chapter 4

From Abstract to Concrete: Operationalization and Measurement

Every quality manifests itself in a certain quantity, and without quantity there can be no quality. To this day many of our comrades still do not understand that they must attend to the quantitative aspect of things—the basic statistics, the main percentages and the quantitative limits that determine the qualities of things. They have no "figures" in their heads and as a result cannot help making mistakes.

Mao Tse-tung

Empirical research is a means of obtaining answers to questions about reality. Our questions may be primarily practical, or they may be principally of academic interest. In either case, they will probably be stated in abstract terms. Yet the answers we want are usually concrete and specific. One of the first problems in research is to devise ways of getting from the abstract level of our questions to some concrete observations that will allow us to answer them.

To take a nonpolitical example, suppose we want to resolve a debate about which of two professional quarterbacks is the greater athlete. Obviously we will need to compare the two in some way to settle the argument. But on what grounds shall we compare them? We want to determine which exhibits more of the qualities of a great athlete, but *athletic greatness* is an abstract concept. In order to evaluate each quarterback in terms of this quality, we will have to *quantify* the concept of athletic greatness. We might agree to count the passes they complete in televised games, to divide that number by the passes they attempted, and to let the resulting quantity stand for athletic greatness. Or, more likely, we might perform several such operations on different aspects of the players' performance so that we can get a more complete picture of how well they play the quarterback position, and then combine them in some way. Once we have these numbers, we will be ready to make concrete comparisons and resolve the dispute.

What we have just described is essentially the process by which we proceed from abstract concept to concrete observation in social science research. It is a crucial phase in the research process, for only if it is done correctly will the information we gather represent evidence about the utility of our theories or provide answers to our questions. The process of selecting observable phenomena to represent abstract concepts is known as **operationalization,** and the specification of steps to take in making observations is called **instrumentation.** The application of an instrument to assign numerical values to cases results in a **measurement,** and it is this measurement that we finally use as evidence in making decisions and answering questions.

In this chapter we describe these processes in detail and discuss the problems that can be encountered in attempting to operationalize and measure concepts. When you complete the chapter, you should be ready to state the explanations you have devised from your search of the literature in a form that will allow you to test them through actual observations. The activities discussed here are the last steps to be taken before devising a research design and entering the field to collect data.

OPERATIONALIZATION: THE LINK BETWEEN THEORY AND OBSERVATION

In Chapter 2 we stressed the importance of having a theory to guide observation. We described the research process principally as a matter of comparing actual observations with the expectations about reality that we derive from our theories in order to judge how much we can rely on the theories as explanations of political phenomena. These expectations are stated in the form of hypotheses, which predict relationships between those variables that represent the concepts in the theory. The object of this chapter is to describe how we can devise observations that will make these comparisons possible. The question is how we can quantify our concepts in order to make precise statements about whether or not our theoretically derived expectations are supported by what we observe.

The problems encountered in doing this in the social sciences are basically the same as those encountered in the physical sciences. A simple example will help make some of the issues clear. Let us say that we want to test the hypothesis that a chemical fertilizer spread in one cornfield will stimulate more growth than the natural nutrients found in another field. Growth is an abstract concept. We cannot see it directly. We need to translate *growth* into an empirically observable variable so that we can determine when one plant has shown more of it than another.

We can let the variable *height attained* represent the concept *growth* because relative heights are empirically observable. But corn plants don't wear signs telling their height; we have to ascertain it for ourselves. But how? We can use human judgment and have observers rate plants in the two fields as tall or short. Such a procedure allows only crude comparisons between plants, however, and is subject to all kinds of errors, because people differ in their perceptions. We need a more precise and dependable means of determining heights if we are to make meaningful comparisons.

What we must do is to translate the variable height into terms of some *measuring instrument* that can be used to yield precise, standardized indications of the extent to which the characteristic is embodied in individual corn plants. We can let

height be represented by an **indicator,** such as *length in inches,* and measure the plants with a tape measure. The readings from the tape measure then become the **values** we assign to plants on the variable *height,* and these values are what we actually compare in attempting to assess the accuracy of our prediction of greater growth in one field than in another.

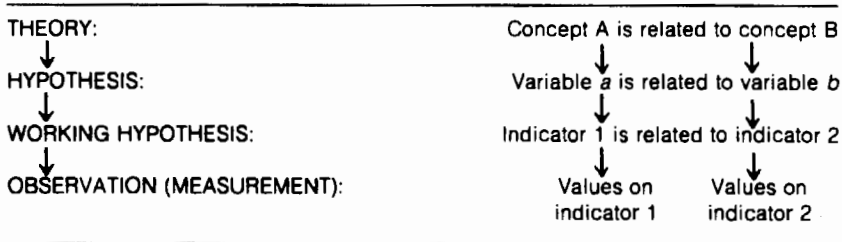
We have moved, then, from the abstract concept *growth* to the variable *height* and then to the indicator *length in inches.* This transformation is what is known as *operationalization,* because we have reduced an abstract concept to a set of values that can be obtained through specifiable operations. We finally make the comparisons on which we will judge the accuracy of our hypothesis about relative growth by comparing the values that result from the measurement process (in this example, the readings from the tape measure). When we speak of **observation** in research, we are referring to *the process of applying a measuring instrument in order to assign values for some characteristic or property of the phenomenon in question to the cases being studied.* In other words, observation means using an instrument to measure a trait or behavior.

This is an important point to understand. It makes clear the significance of operationalization and measurement in the research process. We can never actually compare concepts, even though our theories and often our research questions will be stated in concepts. What we compare are *indicators of concepts.* We cannot compare the growth of plants in the two cornfields in our example. We can compare only the readings we get from the tape measures—the measures produced by the indicator that we have decided to let represent the concept.

This means that *our comparisons can be accurate only to the extent that the indicators selected mirror the concept we intend them to measure.* If we have improperly operationalized our concepts, the relationship between our indicators may not be an accurate reflection of the relationship between the concepts they are supposed to represent. As a result, any conclusions we draw from our observations about the concepts or the theory of which they are a part will be faulty.

Figure 4.1 illustrates this situation. Our theory posits a relationship between two abstract concepts. Our hypothesis predicts a relationship between two empirically observable variables, which we reduce to measurable indicators, and our observations reveal a relationship (or lack of relationship) between two sets of values on these indicators. Obviously we can infer something about the reality of the theoretical relationship only if the variables accurately represent the concepts *and* the indicators ac-

FIGURE 4.1 Operationalization: The relationships of concept, variable, and measure



curately represent the variables. *Operationalization almost inevitably involves some simplification or loss of meaning,* since indicators seldom reflect all that we mean by a concept. Though we almost always have to accept some loss of meaning, we need to operationalize so as to minimize that loss. We have to seek indicators that capture as much of the meaning of the concept as possible and that represent at least some aspects of our concepts as accurately as possible.

We can see the implications of this in our agricultural example. Once we have begun the research, we may realize that there is more to the concept *growth* than height and that the indicator *length in inches* does not fully capture what we want to measure. For instance, it may be that the amount of growth in the two fields is substantially different but all the difference is in stalk diameter, width of leaves, and weight of corn ears; the height of the plants in the two fields may not be noticeably different. In that case, if we look only at height in evaluating the effects of the fertilizer, we will be seriously misled, because the link between the concept (growth) and the variable that represents it (height) is imperfect. The variable used here does not *fully operationalize* the concept it represents. It does not capture all the meaning in the concept, and using it misleads us about the relationship that exists in the real world.

This is an especially common situation in the social sciences since most important social science concepts are **multidimensional** in that they have more than one aspect or component. Our measures of these concepts must reflect their diversity if they are to be useful indicators of the concepts. For example, if we operationalized the concept *democracy* only in terms of the holding of regular elections, we might classify dictatorial regimes that hold elections with only one candidate per office and do not allow freedom of expression as being just as democratic as the nations of Western Europe. To obtain an accurate measure of the degree to which nations are democratic, we obviously need indicators that reflect the various dimensions of the concept.

This example should clarify why operationalization is crucial to theory testing and the research process in general. It is not as easy to explain how to ensure proper operationalization. This is because selecting variables to represent concepts and devising indicators for the variables both involve a good deal of creativity and cannot be reduced to a set of standardized steps that will unerringly produce good measures. What we can do is to point out some of the pitfalls to be avoided in the process and to describe ways of evaluating the adequacy of operationalizations once they have been selected. We do that in the sections on measurement in this chapter.

OPERATIONAL DEFINITIONS

Before moving to a discussion of social science measurement, however, we should consider what is involved in operationalizing a concept. This is done by specifying a set of procedures to be followed or operations to be performed in order to obtain an empirical indicator of the manifestation of a concept in any given case. These procedures then provide an **operational definition** of the concept and its variable counterpart. The process of operationalization essentially reduces to a matter of selecting operational definitions for concepts.

To be useful (that is, to provide valid and reliable measures of our concepts), operational definitions must tell us precisely and explicitly what to do in order to deter-

mine what quantitative value should be associated with a variable in any given case. They should specify a complete set of steps to take in the process of measurement.

We want to be precise in this for at least three reasons. First, we want to be able to tell others exactly what we have done to obtain our measures, so that they can evaluate our work and possibly repeat our study to verify its results in another setting. Second, if we have assistants actually gathering the information, we will want our instructions to them to be detailed and precise enough to ensure that each one takes the measurements in exactly the same way as the others do. If our instructions are vague and our assistants go through slightly different sets of steps in obtaining measures, their results will not be comparable and we will be unable to draw valid conclusions from them. Finally, precise and detailed statements of how to operationalize a variable will help us in evaluating the results we obtain and in eliminating rival explanations of those results that essentially claim that the "findings" have been produced by flaws in the measurement process. (We have more to say about this in subsequent sections of this chapter.)

Thinking through the following hypothetical example should help you appreciate what is involved in developing an operational definition. Suppose a university has hired you to assess the determinants of students' academic performance in order to decide whether or not to alter its admission criteria.

For simplicity, suppose you elect to let the variable *grade point average* stand for the concept *academic performance.* After selecting the colleges and students to be included in the study, you must devise procedures for taking the appropriate measurements and instruct assistants in how to apply those procedures. To secure data on the dependent variable (academic performance), you will have to specify where to find the grades to be used, how to compute a grade point average, and how to record that information on a form. To secure a measure of the independent variables, you will have to provide a precise statement of the questions to be asked and specify exactly how to record the various possible responses to those questions.

When devising operational definitions for variables in research, you should routinely write out a description of the procedures you will use to obtain measurements. Every step should be detailed. This not only provides a record of your research and ensures standardization of measuring procedures but also gives you an opportunity to think through the act of obtaining a measurement in order to discover possible errors that can damage the reliability of results.

Suppose we want to measure the degree to which members of the two major parties support their own party in a state legislature. We can operationalize the concept *party unity* as *voting together on roll call votes* and then use the percentage of the average member's votes that agree with those of the majority of his or her party as our indicator of *voting together.* Having decided to do this, however, we face a number of critical choices in actually operationalizing our variable.

We can get information on how each legislator votes from the records of the legislature, but we will then have to decide which of the many recorded votes to include in our count. Some votes are unanimous (such as a vote to issue a proclamation of praise for some national hero) and do not reflect party unity because they do not involve partisan issues. Including all votes reduces the extent to which our measure reflects our concept. We have to state criteria for selecting votes to include. In order to focus only on controversial issues, therefore, we might, for instance, choose to in-

clude only those roll calls in which at least two-thirds of the legislators vote and in which the losing position gets no less than 30 percent of the vote.

We also have to decide how to devise a procedure for determining how a majority of the party has voted in order to classify each member's votes as consistent or inconsistent with that majority position. We need to decide how to treat abstentions. Do they count as a failure to support the party, or do we exclude them from our count? In addition, we have to specify a procedure for first computing and then averaging the percentages of agreeing votes for each legislator.

With every operationalization, we face similar decisions about exact procedures to follow in obtaining measures. A complete operational definition reveals how we have decided to handle such problems and leaves no ambiguity about what we actually did in taking our measures.

Constructing an operational definition results in the development of an **instrument** for taking measurements. In the physical sciences, such instruments as scales, light meters, and micrometers are used to obtain indicators of the degree to which things exhibit some property. In the social sciences, measuring instruments take different forms. Typical social science instruments include a series of questions on a survey form, instructions on how to make and record observations of certain events such as a debate on the floor of the United Nations, and sets of numbers to be taken from some sourcebook and the rules for combining them into a measure.

Proper instrumentation is as important in the social sciences as it is in the physical sciences. Just as we would not attempt to measure weight with a ruler, we would not want to measure *political alienation* with a series of questions that do not reveal how alienated people feel. In discussing the validity and reliability of measures in the next section, we also suggest some ways to test the instruments developed in the process of operationalization in order to increase our confidence that they measure what we want them to.

MEASUREMENT

We operationalize variables in order to have a way to quantify abstract concepts so that we can make meaningful comparisons between real-world phenomena in terms of the properties suggested by those concepts. This assigning of numerals to represent properties is known as *measurement.*[1] The result of measuring is that we have a *value* to associate with some variable for a given case.[2] This means simply that we can speak with more precision about the extent to which a given unit of observation (for example, a person, a city, a nation, or an organization) exhibits the property represented by the variable being measured. Rather than say that a city has a "bad crime problem," we can

[1] This definition of measurement is expanded in Norman Campbell, *What Is Science?* (New York: Dover, 1952), p. 110.

[2] It is crucial that we appreciate the difference between a variable and its *values.* We recognize a variable because of its capacity for taking on different values. The variable is a concept translated into empirical terms. A value is some magnitude or quality of the variable that individual cases can reflect. For example, 23 years is a value for the variable *age*; $25,000 is a value for the variable *annual income*; 12 percent is a value for the variable *percentage of population foreign-born*; and upper class is a value for the variable *socioeconomic status.*

speak of specific crime rates. Rather than say that a person is a "devoted Republican," we can say that one has scored a 5 on our *strength of party identification* measure.

LEVELS OF MEASUREMENT

Measuring procedures provide a means of categorizing and ordering phenomena. Some procedures, however, produce more precise and detailed distinctions between events than do others. Because of this we speak of various **levels of measurement.** When we say a procedure produces a given level of measurement, we are classifying it according to how much information it gives us about the phenomena being measured and their relationship to one another. The levels of measurement are referred to as *nominal, ordinal,* and *interval.*

Nominal measurement provides the least information about phenomena. It gives us only a set of discrete categories to use in distinguishing between cases. Nominal measurement is obtained by simply naming cases by some predetermined scheme of classification. Nationality is generally "measured" at the nominal level by classifying people as British, Swiss, Brazilian, and so on. However, that "measurement" neither tells us how *much* of the characteristic "nationality" different individuals have nor allows us to rank-order them. Using nominal measurement simply gives us a way of sorting cases into groups designated by the names used in a classificatory scheme.

To be useful, nominal measurement schemes must be based on sets of categories that are **mutually exclusive** and **collectively exhaustive.** This means (1) it must not be possible to assign any single case to more than one category and (2) the categories should be set up so that *all* cases can be assigned to some category. If we want to classify voters in the United States by use of a nominal measuring scheme, we cannot use the categories *Democrat, Republican, liberal,* and *conservative* successfully, because these categories are not mutually exclusive. Because U.S. political parties each appeal to a broad spectrum of voters, it is possible for a person to be both a Democrat *and* a conservative or liberal, or both a Republican *and* a conservative or liberal. The categories do not allow us to differentiate among voters in all cases. Similarly, if we try to categorize voters by party affiliation using only two categories—*Republican* and *Democrat*—we will find that our categories are not collectively exhaustive, because some voters consider themselves independents or members of other parties.

In order to facilitate analysis, we will probably want to substitute a number for each category in a scheme of nominal measurement. It is important to recognize, however, that such numbers have no real meaning in this context; they are simply symbols. Just because we choose to substitute a 5 for the *Republican* category and a 1 for the *Democrat* category, we *cannot* assume that Republicans have five times as much party affiliation as Democrats. Any number can be substituted for any category of a nominal measurement so long as each category has a unique number associated with it.

Ordinal measurement provides more information because it allows us both to categorize and to order, or rank, phenomena. Ordinal measurement allows us to associate a number with each case. That number tells us not only that the case is different from some other cases and like still others with respect to the variable being measured but also how it relates to those other cases in terms of how much of a particular

property it exhibits. With ordinal measurement we can say which cases have more (or less) of the measured quality than other cases, and we can *rank* cases in the order of *how much* of the quality they exhibit. That ranking gives us more detailed and precise information about the cases than we would get from a nominal measurement. The concept *social class* is usually measured at the ordinal level, with individuals being ranked as lower-, middle-, or upper-class.

Interval measurement provides even more information. Not only can we classify and rank-order cases when they have been measured at the interval level, but we can also tell *how much* more (or less) of the measured property they contain than other cases. Ordinal measurement is not based on any standardized unit of the variable in question and does not allow us to tell how far cases are from one another in terms of that variable. It allows us only to say that some have more or less of it than others. Interval measurement is based on the idea that *there is some standard unit of the property being measured.*

Whereas ordinal measures give us only a rough idea of the relationship between cases with respect to a variable, interval measures provide information on the "distance" between cases. The variable *income* is a clear example. Income is usually measured in units of currency (dollars in the United States). Because we can use *standard units* in our measurement, we can say that the difference in income between $10,000 and $11,000 a year is exactly the same as the difference in income between $50,000 and $51,000 a year. We cannot do that with ordinal measurement. If we measure income ordinally by dividing people into such income categories as *under $10,000* and *$10,000 to $19,999*, we can say that one person has more or less income than another, but we cannot say exactly how far apart they are in income because we cannot tell where an individual falls in the category. The income difference between a person in category 1 (under $10,000) and a person in category 2 ($10,000 to $19,999) can be as little as one dollar ($10,000 minus $9,999) or as much as $10,000 ($19,999 minus $9,999), depending on their exact incomes, but we cannot make this distinction from an ordinal measure.

In addition to giving us precise information on the absolute differences between cases, interval measurement lets us make accurate statements about the relative differences between concepts. We can, for instance, agree that 50,000 people is twice as large a population as 25,000 people because we can speak meaningfully of a place that has no population. There is a *zero point* in true interval measures, and it is at least conceivably possible for a case to score zero on such measures. Because there is no meaningful zero point on an ordinal scale, we cannot say, for example, that upper-class people have twice as much "class" as lower-class people, because we don't know what it means to have no class standing.

This suggests an important point about levels of measurement. Nominal-level measurement is the least useful form of measurement when we have to compare phenomena. If we use it when we can use a "higher" (more precise) level of measurement, we may be wasting potentially valuable information. If, in a study of voting behavior, we categorize people only as Republicans, Independents, and Democrats when we can ask a different set of questions and produce a rank ordering of them as strong to weak party identifiers, we may be giving up information that will help us understand the relationships we observe. Ordinal-level measurement is more useful than nominal, but it too has limitations. Interval-level measurement is the most desirable

form of measurement both because of the amount of detail in the information it provides and because of the mathematical procedures it allows us to perform on our data. (We say more about this in Chapters 16, 17, and 18.)

The point is that we should strive for operationalizations that allow interval level measurement whenever possible and appropriate. But how do we decide which level of measurement is appropriate for the particular concepts we want to operationalize? This is a matter of both conceptualization and measurement technology.

In the theory-building stage of research, we must first ask ourselves if there is some continuum underlying the differences we see in cases. If there is, we can devise ordinal and even interval measurements for a concept that might otherwise be measured only by nominal classification. An example will help clarify the significance of this.

Suppose we are studying the effects of immigrants' nationality on the degree of their support for big-city political machines in the early twentieth century in the United States. If we operationalize nationality at the nominal level and categorize city voting precincts' support for the machine, we might get a picture like that presented in Figure 4.2*a*. There is no apparent relationship between nationality and voting behavior, because knowing a precinct's dominant nationality does not help us rank it relative to the others. If we examine our reasoning, however, we might decide that the reason we expect nationality to be related to support for the machine is that countries of origin differ in the opportunities they allow their citizens for political participation. When people have had little experience with democratic politics in their native land, we might reason, they will be more willing to give up to a political boss their right to self-government. If we can follow this reasoning and rank-order the nations of origin by the extent they allow their citizens political participation, we can construct a graph like that shown in Figure 4.2*b*. In that graph, a relationship between nationality and support for the machine is apparent. The ordering of categories on our independent variable helps us discover a pattern in its relationship to the dependent variable.

If we are bold enough, we may even upgrade our measurement of the independent variable to an interval level. For example, we might count the number of legal provisions for political participation in the statutes of each country in question for the years just prior to the beginning of significant immigration to the United States. We can use the resulting numbers to rank nationalities along an interval scale and make even more precise comparisons of independent and dependent variables.

Whether or not we can achieve this upgrading of variables from the nominal to the ordinal or interval level depends both on developing a theoretical rationale for doing so and on the technical possibility of applying the operational procedures that produce the higher-level measurements. Even if we can conceptualize *nationality* in interval terms in our example, we may not have access to the legal records necessary to place countries along the interval scale. In that case, *measurement technology* limits what we can do to strengthen our measures. We may encounter many such cases. We may be investigating the relationship between sex and political behavior, for example, and be able to argue that maleness is a characteristic that people possess in degrees (those with the least of it being females). If we can come up with a set of questions to reveal how much (if any) of this quality people have, we can, in principle, rank people on an ordinal or interval scale of maleness. If, however, we do not have

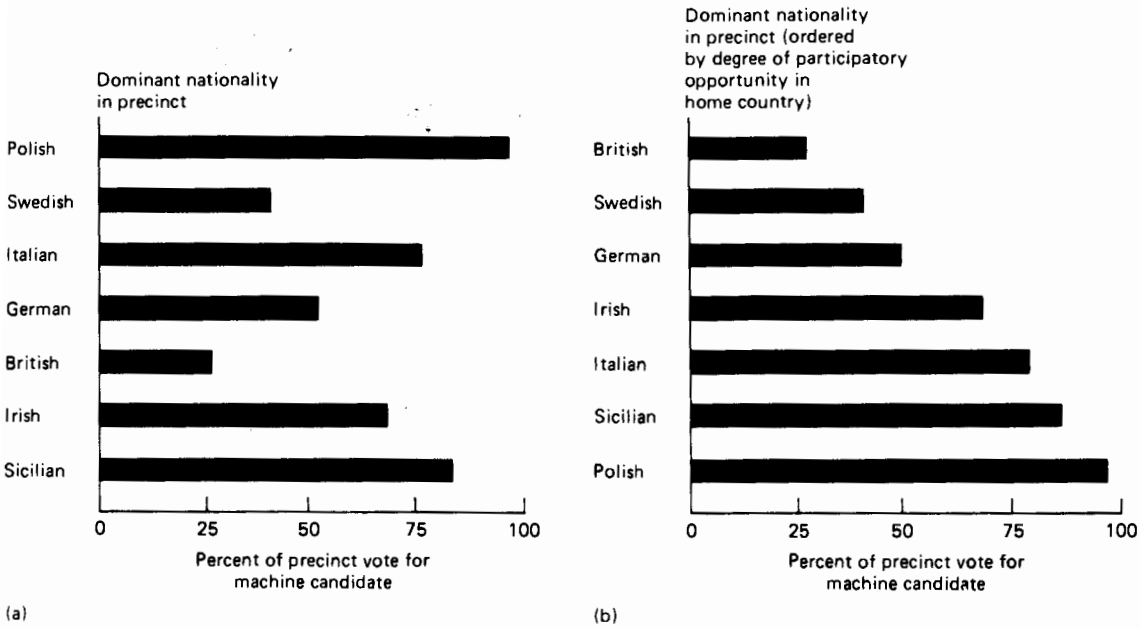


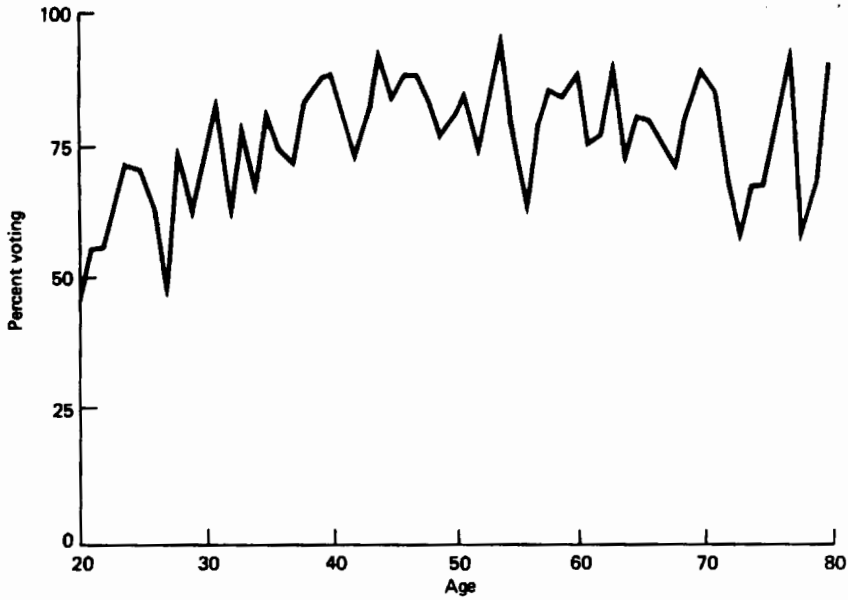
FIGURE 4.2 An example of how level of measurement can affect the interpretation of data

the research funds to do a survey to get people to answer the questions, we may have to rely on the nominal classification *male* or *female* that we find in the records of party membership.

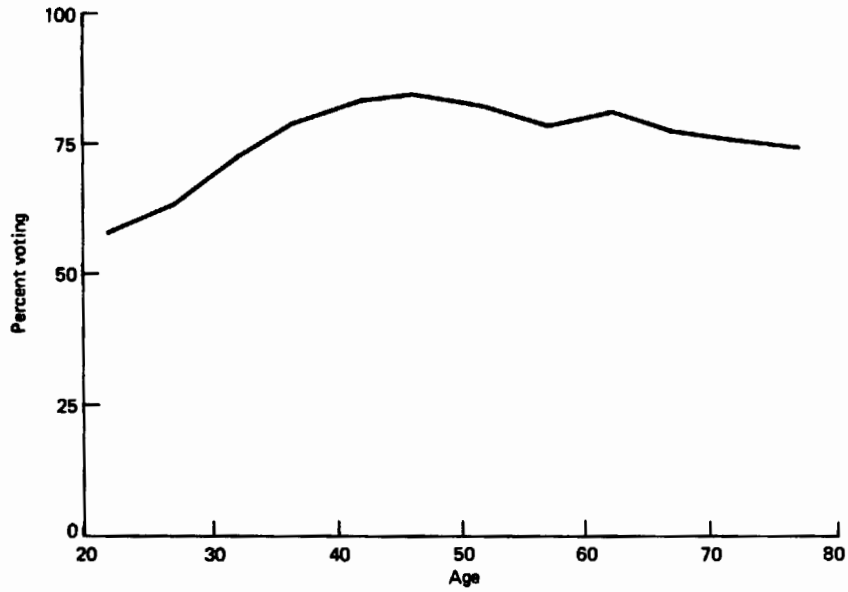
These situational factors make it difficult to set down rules about operationalizing concepts to achieve certain levels of measurement. We can, however, suggest that W. Philips Shively's rule for operationalization is a good one to follow.[3] Shively advises that we use measures as precise as possible given the subject we are studying and that we not waste information by imprecise measurement. This generally means upgrading measurement procedures so that they yield higher levels of measurement whenever possible. Do not settle for an operationalization that produces nominal measurement when ordinal or interval measurement is theoretically defensible and technically possible.

Having said this, we need to add a qualification to the general rule. There are cases in which too much precision in measurement is actually undesirable. Shively offers an example of this. Figure 4.3 is adapted from his work. In it we see the relationship between age and voting in the 1968 presidential election presented in two different ways. In Figure 4.3*a*, age is measured in single years. Because there are so few people in each age group (for example, 21-22, 35-36, 50-51), the chart reveals no clear pattern in the relationship—between the two variables. In Figure 4.3*b*, age is measured less precisely, in five-year groupings. With more cases in each group, we can see that there is a broad pattern to the relationship, with voting likelihood increasing to age 50 and then generally declining.

[3] W. Philips Shively, *The Craft of Political Research*, 2d ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1980), pp. 66, 80.



(a) Age and participation in 1968 election: age measured by years



(b) Age and participation in 1968 election: age measured by half-decades

FIGURE 4.3 An example of the effect of grouping data on interpretation

SOURCE: W. Philips Shively, *The Craft of Political Research*, 2d ed. Englewood Cliffs, N.J.: Prentice-Hall, 1980, pp. 67–68. Reprinted with permission.

By giving up some precision in our measurement, we have gained a greater ease of analysis. This is a good trade as long as we do not go so far in the direction of less precision that we again lose sight of relationships. If we use 20-year groupings to measure age, we will see little difference in the percentage of each age group that votes and might falsely conclude that age is unrelated to the likelihood of voting. Because we generally do not know in advance of actual data analysis how much precision will be needed to allow us to discover relationships, we should follow the rule of operationalizing our concepts as precisely as possible. We can always discard unnecessary precision by "collapsing categories" (moving to larger units of differentiation) if we find it necessary. But if we do not collect the information in the first place, we cannot draw on it later.

WORKING HYPOTHESES

Measurement assigns values to cases with respect to given variables. These values are what we use to represent concepts when we compare observations. Before we can understand the implications our observations have for our theories, we have to translate our hypotheses concerning relations between variables into **working hypotheses,** which state the expected relationships between measures or indicators. The next-to-last line in Figure 4.1 suggests the form that working hypotheses take. These hypotheses force us to state what linkages between indicators and variables we believe our operationalization has produced.

Consider an example from the study of international relations. Suppose we are interested in a theory of dominance in the international sphere. Working from the theoretical proposition *The more dominated a nation is, the more conformist its foreign policy will be,* we can hypothesize as follows: *As a nation's economic dependency increases, its support for the international policies of its patron state will increase.* We can operationalize *economic dependency* as the percentage of the nation's exports that go to the patron country. A percentage of exports becomes our indicator of the independent variable *dependency. Support* can be measured by the percentage of votes in the United Nations General Assembly in which the client nation's vote differs from that of the patron state. A percentage of votes in the United Nations becomes our indicator of the dependent variable *support for the patron state's policies.* We can now set out a working hypothesis stating the negative relationship we expect between indicators: *As the percentage of exports going to the patron state increases, the percentage of votes in the United Nations that disagree with the patron state decreases.*

This working hypothesis tells us what observations are consistent with our hypothesis and our theory. It also suggests the relationship we envision between our variables and our indicators. That relationship is diagramed in Figure 4.4.

The diagram shows how important it is that we think through the relationship between our measures and our variables. The relationship predicted by the proposition and the hypothesis is a positive one. But the relationship predicted by the working hypothesis is *negative.* This is because the relationship between the dependent variable and its indicator is negative. That means is that because of the way we have operationalized our dependent variable, a *negative* relationship between indicators

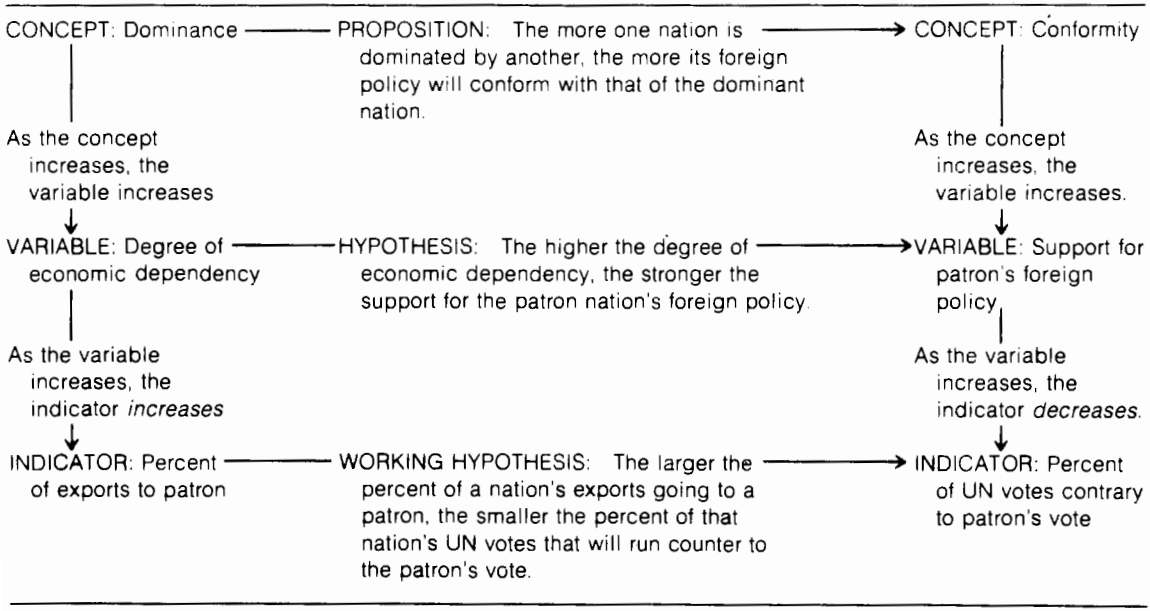


FIGURE 4.4 Specification of the relationships of concepts, variables, and indicators

will provide evidence for a hypothesis and a theoretical proposition that predict *positive* relationships between concepts and variables. We have to be aware of this if we are to avoid misinterpreting our data and if we are to draw accurate conclusions about the utility of our theory from our observations.

Being clear about the relationship between our indicators and the variables and concepts they represent is so important that some social scientists argue that, in addition to our theories about political phenomena, we should be able to state a **measurement theory** that sets out *why we would expect our indicators to be related to our concepts.*[4] Why should we expect economic dependency to be related to concentration of exports? What is there about the distribution of exports that makes it a reflection of what we mean when we refer to dependency? These are the types of questions a well-developed measurement theory helps us answer. A measurement theory consists of the assumptions that explain why our indicators should change values as the degree to which cases manifesting our concepts change.

Indicators cannot be casually selected but must be chosen as a result of careful reasoning about the way things are related in the world. That reasoning is much like what we go through in constructing theories about political phenomena. The conclusions we reach may be wrong. The properties we refer to when we use a concept may in fact be unrelated to the indicators we decide to use as empirical measures for that concept.

This issue of whether or not there is any correspondence between our concepts and variables on the one hand and our indicators or measures on the other is the cen-

[4] For example, see Hubert M. Blalock, Jr., "The Measurement Problem: A Gap between the Languages of Theory and Research," in *Methodology in Social Research,* Hubert M. Blalock, Jr., and Ann Blalock, eds. (New York: McGraw-Hill,1968), pp. 5-27.

tral problem of measurement in science. The question of whether changes in our indicators are actually the result of changes in the concepts they represent gives rise to the problems of reliability and validity that we discuss in subsequent sections of this chapter. The important point here is that the idea of a measurement theory suggests the way we should approach these problems in our research.

Every operationalization of a concept is, in essence, a hypothesis. When we operationalize and say, "Let this indicator stand for that concept," we are hypothesizing that the things we mean when we use the concept are actually reflected in the indicator we select. That hypothesis may or may not be supported by observations. We can never take the adequacy of our measures for granted but must examine our observations in order to find evidence that they represent what we mean by our concepts. The search for valid and reliable measurement procedures in the social sciences is in many ways a process of hypothesis testing. We must be ready to admit mistakes and begin again when the evidence suggests that our indicators fail to reflect our concepts. The testing of measures occurs primarily in our attempts to assess the validity and reliability of the indicators we select.

MEASUREMENT ERROR: THE ENEMY

The process of measurement results in the assignment of a variety of values to different cases on the basis of how they score on our indicators. The differences in the scores can all be attributed to two basic sources. One source is the extent to which the cases really exhibit different degrees or aspects of the property in which we are interested. Different scores occur when our measures actually pick up those differences. In this case, *real* differences in the concept are reflected in our measures. The other source of differences in the scores is the extent to which something about the measure itself or the setting in which it is applied causes different cases to get different scores. In that instance, our measures are showing differences between cases that are *not* real—in the sense that they do not reflect authentic differences in the concept we want to measure. The differences we observe when this happens result from inconsistencies in the procedure for measurement.

If our measures were perfect, they would reveal only the first kind of differences between cases. But our measures are rarely if ever flawless. Differences in the values assigned to cases inevitably reflect not only real differences in the degree to which those cases manifest the concept but also "artificial" differences created by the measurement process. Any differences in the values assigned to cases that are attributable to anything other than real differences are known as **measurement error**. They are not real differences between cases but differences that are erroneously recorded because of flaws in the measuring process.

This distinction between true variations in scores and variations due to measurement error is similar to the distinction between differences in objects apparent to the unaided eye and differences we see when we look only at their reflection in a mirror. To the extent that the mirror distorts the images, it either masks differences we would see with the naked eye or creates an impression of differences we would not otherwise perceive. In the social sciences, we can rarely see our key concepts directly and must rely on measurement procedures analogous to the mirror to reflect these con-

cepts in any given case. Consequently, the accuracy of our impressions of the world depends on the precision with which our measures reflect reality.

What are some of the sources of distortion in the images our measures provide? We need to know the answer to this question if we are to control measurement error or recognize it when it is present in our data. We can list several of the primary sources of measurement error by identifying common sources of differences in the scores assigned to cases *other than true differences in the characteristics we want to measure.*[5]

1. Differences in the distribution of other, relatively stable characteristics among the cases that are unintentionally revealed by our measures: For example, the questions representing our measure of political ideology may require a given level of intelligence to interpret and answer. If this is the case, responses will reflect not only differences in people's political ideology but also differences in their intelligence. When we look at the resulting data, the effects of intelligence and political ideology will be confused, and we will be unable to distinguish differences in scores that reflect ideological differences from those that reflect differences in intelligence. Similarly, other characteristics of our units of analysis (such as the regional location of cities, the cultural traits of nations, or the sources of documents) can be inadvertently reflected in our measures and distort our perceptions of the manifestation of the target concepts. When we can identify and measure these "contaminating" influences, we should check to see whether holding their values constant wipes out, reduces, or increases differences in the scores that cases receive on our measures.[6]
2. Differences in the distribution of temporary characteristics among the cases that are reflected in our measures: A person's mood or state of health can affect the way one responds to items on a questionnaire. The recent political history of cities (the revelation of corruption among public officials, for instance) can create systematic but temporary differences in the way those cities' citizens answer survey questions. A massive natural disaster can produce a drastic but temporary change in the statistics we are relying on to indicate the level of economic development. The effects of such temporary "abnormalities" are more difficult to identify and control than the effects of the stable characteristics in our cases. The only approaches for guarding against them are being alert to signs that individual cases are subject to such transient influences (for example, studying the recent political history of the cities included in our sample or advising our interviewers not to attempt to interview a person who is temporarily bedridden) and following the proce-

---

[5] A more detailed discussion of the points that follow is found in Claire Selltiz, Lawrence S. Wrightsman, and Stuart W. Cook, *Research Methods in Social Relations,* 3d ed. (New York: Holt, Rinehart and Winston, 1976), pp. 165-68. Their discussion comes from the tradition of psychology and pertains primarily to research in which the subjects are people. Many of the principles, however, are transferable to the broader range of research situations we are considering.

[6] Useful discussions of how to do this are found in Donald F. Campbell, "Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity," *American Psychologist,* 15 (August 1960), pp. 546-53; and Morris Rosenberg, *The Logic of Survey Analysis* (New York: Basic Books, 1968).

dures for checking the reliability of measures described in the Reliability section of this chapter.

3. Differences in subjects' interpretation of the measuring instrument: This is a problem only when people must respond directly to questions, as opposed to when the researcher constructs measures by observing behavior. If our questions are ambiguously worded, the different interpretations our respondents place on them can produce differences in their scores on the measures composed of those questions. Suppose, for instance, we are careless enough to ask the question *Did you vote in the last election?* in a study of voting behavior. If some of the interviewees do not know that a city election has been held the prior week, they may answer that they *have* voted, because they think the question refers to the last national election, even though they have not voted in the election to which we intend our questions to refer. We can guard against this source of unintended differences in scores on our measurements by pretesting questions (as discussed in Chapter 7) and testing our measures for reliability.
4. Differences in the setting in which the measure is applied: This is also a source of measurement error principally in research that relies on individuals' responses to questions as its measures. One well-established fact in survey research, for example, is that the race, sex, and age of interviewers can affect responses. Answers (and therefore scores on measures) can differ from interview to interview on the basis of the characteristics of the interviewer alone. Similar problems can arise outside survey research. We may, for instance, make the mistake of doing a content analysis of one country's domestic newscasts and another's newscasts intended only for foreign nations. We will then be applying the same instrument in very different settings and can expect some differences in scores from this fact alone. We can avoid this source of measurement error only by making every effort to see that the situations in which our measures are applied are standardized.
5. Differences in the administration of the measuring instrument: The scores assigned to cases can differ as a result of a variety of errors that occur in collecting and recording information. Interviewers may misunderstand instructions and ask questions in ways the researcher might not intend. Poor lighting may cause a respondent to mismark a questionnaire. Pencils can break and pens run out of ink at crucial moments so that observers fail to record key events in a group interaction. A bored or tired coder might change or ignore the instructions for coding items in a content analysis. These kinds of variation in the administration of measuring instruments cause differences in scores independent of any differences in real values for the variable under investigation. Beyond employing only dependable assistants, the primary way to guard against such sources of measurement error is through *pretesting* our instruments. A trial run will help us discover potential "mechanical" problems with the instrument (such as insufficient space for recording typical answers on a coding form) and human factors that may affect results (such as length of time observers can work without fatigue).
6. Differences in the processing and analysis of data: Information has to undergo a great deal of handling before it can be analyzed. It often changes form

several times. For example, interviewers may record responses by writing down every word an interviewee says in answer to a question. Those written passages may subsequently be reduced to a single number as responses get coded. The written number may be transferred to a computer file as an entry in the appropriate column, and that punch may then be translated into a *bit* on a magnetic disc or tape. In each of these steps, data analysis has been made simpler, but with each step there is a chance of errors that can cause cases to appear to differ on a variable when they do not. The possibility of such errors makes it a good idea to always double- and triple-check each transformation of data and to keep the original form for future reference.

7. Differences in the way individuals respond to the form of the measuring instrument: This is especially a problem when our units of analysis are people. Measuring instruments can take such different forms as oral interviews, questionnaires to be filled out by the subject, and observation by a trained researcher. The different forms place different demands on the people under study. An interview requires ease of speaking, and a questionnaire requires an ability to read and write, for example. If people differ in these abilities, their scores may differ even when the people are actually alike on the variable being operationalized. The best guard against this source of measurement error is the use of more than one form of measure to operationalize each concept. We say more about this in the Validity section of this chapter.

All of these factors can introduce measurement error into our research. The various errors that arise from these seven sources are generally categorized as either *systematic* or *random* errors. **Systematic errors** are those that arise from a confusion of variables in the world (as discussed in item 1 in the preceding list) or from the nature of the instrument itself. They appear in each use of the instrument and are constant among cases and studies in which the same measure is used. Constant errors cause our results to be *invalid,* in that the differences (or similarities) our measures seem to reveal are not accurate reflections of the differences we think we are measuring. **Random errors** affect each application of the instrument differently. They occur as a matter of chance and are due to transient characteristics in our cases, situational variations in application of the instrument, mistakes in administration and processing, and other factors that vary from one use of the instrument to the next. They make our measures invalid in much the same way that systematic errors do. Random errors also make our measures *unreliable,* in that we cannot consistently get the same results when we use the measure if random errors are occurring.

How are we to avoid having measurement errors so distort our results as to render our research useless or misleading? To answer that question we must give careful attention to the issues of validity and reliability.

VALIDITY

We can seldom obtain direct measures of the concepts used in social science theories. Such concepts as power, democracy, and representation cannot be quantified as simply as such concepts as length and weight. We have to use indicators that correspond

only indirectly to the concepts they represent. There is always a chance then that the indicators we choose will not adequately reflect the concepts we want to measure. **Validity** is the term we use to refer to *the extent to which our measures correspond to the concepts they are intended to reflect.* To ask about a measure's validity is to ask if we are in fact measuring what we think we are measuring when we use it. Achieving validity is often viewed as the basic problem of measurement in the social sciences.

To be valid, a measure must be both *appropriate* and *complete.* If, for example, we are interested in comparing the quality of public education in different cities, we may be tempted to use the number of teachers in those cities' schools as an indicator of the quality of educational services. This measure is *inappropriate,* because the number of personnel in a school system is determined largely by the number of students and the size of the city and may have little to do with the quality of education. If we use the ratio of students to teachers as our indicator of educational services, we will have a more appropriate measure, in that differences caused by city size will be reduced or eliminated. The measure, however, will still be *incomplete.* Education involves more than teachers. It also involves school buildings, films, books, labs, and a variety of other factors. Looking at any one of these factors by itself might leave us with a false impression of the total quality of educational services. A school system may have a highly desirable student-teacher ratio but inadequate facilities and learning materials. It is a mistake to say that such a school system is equal to one with an identical student-teacher ratio *and* excellent facilities and learning materials. If we are to achieve validity, we must strive to construct measures that are both appropriate and complete.

This raises two questions: How can we create measures that are complete and appropriate, and how can we tell whether we have succeeded in doing so?

The answer to the first question begins with the operationalization process. We can define *validity* as the extent to which differences in scores on a measure reflect *only* differences in the distribution of values on the variable we intend to measure. Since we can probably never achieve complete and total validity, our goal should be to select measures that are susceptible to as few influences as possible other than differences in our target variable. This requires that we think carefully through the processes that surround our measures in search of possible causes of variations in scores. We are essentially concerned at this point with guarding against the effects of systematic error.

Consider this example. We may want a measure of the extent to which the citizens of different nations agree with the policies of their government. We decide to rely on answers to a series of survey questions as our indicator of agreement or disagreement. We hope that differences in citizens' actual opinions are the only source of differences in their responses to these questions. A moment's reflection, however, alerts us to another possible source of variation. If some of the nations included in our study have authoritarian governments that use secret police to repress dissent and that regard any criticism of their policies as acts of treason, their citizens may well be afraid to express disagreement with their government in an interview. In this case, scores on our measure may be determined at least as much by the attitude of each nation's government toward dissent as by the opinions of those being interviewed. The strong possibility of this type of measurement error makes the survey questions an inappropriate operationalization.

Similarly, we must be concerned with completeness early in the research process. If we want to measure the relative influence of different interest groups in

the state legislature, we may think of using newspaper reports of interest group appearances before legislative committees as our indicator. We must ask ourselves, however, whether giving testimony in public hearings is all there is to political influence. This activity is legitimately considered a *part* of the influence process, but there are so many other means of exercising influence that a measure that relies exclusively on the giving of testimony as an indicator of influence is incomplete.

Achieving appropriate and relatively complete operationalizations then depends both on knowing a good deal about the subject of our study and on conducting a careful, logical analysis of alternative operationalizations. We can check the validity of our measures in order to determine whether or not we have developed sound measures only *after* we have collected data, however. The process of evaluating the validity of our measures is referred to as **validation.**

There are four basic approaches to validation. The first is often called **pragmatic validation,** because it involves assessing the validity of a measure from evidence of how well it works in allowing us to predict behaviors and events. For example, say that we devise a measure of how appealing candidates for public office are to voters. We can get some indication of the validity of this measure by applying it to all the candidates for seats in the U.S. Senate in a given election year and predicting their chances of being elected on the basis of their relative scores on our voter appeal measure. The more successful we are at predicting the candidates' electoral fate, the more confident we become that we have a valid measure, one that accurately reflects the intended concept. Measures that allow us to predict future events accurately are said to have **predictive validity.**

Pragmatic validation requires that there be some alternative indicator of variables that we feel fairly certain is a valid reflection of them. We check our measures against this alternative as we might check the accuracy of verbal reports of age against birth certificates. Unfortunately there are seldom any clearly valid alternative indicators for the concepts used in social science research. As a result, we generally have to rely on the second type of validation—**construct validation.**

Construct validation is achieved by *inferring* the validity of a measure from evidence of the extent to which actual relationships between scores of various measures are consistent with what we expect from the theory that has led us to use a given indicator. This involves two lines of reasoning.

First we might say to ourselves, "If concept X has a positive relationship to concept Y and a negative relationship to concept Z (as our theory says it does), then it will also be true that scores on a valid measure of X will have a positive relationship to scores on a valid measure of Y and a negative relationship to scores on a valid measure of Z." We cannot validate the measure by comparing scores on it to scores on some other measure of the same variable that we know to be valid (as in the case of the birth certificate). We can, however, judge its validity by the extent to which using it as an indicator of our variable produces the kinds of relationships that our theory leads us to expect between that variable and other variables.

As an example, consider a study of international alliances. We might create a measure of the strength of an alliance based on a content analysis of newspaper articles from the countries involved. Is what the newspapers of one nation say about another nation a valid indicator of the strength of the alliance between the two countries? We might get an idea of whether it is by reasoning as follows: "Our theory tells

us that the stronger an alliance between two nations is, the more often they will vote together in the United Nations and the fewer restrictions they will place on trade with each other. Therefore, scores on a valid measure of *strength of alliance* will be positively related to scores on measures of *voting together in the United Nations* and negatively related to scores on measures of *number of trade barriers.*" We then proceed to do the data analysis necessary to see whether this expectation is supported by our observations. If the relationships are as expected, we will have greater confidence in the validity of our measure of *strength of alliance.* If they are not as we have expected, we will question whether we have a sound measure of this concept.

What we have just described is often referred to as **external validation.** It involves comparing scores on the measure being validated with scores on measures of *other* variables. To use this method of validation, of course, we have to include measures of the other variables in our research. This means that *we have to begin thinking about ways to validate our measures early in the research process.* Certainly by the time we are ready to develop a research design, we have to know how we will want to check the validity of our measures so that we can be certain to gather any other information we will need.

Our efforts at external validation will produce convincing evidence about the validity of our measure of one variable only if we can have a high degree of confidence in the validity of the measures we use for the other variables. In the previous example, for instance, we could not conclude anything about the validity of our measure of *strength of alliance* from the relationships between scores on it and scores on the other two variables if we did not believe that our indicators of *voting together* and *trade barriers* were valid. Because it is often difficult to find clearly valid indicators of variables to which our key variable should be related, external validation procedures must be used with caution. This is very much like testing a hypothesis. No single result guarantees the validity (or invalidity) of the measure. Rather, as instances of successful validation attempts accumulate, our confidence in the validity of our measure grows. For that reason, it is wise to seek out as many theoretically predictable relationships as possible to use in external validation. The more different tests of validity we have, the stronger our case will be.

This same logic applies to the second type of construct validation—**internal** or **convergent validation.** This type of validation involves devising several measures of the *same* variable and comparing scores on these various measures. We reason that if each of the indicators provides a valid measure of the concept in question, the scores individual cases receive on the measures should be closely related. If A, B, and C are all valid measures of X, then any individual's scores on A, B, and C should be highly similar.

For instance, suppose that we want an indicator of the quality of street lighting in residential neighborhoods as part of a study of the distribution of public services. We might want to use citizens' perceptions of the adequacy of street lighting (as revealed in survey interviews) as that indicator. We can ask a sample of citizens in a neighborhood how adequate they think the streetlights in their area are and take the average evaluation as our measure of *quality of street lighting.* In order to perform an internal validation, we may also measure street lighting quality (1) by using a light meter to get a physical measure of the brightness and distribution of lighting, (2) by having trained observers rate the lighting, and (3) by having citizens compare their street lighting with that pictured in a series of photos showing streets with different qualities of

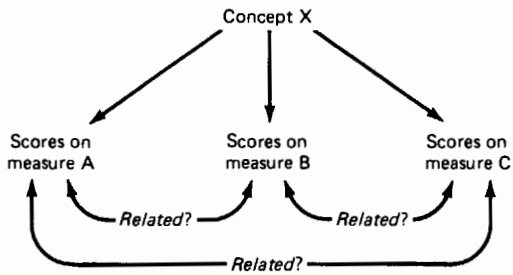
lighting and then averaging their rankings to get a measure for the neighborhood. This gives us four measures of the variable. If each is valid, all should be strongly related. We can check this with appropriate statistics. If we find that scores on the measure based on responses to interview questions are weakly related to scores on the other three measures *and* that scores on those other measures are strongly related to one another, we will have reason to suspect that our first measure is not valid.

This is much like weighing the same object on three different scales. If each of the scales gives an accurate weight and we have no reason to assume that the object has changed weight in the course of the test, we expect the weights obtained from the three scales to be identical. If one gives a different weight, we suspect it of being out of adjustment.

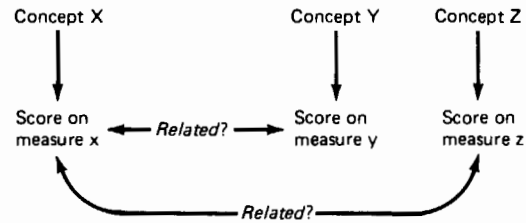
Figure 4.5 suggests the differences between internal and external forms of construct validation. In Figure 4.5*a*, we see that internal validation is achieved by checking the correspondence of scores on several different measures of the *same* concept. The more closely they correspond, the more justified we feel in claiming that any one of the measures is valid. In Figure 4.5*b*, we see that external validation involves determining whether our measure of one variable shows it to be related to *other* variables as we expect it to be from our theory. If the expected relationships do not appear, we have reason to suspect that the indicator we have selected does not provide valid measures of the concept. (In Chapter 17, we discuss the statistics that can be used to determine how strongly various measures are actually related.)

The same caution that applies to the use of external validation procedures applies to the use of internal validation. We cannot always be certain that our alternative measures of the key concept are valid, and we should therefore always be careful about concluding that a measure is valid or invalid from any one test of validity. We can significantly increase our confidence in the results of an internal validation if we follow a simple rule: *The alternative measures of the concept should be based on as many different types of operationalization as possible.*

In the street lighting example, our measures come from four distinct types of operationalization: citizens' verbal ratings, physical measurements, observers' judgments, and citizens' selection of photographs. Each of these represents a different *mode* of operationalization. The more different modes we can use and the more independent they are of each other, the more confidence we can place in our validation. Why? The logic is as follows. The principal source of invalidity is systematic and random measurement



(a) Internal (convergent) validation



(b) External validation

FIGURE 4.5 Forms of construct validation

error. Different measures are subject to different kinds of measurement error. The more indicators we have for any variable and the more they differ from one another, the less likely it is that all the indicators will be affected by the same measurement error. If this is true, we will have a better chance of both recognizing measurement error as a source of differences in the scores on any one of our measures and getting an accurate measure of our variable if we use **multiple indicators.**[7]

For instance, the factors that may make our physical measure of street lighting quality invalid (such as a faulty light meter) are likely to be quite unrelated to any factors that might introduce systematic errors into the measure based on citizens' evaluations (such as a tendency for people to claim out of a sense of community pride that public services in their neighborhood are as good as those in other areas). If we use only one mode of measurement, any source of measurement error may affect the scores on each measure, giving us a consistently invalid indicator and not allowing meaningful comparisons among measures. If, for example, we rely only on the physical measure of lighting but take readings in several different ways (say, on the sidewalk, on the curb, and in the street), then any flaw in the measuring instrument (the light meter, in this case) will affect all measures and none can be used to check another.

This logic suggests the great value of having multiple indicators for our variables. The availability of multiple measures not only gives us an opportunity to *test* the validity of our indicators but also *improves our chances of obtaining a valid measure* of our variables in the first place. Multiple measures can actually increase the validity of measurement by allowing us to combine the results of several different measurement procedures so as to produce a composite score that is more likely to be a valid reflection of the actual value of our variable than is any of the measures taken alone. Such a composite score is more likely to be a valid measure because there is a chance that the errors that cause each measure to be invalid will cancel out when the results of several measurement procedures are combined.

The principle here is much like weighing an object on many different scales, as described earlier. Because scales are not perfect, each one may show a slightly different weight—one a little too heavy, another a little too light. If we weigh the object on enough scales, however, the laws of probability tell us that there is a very good chance that these small errors will cancel each other out, giving us an *average* weight that is correct. Similarly, if we operationalize our concepts in several different ways so that the measurement error associated with each operationalization is independent of that involved in all other operationalizations, we stand a good chance of securing an accurate measure of our concept by combining the various scores. (In Chapter 9, the sections on scaling and indexing describe some possible methods of combining scores to produce a composite measure.)

A third approach to validation is referred to as **discriminant validation.** When we ask whether a measure exhibits *discriminant validity,* we are essentially asking whether using it as an indicator of a given concept allows us to distinguish that concept from other concepts. For example, we might want to measure the concept *trust in political officials* through a series of questions in a survey. If we also have on the

[7] Excellent discussions of the logic of multiple operationalizations are presented in David C. Leege and Wayne L. Francis, *Political Research* (New York: Basic Books, 1974), chap. 5, and John L. Sullivan and Stanley Feldman, *Multiple Indicators* (Beverly Hills, Calif.: Sage, 1979).

TABLE 4.1 Types of validation

Pragmatic Validation	Construct Validation	Discriminant Validation	Face Validation
Check results obtained from use of the indicator against results obtained from use of another indicator that is known to be a valid measure of the concept, or test the *predictive validity* of the indicator by using it to predict events that reflect the concept being measured.	*Internal (convergent) validation:* Infer validity of the indicator from its relationship to other indicators of the *same* concept using *multiple indicators.* *External validation:* Infer validity of the indicator from its relationship to indicators of *other* concepts to which the concept being measured should *theoretically* be related.	Infer validity of the indicator from the degree to which it is *un*related to indicators of other concepts that are theoretically distinct from the concept being measured.	Assume validity from the self-evident character of the indicator. (Can knowledgeable persons be persuaded that this is a valid indicator of the concept?)

questionnaire a series of questions designed to measure *trust in people (in general),* by comparing the scores on the two measures we can ask whether our first set of questions actually reflects simply another way of measuring trust in people. If the scores are highly similar, we say that the political trust measure does not have discriminant validity because it does not permit us to distinguish the concept of trust in political officials from the concept of trust in people.

A final approach to validation relies on the concept of **face validity.** Some measures are based on such direct observation of the behavior in question that there seems to be no reason to question their validity; such a measure seems valid "on the face of it." For example, suppose we want to measure compliance with a state law requiring each business establishment to display its operating license on its front door. Having trained observers simply note the presence or absence of such licenses seems to provide an obviously valid measure of compliance. Though we should always ask ourselves if the measures we have selected appear valid on their face, it is generally a mistake to rely on face validity alone to ensure accurate results from our research. We should attempt to ascertain the validity of our measures through established procedures, such as those already described.

These four types of validation are summarized in Table 4.1. If we are to draw accurate conclusions from our research, we must have valid measures. But to be valid, our measures must also be reliable.

RELIABILITY

When we ask about the validity of a measure, we are asking how closely the values it yields correspond to the true values of the variable being measured. When we ask about the **reliability** of a measure, we are asking how stable the values it yields are. Can we get the same value for any given case when we apply the measure several dif-

ferent times, or does each application result in the assignment of a different value to each case? If we do not get substantially the same value for any given case from successive applications of a measure, that measure is *unreliable* as an indicator of the concept. Rulers are made of inelastic materials in order to ensure reliability. If they were made of elastic materials, they might very well show different lengths for the same object—even when the object's true length has not changed—simply because the ruler stretches and contracts.

If a measure is unreliable, it cannot be valid, because at least some of the differences in the scores assigned to cases result from measurement errors rather than from true differences between cases. Recall our example of the study of street lighting. What if the light meter we use is so sensitive that in addition to recording the light from the streetlights, it picks up light from the moon? Then the values assigned to each street on the variable *quality of street lighting* will depend both on the brightness of the street lights *and* on such random factors as the fullness of the moon and the density of the cloud cover. To the extent that these random factors influence our results, the measure will not be a valid reflection of actual differences in the quality of street lighting. In this case, unreliability produces invalidity.

A measure may be quite reliable and yet invalid. Recall our example of the study of the extent to which people in different nations agree with the policies of their government. We said that survey questions may give invalid measures because people in authoritarian countries are afraid to tell the truth about their opinions. Because this factor produces a systematic rather than a random error, the questions might produce very stable results. No matter how many times they are asked, people might give the same "safe" responses. This does not, however, make the measure valid.

A measure may then *be reliable without being valid, but it cannot be valid without being reliable.* Whereas validity is challenged by both systematic and random error, reliability is jeopardized only by random error. This means that if a measure has been convincingly validated in prior studies we can use it without being worried about its reliability; it has to be reliable if it is valid. But demonstrating reliability does not guarantee validity.

How do we guard against unreliability? How do we determine whether or not a given measure is reliable? Preventing unreliability depends on our being aware of the various sources of random measurement error described earlier in this chapter and doing what we can to control them. This involves thinking through the actual measurement process and pretesting our measuring instruments to discover previously unrecognized causes of random error.

It is often quite difficult to determine whether or not we have devised a reliable measure in the social sciences. This is because the true value of the variables with which we are concerned can change dramatically with time and circumstance—people change their opinions in response to experience, nations alter the way they allocate resources between social services and defense efforts in response to perceived military threats, and so on. When real values are changing, it is hard to distinguish the effects of random measurement error from genuine fluctuations in the concepts being measured. This means that tests of reliability should be conducted over as short a time span as possible.

There are essentially three broad methods of assessing the reliability of measures in the social sciences. The first is the *test-retest method.* Here the same measure is applied to the same set of cases again and again, over time. To the extent that cases get

the same score each time, the measure is considered reliable. A difficulty with this technique arises when our measure involves interviewing people (as opposed to measuring inanimate objects or making concealed observations of people). If we repeat questions in a short time, interviewees may remember their first answer and, in an effort to be consistent, repeat that answer rather than respond truthfully in answering the question. If this happens, we cannot get an accurate picture of the questions' reliability as an indicator of the concept. In an effort to avoid this test effect, we might let a good deal of time pass before asking the questions a second time. If we do that, however, we will run into another problem: true values on the variable may have changed with the passage of time, and we may be unable to distinguish differences in scores caused by unreliability in the measure from actual changes in the variable.

Because of that difficulty, a second type of reliability test has been developed: the *alternative form method.* Different forms of the measure are applied to the same group of cases, or the same measure is applied to different groups *at the same time.* In this way there can be no test effect, because no case will be measured more than once, and, because no time lapses between applications of the measure, actual changes in the variables under study cannot affect the results. The success of this strategy, however, depends on the alternative forms of the measure being perfectly comparable to each other as a measure of the concept, or on the two groups being virtually equivalent with respect to the distribution of the variable being measured. If we can assume that these conditions are met, the more the scores on the two measures, or the scores of the two groups, are alike, the more confidence we have in the reliability of our measure. If we cannot come up with comparable measures or groups, however, we cannot use the method properly.

The final basic approach to testing the reliability of a measure is known as the *subsample method.* In it we draw one sample of cases and divide it into several subsamples in such a way that each is highly similar to the others in composition. We then apply the same measure to all subsamples and use the similarity or difference of responses from subsample to subsample as an indicator of the reliability of the measure. Because we use the same measure, we do not have to be concerned about comparability as in the alternative form method, and because we can rely on sampling theory to ensure the equivalence of our subsamples, we do not have to worry that the groups selected for measurement will not be sufficiently alike. Because no case is measured twice, we can discount the test effect as a threat to the accuracy of our reliability test, and because the measures are administered simultaneously, actual changes in the variable cannot create problems for this method, as they can for the test-retest method. However, use of the subsample method depends on our being able to draw a large enough sample that we can divide it and still have subsamples large enough for our statistical tests to be meaningful. This is not always possible and can represent a barrier to the use of the subsample method in testing reliability.

A variety of statistical procedures are available to use in interpreting the results of each of these tests of reliability.[8]

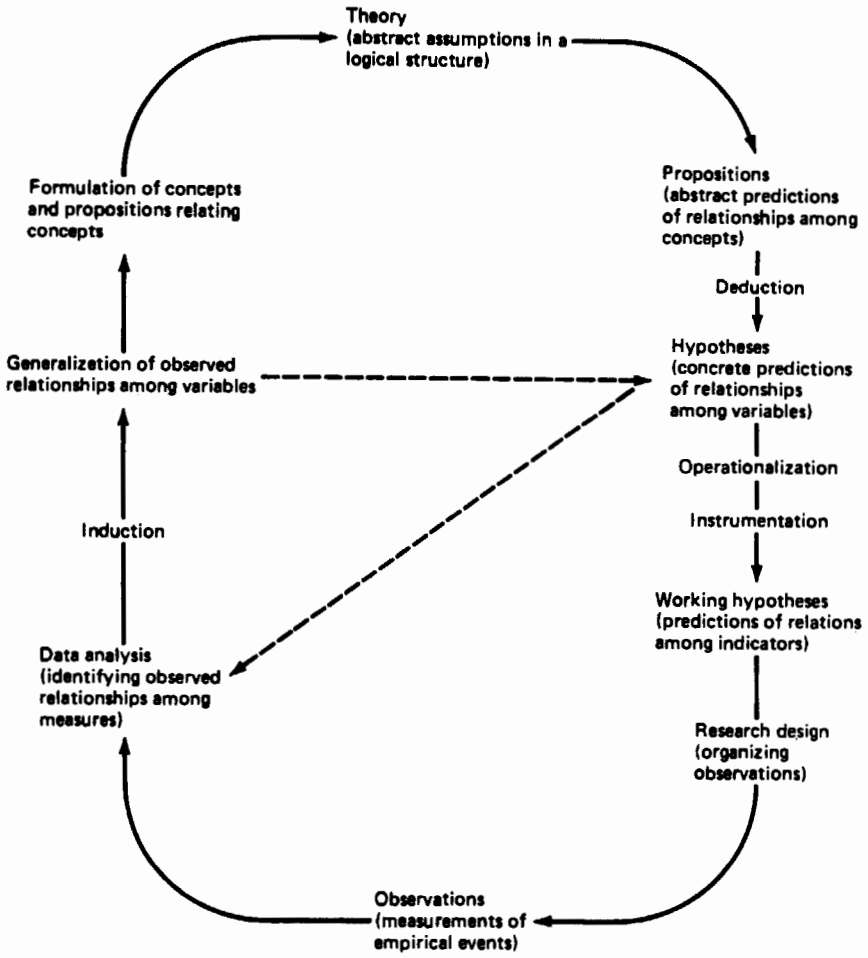
---

[8] Some of these are described in George W. Bohrnstedt, "Reliability and Validity Assessments in Attitude Measurement." In Gene F. Summers, ed., *Attitude Measurement* (Skokie, Ill.: Rand McNally, 1970), pp. 80–99.

There are many variations on these methods. Which one is most appropriate for any given research project will depend both on the time and resources available to complete the research and on the nature of the study. For instance, if we want to measure street lighting by having trained observers rate the lighting on various blocks, we can easily use the test-retest method without concern about a test effect. Street lighting will not change simply because it is measured by someone, and so we can have different observers independently rate the same street on the same night. We cannot have the same confidence in this method if our measure of street lighting quality is based on citizens' responses to interview questions.

Regardless of the reliability test we choose to use, it is important to establish the reliability of our measures *before* actually beginning research. This involves *pretesting* the measure by collecting some data exclusively for the purpose of assessing the instruments we will use in the final study. If we fail to do this, we may find only *after*

FIGURE 4.6 A model of the research process



the study is complete that our measures of key variables are unreliable (and therefore invalid). This means that we will not be able to place any faith in the results of the research and that our energies will have been partially or totally wasted. *Pretests of measures' validity and reliability should be part of any research project that either uses measures that have not been convincingly validated elsewhere or relies on measures that have been validated only in settings very different from those in which they will be used.*

CONCLUSION

At this point we have introduced all the basic elements of the research process. Figure 4.6 pictures their relationships to one another. The operationalization of our concepts through the development of measurable indicators prepares us to enter the field to make the observations on which we will base our conclusions. Before we can make those observations, however, we need a "plan of attack"—a scheme for making the observations in a way that will maximize the number of conclusions we can confidently draw from them. This plan, or *research design,* is the subject of Chapter 5.

SUGGESTIONS FOR FURTHER READING

Most explanations of measurement in the social sciences are found in literature that reports research results or develops sophisticated measurement techniques. General introductions to the subject are rare. We can, however, suggest some useful source material in addition to the works cited in this chapter's notes. Useful discussion of the logical foundations of measurement are offered in Fred N. Kerlinger, *Foundations of Behavioral Research* (New York: Holt, Rinehart and Winston, 1964), and in Abraham Kaplan, *The Conduct of Inquiry* (San Francisco: Chandler Publishing Co., 1964). Some of the practical problems encountered in measurement are illuminated by John M. Johnson, *Doing Field Research* (New York: Free Press, 1975), and one of the best introductions to measurement strategies is W. Philips Shively, *The Craft of Political Research,* 2d ed. (Englewood Cliffs, N.J.: Prentice-Hall, 1980). Some more advanced approaches to measurement can be found in Hubert M. Blalock, Jr., ed., *Measurement in the Social Sciences* (Chicago: Aldine, 1974), and Hubert M. Blalock, Jr., *Conceptualization and Measurement in the Social Sciences* (Beverly Hills, Calif.: Sage, 1982). A variety of examples of measurement strategies are found in George W. Bohrnstedt and Edgar F. Borgatta, eds., *Social Measurement* (Beverly Hills, Calif.: Sage, 1981). A brief overview of methods for testing validity and reliability is Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment* (Beverly Hills, Calif.: Sage, 1979).

RESEARCH EXERCISES

1. Using one of the political science periodicals given in Chapter 3, locate an article that reports the results of an empirical investigation. Identify at least two principal concepts from the article, and write down how each concept was operationalized. If the article uses only one operationalization for each concept, describe at least two more for each concept. If more than one operationalization is used for each concept in the article, describe at least

one alternative operationalization for each. If possible, your operationalizations should rely on indicators that are in a different form from those used in the article.

2. Select another scholarly article reporting the results of empirical research. Identify at least two of the major concepts employed, and describe how they were operationalized. Then state a line of reasoning that will lead us to expect the indicator selected for each variable to change when values of the variable change. In other words, state a *measurement theory* justifying the use of that indicator.
3. Using the variables from the article selected for Exercise 2, devise at least two alternative measurement theories showing how changes in each indicator employed in the article can result from changes in some variable *other* than the one it is used to represent. In other words, identify at least two possible sources of invalidity for each indicator.
4. Using one of the articles selected for any of the above exercises, identify two key measures employed in it. Tell which of the three broad methods for testing the reliability of measures is appropriate to use in checking the reliability of each of these measures, and justify your choice. Describe how you would set up the observations necessary to implement the test of each measure. For example, you might write, "I would include at the beginning and end of each of the interviews a version of the question on which the measure is based and I would compare the answers each respondent gives to the two questions."

TERMS INTRODUCED IN THIS CHAPTER

operationalization
instrumentation
measurement
indicator
values
observation
multidimensional
operational definition
instrument
levels of measurement
nominal measurement
mutually exclusive
collectively exhaustive
ordinal measurement
interval measurement
working hypotheses
measurement theory
measurement error
systematic errors
random errors
validity
validation
pragmatic validation
predictive validity
construct validation
external validation
internal (convergent) validation
multiple indicators
discriminant validation
face validity
reliability