unknown. For example, consider the systematic error that can be introduced by the order in which respondents answer a pollster's questions. Imagine asking people the following two questions about abortion: (1) "Do you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children?" (2) "Do you think it should be possible for a pregnant woman to obtain a legal abortion if there is a strong chance of serious defect in the baby?" The first item receives a substantially higher percentage of "yes" responses when it is asked first than when it is asked after the second item.[21] A palliative is available for such question-order effects: Randomize the order in which the questions appear in a survey. In this way, systematic measurement error is transformed into random measurement error. Random measurement error may not be cause for celebration among survey designers, but, as we have seen, random error is easier to deal with than systematic error.[22]

The second approach to evaluating validity, construct validity, assesses the association between the measure of a concept and another concept to which it should be related. This is a reasonable approach to the problem. For example, if the GRE is a valid measure of students' readiness for graduate school, then GRE scores should be strongly related to subsequent grade point averages earned by graduate students. If the GRE is an inaccurate measure of readiness, then this relationship will be weak.[23]

Here is an example from political science. For many years, the American National Election Study has provided a measurement of the concept of party identification, the extent to which individuals feel a sense of loyalty or attachment to one of the major political parties. This concept is measured by a 7-point scale. Each person is classified as a Strong Democrat, Weak Democrat, Independent-leaning Democrat, Independent–no partisan leanings, Independent-leaning Republican, Weak Republican, or Strong Republican. If we apply the face validity approach, this measure is difficult to fault. Following an initial gauge of direction (Democrat, Independent, Republican), interviewers meticulously lead respondents through a series of probes, recording gradations in the strength of their partisan attachments: strongly partisan, weakly partisan, independent-but-leaning partisan, and purely independent.[24] Durable unintended characteristics are not readily apparent in this measurement strategy. But let's apply the construct validity approach. If the 7-point scale accurately measures strength of party identification, then it should bear predictable relationships to other concepts.

For example, we would expect strongly partisan people, whether Democrats or Republicans, to engage in much campaigning during an election—displaying bumper stickers, wearing buttons, attending rallies, or perhaps donating money to one of the parties. By the same token, we would expect weak partisans to engage in fewer of these activities,

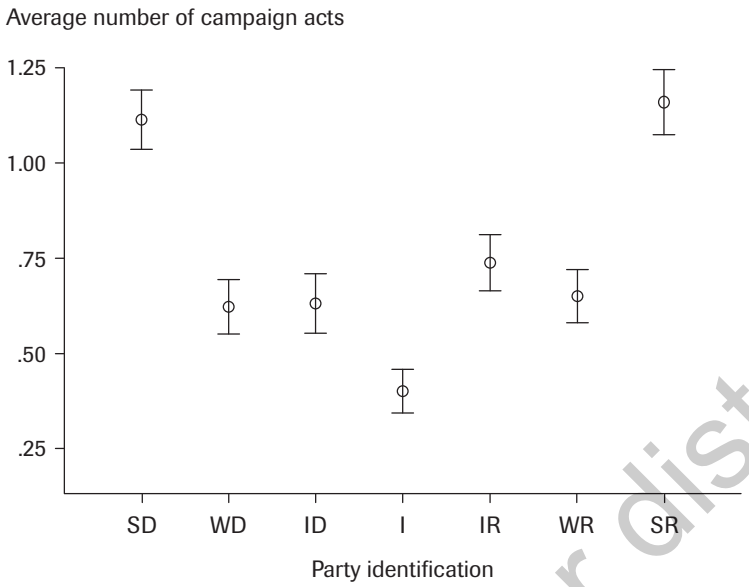
Independent leaners fewer still, and Independents the fewest of all. That is the logic of construct validity. If the 7-point scale is a valid measure of partisan strength, then it should relate to clearly partisan behaviors (campaign activities) in an expected way. How does the concept of party identification fare in this test of its validity?

Figure 1-1 shows the empirical relationship between the 7-point party identification measurement and a measurement of campaigning. The values of party identification appear on the horizontal axis. The vertical axis records the average number of campaign activities engaged in by each partisan type during the 2012 election. This particular graphic form is an error bar chart, because it also displays the 95 percent confidence interval for each mean, an indicator of the amount of random measurement error contained in each estimate. If one mean's error bar overlaps with another mean's error bar, the two means are equivalent, statistically speaking. (Error bar charts are covered in Chapter 7.) Notice that, as expected, people at the strongly partisan poles, Strong Democrats and Strong Republicans, were the most likely to report campaign behavior. And, again as expected, pure Independents were the least likely to engage in partisan activity. But, beyond these expectations, is anything amiss here? Notice that Weak Democrats, measured as having stronger party ties than Independent-leaning Democrats, were about as likely to campaign as were Independent-Democratic leaners. A similar comparison at the Republican side of the scale—Weak Republicans compared with Independent-Republican leaners—shows the same thing: Weak partisans and people measured as Independents with partisan leanings demonstrated no meaningful difference in an explicitly partisan behavior, campaigning.

Scholars who have examined the relationships between the 7-point scale and other concepts also have found patterns similar to that shown in Figure 1-1.[25] In applying the construct validity approach, we can use empirical relationships such as that displayed in Figure 1-1 to evaluate an operational measure. What would we conclude from this example about the validity of this measurement of partisanship? Clearly the measure is tapping some aspect of the intended characteristic. After all, the scale "behaves as it should" among strong partisans and pure Independents. But how would one account for the unexpected behavior of weak partisans and independent leaners? What durable unintended characteristic might the scale also be measuring? Some scholars have suggested that the scale is tapping two durable characteristics—one's degree of partisanship (the intended characteristic) and one's degree of independence (an unintended characteristic)—and that the two concepts, partisanship and independence, should be measured separately.[26] Others have argued that a fundamental mismatch exists between the concept of party identification and the questions used to measure it, and that a new survey protocol is needed.[27] There is, to put it mildly, spirited debate on this and other questions about the measurement of party identification.

Rest assured that debates about validity in political science are not academic games of "gotcha," with one researcher proposing an operational measure and another researcher marshaling empirical evidence to shoot it down. Rather, the debate is productive. It is centered on identifying potential sources of systematic error, and it is aimed at improving the quality of widely used operational measures. It bears emphasizing, as well, that although the problem of validity is a concern for the entire enterprise of political analysis, some research is more prone to it than others. A student of state politics could obtain a valid measure of the concept of state-supported education fairly directly, by calculating a state's per-capita spending on education. A congressional scholar would validly measure the concept of party cohesion by figuring out, across a series of votes, the percentage of times a majority of Democrats opposed a majority of Republicans. In these examples, the connection between the

Figure 1-1 The Relationship between Party Identification and Campaign Activity



Source: 2012 American National Election Study.

Note: Figure displays means of a campaign activity scale created by summing the number of acts engaged in by respondents: trying to convince someone how to vote, attending a campaign rally, displaying a bumper sticker or sign, contributing money to a candidate, contributing money to a political party. Displayed means are as follows: Strong Democrats, 1.11; Weak Democrats, .62, Independent-Democrats, .63; Independents, .40; Independent-Republicans, .74, Weak Republicans, .65; Strong Republicans, 1.16. Based on 2012 ANES variables: mobilpo_rmob, mobilpo_rally, mobilpo_sign, mobilpo_ctbcand, and mobilpo_ctbpty.

concept and its operational definition is direct and easy to recognize. By contrast, researchers interested in individual-level surveys of mass opinion, as the above examples illustrate, often face tougher questions of validity.

SUMMARY

In this chapter we introduced the essential features of concepts and measurement. A concept is an idea, a mental image that cannot be measured or quantified. A main goal of social research is to express concepts in concrete language, to identify the empirical properties of concepts so that they can be analyzed and understood. This chapter described a heuristic that may help you to clarify the concrete properties of a concept: Think of polar-opposite subjects, one of whom has a great deal of the concept's properties and the other of whom has none of the properties. The properties you specify should not themselves be concepts, and they should not describe the characteristics of a different concept. It may be, as well, that the concept you are interested in has more than one dimension.

This chapter described how to write a conceptual definition, a statement that communicates variation within a characteristic, the units of analysis to which the concept applies, and how the concept is to be measured. Important problems can arise when we measure a concept's empirical properties—when we put the conceptual definition into operation. Our measurement strategy may be accompanied by a large amount of random

measurement error, error that produces inconsistently incorrect measures of a concept. Random error undermines the reliability of the measurements we make. Our measurement strategy may contain systematic measurement error, which produces consistently incorrect measures of a concept. Systematic error undermines the validity of our measurements. Although measurement problems are a persistent worry for social scientists, all is not lost. Researchers have devised productive approaches to enhancing the reliability and validity of their measures.

Take a closer look.
edge.sagepub.com/pollock

KEY TERMS

aggregate-level unit of analysis (p. 7)
alternative-form method (p. 12)
concept (p. 1)
conceptual definition (p. 2)
conceptual dimension (p. 5)
conceptual question (p. 2)
concrete question (p. 2)
construct validity (p. 14)
Cronbach's alpha (p. 13)
cross-sectional study (p. 13)
ecological fallacy (p. 7)
face validity (p. 14)
Hawthorne effect (p. 10)
individual-level unit of analysis (p. 6)
multidimensional concept (p. 5)
operational definition (p. 3)
panel study (p. 13)
random measurement error (p. 9)
reliability (p. 11)
split-half method (p. 13)
systematic measurement error (p. 9)
test-retest method (p. 12)
unit of analysis (p. 6)
validity (p. 11)

EXERCISES

1. Suppose you wanted to study the role of religious belief, or religiosity, in politics and society. You would begin by setting up an inventory of empirical properties, contrasting the mental images of a religious person and a nonreligious person.

A religious person:	A nonreligious person:
a. Regularly prays	a. Never prays
b.	b.
c.	c.

   A. Item a, "regularly prays/never prays," provides a good beginning for the inventory. Think up and write down two additional items, b and c.

   B. As discussed in this chapter, a common problem in developing an empirical inventory is that we often come up with items that measure a completely different concept. For example, in constructing the liberal-conservative inventory, we saw that "has low income"/"has high income" did not belong on the list, because income and ideology are different concepts. For each item you chose in part A, explain why you think each property is a measure of religiosity and does not measure any other concept.

   C. Using one of your items, b or c, write a conceptual definition of religiosity. In writing the conceptual definition, be sure to use the template presented in this chapter.

2. *Finding 1:* An examination of state-level data on electoral turnout reveals that, as states' percentages of low-income citizens increase, turnout increases. *Conclusion:* Low-income citizens are more likely to vote than are high-income citizens.

   A. For the purposes of this exercise, assume that Finding 1 is correct—that is, assume that Finding 1 describes the data accurately. Is the conclusion supported? Making specific reference to a problem discussed in this chapter, explain your answer.

   B. Suppose that, using individual-level data, you compared the voting behavior of low-income citizens and high-income citizens. *Finding 2:* Low-income citizens are less likely to vote than high-income citizens. Explain how Finding 1 and Finding 2 can both be correct.

3. This chapter discussed the Hawthorne effect, a measurement problem that can arise in the study of human subjects.

   A. Using an example other than the one used in this chapter, describe a measurement situation in which the Hawthorne effect would arise.

   B. What sort of measurement error is the Hawthorne effect—random measurement error or systematic measurement error? Explain your answer.

4. Four researchers, Warren, Xavier, Yolanda, and Zelda, have devised different operational measures for gauging individuals' levels of political knowledge. Each researcher's operational measure is a scale ranging from 0 (low knowledge) to 100 (high knowledge). For the purposes of this exercise, assume that you know—but the researchers do not know—that the "true" level of knowledge of a test respondent is equal to 50. The researchers measure the respondent four times. Here are the measurements obtained by each of the four researchers:

   Warren: 40, 60, 70, 45

   Xavier: 48, 48, 50, 54

   Yolanda: 49, 50, 51, 50

   Zelda: 45, 44, 44, 46

   A. Which researcher's operational measure has high validity and high reliability? Explain.

   B. Which researcher's operational measure has high validity and low reliability? Explain.

   C. Which researcher's measure has low validity and high reliability? Explain.

   D. Which researcher's measure has low validity and low reliability? Explain.

5. Two candidates are running against each other for a seat on the city commission. You would like to obtain a valid measurement of which candidate has more preelection support among the residents of your community. Your operational measure: Obtain a precise count of bumper stickers on vehicles parked at a nearby shopping mall. The candidate with a greater number of bumper stickers will be measured as having greater preelection support than the candidate having fewer bumper stickers.

   A. This measurement strategy has low face validity, because it clearly measures unintended characteristics—characteristics other than preelection support for the two candidates. Describe two reasons why the measurement strategy will not produce a valid measurement of residents' preelection support for the candidates.

   B. Putting your bumper-sticker counting days behind you, you resolve to develop a more valid measurement of preelection support for the candidates. Describe a measurement strategy that would produce a valid measure of preelection support for candidates running for election.

C. This chapter discussed construct validity—the idea that a measurement instrument is valid if it has expected relationships with other concepts. Describe a relationship that would help establish the construct validity of the measurement you described in part B.

6. Political scientists who are interested in survey research often must deal with tough issues of validity. A survey question may not be valid, because it does not tap the intended characteristic. Sometimes, too, a survey question may gauge an unintended characteristic. Below are three hypothetical survey questions. For each one: (i) Describe why the question is not a valid measure of the intended characteristic. (ii) Rewrite the survey question to improve its validity.

   Example. Intended characteristic: Whether or not an individual voted in the 2012 election. Survey question: "Thinking about the 2012 election, some people did their patriotic duty and voted in the election. Others were too unpatriotic to vote. How about you—did you vote in the 2012 election?"

   (i) This question is not valid because it also measures individuals' feelings of patriotism. Not wanting to appear unpatriotic, citizens who did not vote may respond that they did vote.

   (ii) Improved rewrite: "Thinking about the 2012 election, some people voted in the election, whereas other people did not vote. How about you—did you vote in the 2012 election?"

   A. Intended characteristic: Respondents' position on the trade-off between protecting the environment and creating jobs. Survey question: "Which of the following comes closest to your opinion about the trade-off between protecting the environment and creating jobs: 'We should regulate business to protect the environment and create jobs.' OR 'We should not regulate business because it will not work and will cost jobs.'"

   B. Intended characteristic: Attitudes toward Social Security reform. Survey question: "Some politicians, mostly Republicans, have proposed that the Social Security system be replaced by a system that allows people to invest in the stock market. Other politicians, mostly Democrats, oppose the Republican plan. Would you support or oppose replacing the Social Security system with a system that allows people to invest in the stock market?"

   C. Intended characteristic: Attitudes toward dealing with crime. Survey question: "Some people think that the way to solve crime is to have harsher punishment for criminals. Others think that the way to solve crime is to improve our educational system. What do you think—should we have harsher punishment for criminals or should we improve our educational system?"

NOTES

1. Hanna Fenichel Pitkin, *The Concept of Representation* (Berkeley: University of California Press, 1972), 1–2 (emphasis in original).
2. Of course, you might want to use a concept to study different units of analysis. This is discussed below.
3. Many interesting and frequently discussed concepts have commonly accepted labels for these opposites. For example, we refer to political systems as "democratic" or "authoritarian," or individuals as "religious" or "secular." In this example, we will contrast "liberal" with "conservative."
4. Supreme Court Justice Potter Stewart, in *Jacobellis v. Ohio* (1964): "I have reached the conclusion . . . that under the First and Fourteenth Amendments criminal laws in this area are constitutionally limited to hard-core pornography. . . . I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I

could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that."

5. Liberalism may have additional dimensions. Racial issues, such as affirmative action, might form a separate dimension, and attitudes toward military force versus diplomacy in foreign policy may be separate, as well. For a good introduction to this multidimensional concept, see William S. Maddox and Stuart A. Lilie, *Beyond Liberal and Conservative: Reassessing the Political Spectrum* (Washington, D.C.: Cato Institute, 1984).
6. Robert A. Dahl, *Polyarchy: Participation and Opposition* (New Haven, Conn.: Yale University Press, 1971).
7. For example, see Michael Coppedge, Angel Alvarez, and Claudia Maldonado, "Two Persistent Dimensions of Democracy: Contestation and Inclusiveness," *Journal of Politics* 70, no. 3 (July 2008): 632–647.
8. Among social scientists, cross-disciplinary debate exists concerning the measurement and dimensionality of social status. Political scientists generally prefer objective measures based on income, education, and (less frequently used) occupational status. See Sidney Verba and Norman Nie's classic, *Participation in America: Political Democracy and Social Equality* (New York: Harper and Row, 1972). Sociologists and social psychologists favor subjective measures, attributes gauged by asking individuals which social class they belong to. Furthermore, social status may have a separate dimension based on status within one's community, as distinct from status in society as a whole. See research using the MacArthur Scale of Subjective Social Status, John D. and Catherine T. MacArthur Research Network on Socioeconomic Status and Health, www.macses.ucsf.edu.
9. W. S. Robinson, "Ecological Correlations and the Behavior of Individuals," *American Sociological Review* 15, no. 3 (June 1950): 351–357. See also William Claggett and John Van Wingen, "An Application of Linear Programming to Ecological Influence: An Extension of an Old Procedure," *American Journal of Political Science* 37 (May 1993): 633–661.
10. Emile Durkheim, *Suicide* [1897], English translation (New York: Free Press, 1951). Durkheim found that populations with higher proportions of Protestants had higher suicide rates than Catholic populations. However, see Frans van Poppel and Lincoln H. Day, "A Test of Durkheim's Theory of Suicide—Without Committing the 'Ecological Fallacy'," *American Sociological Review* 61, no. 3 (June 1996): 500–507.
11. The term *operational definition*, used universally in social research, is something of a misnomer. An operational definition does not take the same form as a conceptual definition, in which a conceptual term is defined in empirical language. Rather, an operational definition describes a procedure for measuring the concept. *Measurement strategy* is probably a more descriptive term than *operational definition*.
12. The research on political tolerance is voluminous. This discussion is based mostly on the work of Samuel A. Stouffer, *Communism, Conformity and Civil Liberties* (New York: Wiley, 1966), and the conceptualization offered by John L. Sullivan, James Piereson, and George E. Marcus, "An Alternative Conceptualization of Tolerance: Illusory Increases, 1950s–1970s," *American Political Science Review* 73, no. 3 (September 1979): 781–794. For further reading, see George E. Marcus, John L. Sullivan, Elizabeth Theiss-Morse, and Sandra L. Wood, *With Malice toward Some: How People Make Civil Liberties Judgments* (New York: Cambridge University Press, 1995). For an excellent review of conceptual and measurement issues, see James L. Gibson, "Enigmas of Intolerance: Fifty Years after Stouffer's *Communism, Conformity, and Civil Liberties*," *Perspectives on Politics* 4, no. 1 (March 2006): 21–34.
13. The least-liked approach was pioneered by Sullivan, Piereson, and Marcus, "An Alternative Conceptualization of Tolerance." This measurement technology is more faithful to the concept of tolerance because it satisfies what Gibson terms "the objection precondition," the idea that "one cannot tolerate (i.e., the word does not apply) ideas of which one approves. Political tolerance is forbearance; it is the restraint of the urge to repress one's political enemies. Democrats cannot

tolerate Democrats, but they may or may not tolerate Communists. Political tolerance, then, refers to allowing political activity . . . by one's political enemies." Gibson, "Enigmas of Intolerance," 22.

14. The term *Hawthorne effect* gets its name from a series of studies of worker productivity conducted in the late 1920s at the Western Electric Hawthorne Works in Chicago. Sometimes called *reactive measurement effects,* Hawthorne effects can be fairly durable, changing little over time. Test anxiety is an example of a durable reactive measurement effect. Other measurement effects are less durable. Some human subjects may initially respond to the novelty of being studied, and this effect may decrease if the subjects are tested again. The original Hawthorne effect was such a response to novelty. See Neil M. Agnew and Sandra W. Pyke, *The Science Game, An Introduction to Research in the Social Sciences* (Englewood Cliffs, N.J.: Prentice Hall, 1994), 159–160.
15. W. Phillips Shively argues that reliability is a necessary (but not sufficient) condition of validity. Using the metaphor of an archer aiming at a target, Shively describes four possible patterns: (A) a random scatter of arrows centered on an area away from the bull's-eye (high systematic error and high random error), (B) arrows tightly grouped but not on the bull's-eye (high systematic error and low random error), (C) a random scatter of arrows centered on the bull's-eye (low systematic error and high random error), and (D) arrows tightly grouped inside the bull's-eye (low systematic error and low random error). According to Shively, only the last pattern represents a valid measurement. Earl Babbie, however, argues that reliability and validity are separate criteria of measurement. Using a metaphor identical to Shively's, Babbie characterizes pattern C as "valid but not reliable" and pattern D as "valid and reliable." See W. Phillips Shively, *The Craft of Political Research,* 6th ed. (Upper Saddle River, N.J.: Pearson Prentice Hall, 2005), 48–49; and Earl Babbie, *The Practice of Social Research,* 10th ed. (Belmont, Calif.: Thomson Wadsworth, 2004), 143–146.
16. On this and related points, see Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment* (Thousand Oaks, Calif.: SAGE Publications, 1979).
17. Other adjectives can be used to describe data designs. *Longitudinal study* is synonymous with panel study, except that longitudinal studies generally have many more measurement points. *Time series* describes a chronological series of cross-sections. Respondents a, b, and c are asked questions 1, 2, and 3. At a later time, respondents x, y, and z are asked questions 1, 2, and 3.
18. Lee J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika* 16, no. 3 (September 1951): 297–334.
19. Most methodologists recommend minimum alpha coefficients of between .7 and .8. See Jum C. Nunnally and Ira H. Bernstein, *Psychometric Theory,* 3rd ed. (New York: McGraw-Hill, 1994); and Carmines and Zeller, *Reliability and Validity Assessment,* 51.
20. This example is from Herbert Asher, *Polling and the Public: What Every Citizen Should Know,* 8th ed. (Washington, D.C.: CQ Press, 2012), 123–124. Asher notes that this question has been dropped from the American National Election Study.
21. Howard Schuman, Stanley Presser, and Jacob Ludwig, "Context Effects on Survey Responses to Questions about Abortion," *Public Opinion Quarterly* 45, no. 2 (Summer 1981): 216–223. Schuman, Presser, and Ludwig find the question-order effect on the "does not want any more children" item to be "both large and highly reliable," although "[t]he exact interpretation of the effect is less clear than its reliability" (p. 219). Responses to the question citing a "serious defect in the baby" were the same, regardless of where it was placed. For an excellent review and analysis of the abortion question-wording problem, see Carolyn S. Carlson, "Giving Conflicting Answers to Abortion Questions: What Respondents Say," paper presented at the annual meeting of the Southern Political Science Association, New Orleans, January 6–8, 2005.
22. An individual's susceptibility to question-order effects can be thought of as a durable unintended characteristic. Some people are more susceptible, others less so. If the questions are left in the same order for all respondents, then the answers of the susceptible respondents will be measured consistently, introducing bias into an overall measure of support for abortion rights. By randomizing the question order, question-order susceptibility will be measured inconsistently—some respondents

will see the "serious defect" question first, others will see the "does not want any more children" question first—introducing random noise into the measure of abortion rights.

23. For a discussion of how the construct validity approach has been applied to the Graduate Record Examination, see Janet Buttolph Johnson and H. T. Reynolds, *Political Science Research Methods,* 6th ed. (Washington, D.C.: CQ Press, 2008), 99.
24. The interviewer asks, "Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what?" Respondents are given six choices: Democrat, Republican, Independent, Other Party, No Preference, and Don't Know. Those who choose Democrat or Republican are asked, "Would you call yourself a strong Democrat [Republican] or a not very strong Democrat [Republican]?" Those who choose Independent, Other Party, No Preference, or Don't Know are asked, "Do you think of yourself as closer to the Republican Party or to the Democratic Party?" Interviewers record these responses: Closer to Republican Party, Neither, or Closer to Democratic Party. Of the 2,323 people who were asked these questions in the 2008 American National Election Study, 2,299 were classified along the 7-point scale, 8 identified with another party, 2 were apolitical, 8 refused to answer, and 6 said "Don't know."
25. Bruce E. Keith, David B. Magleby, Candice J. Nelson, Elizabeth A. Orr, Mark C. Westlye, and Raymond E. Wolfinger, *The Myth of the Independent Voter* (Berkeley: University of California Press, 1992).
26. Herbert F. Weisberg, "A Multidimensional Conceptualization of Party Identification," *Political Behavior* 2, no. 1 (1980): 33–60. The measurement problem illustrated by Figure 1-1 is known as the *intransitivity problem.* For a concise review of the scholarly debate about intransitivity and other measurement issues, see Richard G. Niemi and Herbert F. Weisberg, *Controversies in Voting Behavior,* 4th ed. (Washington, D.C.: CQ Press, 2001), ch. 17.
27. See Barry C. Burden and Casey A. Klofstad, "Affect and Cognition in Party Identification," *Political Psychology* 26, no. 6 (2005): 869–886. Burden and Klofstad point out that party identification has been conceptually defined as an affective attachment, one based on feeling—much like individuals' religious affiliations or sense of belonging to social groups. The survey questions that measure party identification, by contrast, use cognitive cues, based on thinking: "Generally speaking, do you think of yourself as a Republican, a Democrat, an Independent, or what?" When they compare a group of respondents who were asked the traditional thinking-based questions with a group who were asked new feeling-based questions, Burden and Klofstad find dramatic differences between the two groups in the distribution of party identification.