CHAPTER 4

The Building Blocks of Social Scientific Research:

Measurement

In the previous chapter we discussed the beginning stages of political science research projects: the choice of research topics, the formulation of scientific explanations, the development of testable hypotheses, and the definition of concepts. In this chapter we show how to test empirically the hypotheses we have advanced. This entails understanding a number of issues involving the **measurement**, or systematic observation and representation by scores or numerals, of the variables we have decided to investigate.

In Chapter 1 we said that scientific knowledge is based upon empirical observation. In this chapter we confront the implications of this fact. If we are to test empirically the accuracy and utility of a scientific explanation for a political phenomenon, we will have to observe and measure the presence of the concepts we are using to understand that phenomenon. Furthermore, if this test is to be an adequate one, our measurements of the political phenomena must be as accurate and precise as possible. The process of measurement is important because it provides the bridge between our proposed explanations and the empirical world they are supposed to explain.

The researchers discussed in Chapter 1 measured a variety of political phenomena. Steven Poe and C. Neal Tate measured population and economic growth as well as a number of other factors to see if they had an impact on the incidence of state terrorism.[1] Bruce Bueno de Mesquita, Randolph Siverson, and Gary Woller measured the outcomes and costs of international wars to study their post-war effects.[2]

Benjamin Page and Robert Shapiro measured the change in public opinion on more than three hundred political issues and the subsequent change, if any, in public policy on those same issues.[3] Jeffrey Segal and Albert Cover measured both the political ideologies and the written opinions of Supreme Court justices in cases involving civil rights and liberties.[4] And B. Dan Wood and Richard Waterman measured the decisions of several bureaucratic agencies to determine if they were influenced by presidential and congressional intervention.[5] Richard Fording measured growth in welfare rolls to see if

it was related to the amount of civil unrest.[6] Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino measured the intention to vote reported by participants in their experiments to see if it was affected by exposure to negative campaign advertising.[7] In each case some political behavior or attribute was measured so that a scientific explanation could be tested.

Devising Measurement Strategies

As we pointed out in Chapter 3, researchers must define the concepts they use in their hypotheses. Researchers also must decide how they are actually going to measure the presence, absence, or amount of these concepts in the real world. Political scientists refer to this process as providing an **operational definition** of their concepts—in other words, deciding what kinds of empirical observations should be made to measure the occurrence of an attribute or behavior.

Let us return, for example, to the researcher trying to explain the existence of democracy in different nations. If the researcher were to hypothesize that higher rates of literacy make democracy more likely, then a definition of two concepts—literacy and democracy—would be necessary. The researcher could then develop a strategy, based on these two definitions, for measuring the existence and amount of both attributes in a number of nations.

Suppose literacy was defined as "the completion of six years of formal education," and democracy was defined as "a system of government in which public officials are selected in competitive elections." These definitions would then be used to develop operational definitions of the two concepts. These operational definitions would indicate what should be observed empirically to measure both literacy and democracy, and they would indicate specifically what data should be collected to test the researcher's hypothesis. In this example, the operational definition of literacy might be "those nations in which at least 50 percent of the population has had six years of formal education, as indicated in a publication of the United Nations," while the operational definition of democracy might be "those countries in which the second-place finisher in elections for the chief executive office has received at least 25 percent of the vote at least once in the past eight years."

When a researcher specifies the operational definition of a concept, the precise meaning of that concept in a particular research study becomes clear. In the preceding example, we now know exactly what the researcher means by literacy and democracy. Since different people often mean different things by the same concept, operational definitions are especially important. Someone might argue that defining literacy in terms of formal education ignores the possibility that people who complete six years of formal education might still be unable to read or write well. Similarly, it might be argued that defining

democracy in terms of competitive elections ignores other important features of democracy such as freedom of expression and citizen involvement in government actions. In addition, the operational definition of competitive elections is clearly debatable. Is the "competitiveness" of elections based on the number of competing candidates, the size of the margin of victory, or the number of consecutive victories by a single party in a series of elections? Unfortunately, operational definitions are seldom absolutely correct or absolutely incorrect; rather, they are evaluated in terms of how well they correspond to the concepts they are meant to measure.

It is useful to think of the operational definition as the last stage in the process of defining a concept precisely. We often begin with an abstract concept (such as democracy), then attempt to define it in a meaningful way, and finally decide in specific terms how we are going to measure it. At the end of this process we hope to attain a definition that is sensible, close to our meaning of the concept, and exact in what it tells us about how to go about measuring the concept.

Let us consider another example: the researcher interested in why some individuals are more liberal than others. The concept of liberalism might be defined as "believing that government ought to pursue policies that provide benefits for the less well off." The task then is to develop an operational definition that can be used to measure whether particular individuals are liberal or not. The researcher might decide that anyone is a liberal who agrees in a public opinion poll with the following statement: "The federal government should increase the amount of money spent on food stamp and free lunch programs."

An abstract concept, liberalism has now been given an operational definition that can be used to measure the concept for individuals. This definition is also related to the original definition of the concept, and it indicates precisely what observations need to be made. It is not, however, the only operational definition possible. Others might suggest that questions regarding affirmative action, school vouchers, the death penalty, welfare benefits, and pornography could be used to measure liberalism. The important thing is to think carefully about the operational definition you choose and to try to ensure that that definition coincides closely with the meaning of the original concept.

Examples of Political Measurement

Let us take a closer look at some operational definitions used by political science researchers. Page and Shapiro's article on the relationship between public opinion and public policy contains a number of important decisions about the measurement of abstract political concepts.[8] Ever since democratic theorists first articulated the normative position that in a democracy there should be a relationship between popular preferences and the decisions made by political elites, researchers have been curious about whether such a rela-

tionship exists in modern nation-states and, if so, how strong it is. Observing the relationship requires devising proper measures of both popular preferences and public policy decisions, a requirement that effectively stymied empirical research on opinion-policy relationships until recently because of the difficulty in finding appropriate measures of popular preferences and of corresponding public policy decisions over an extensive enough period of time.

Page and Shapiro approached this problem by relying on the thousands of public opinion survey questions that have been asked of the American adult population since 1935 for their measures of popular preferences. Since they were interested in the effect of *change* in the distribution of public opinion, they decided to select those survey questions that had been asked in identical form at two or more points in time, and on which there had been at least a 6-percentage-point change in public opinion from one survey to another. This left them with 357 instances of public opinion change on public policy questions between 1935 and 1979.

To test the relationship between opinion and policy change, Page and Shapiro then had to devise measures of public policy that corresponded to the measures of public opinion. And to determine whether opinion change *precedes* policy change (recall our discussion in Chapter 3 on the importance of establishing the temporal sequence when evaluating causation), the policy measures needed to cover a period of time at least as broad as the measures of public opinion covered. So Page and Shapiro sought public policy measures that began two years before the first measure of public opinion and continued until four years after the final survey (to allow for a time lag in the policy-making process).

Developing measures of public policy that correspond with public opinion surveys is not an easy task. Suppose, for example, we decide to conduct a survey that asks whether the populace favors increasing, decreasing, or maintaining the government's defense spending. What would be an appropriate corresponding measure of public policy? Would we use the entire budget of the Defense Department? Should we include military personnel salaries, benefits, and pensions, or just expenditures on weapons systems? Should we use authorized budget figures (as a measure of willingness to spend), or the amount actually expended? Should we include expenditures on NASA, the CIA, and military foreign aid? Should we correct the measure of expenditures for inflation? Should we express expenditures as a percentage of the total budget, or as a percentage of the Gross National Product? These and other questions must be answered before we can settle on an appropriate measure of just this one public policy. Clearly, the measurement decisions required by the hypothesis relating public opinion and public policy are both numerous and difficult.

Fording's investigation of the impact of mass insurgency on welfare generosity required that he measure both welfare expansion and the extent of ri-

oting. As a measure of welfare expansion, Fording used the annual growth in state AFDC (Aid to Families with Dependent Children) recipient rates. The AFDC program was used because data for the program were reported for each state on a regular basis. Data for another welfare program among the states, General Assistance, in contrast, were not regularly reported. Fording encountered a more complex measurement problem for the concept of mass insurgency. Previous researchers had used the number of riots as the indicator of mass insurgency, with a riot being defined as a spontaneous incident of collective violence, usually involving no less than twenty to thirty people. Fording chose to operationalize insurgency as "any act of violence on behalf of blacks or minorities, either spontaneous or planned, that is either framed as, or can be construed as politically motivated. Violent acts include rock throwing, vandalism, arson, looting, sniping, or beating of whites."[9] He used the number of incidents rather than the severity of violent incidents since previous studies had shown that measures of the severity or intensity of violence were highly correlated with the number of violent incidents. Data for violence were obtained from a variety of sources including the *New York Times*, the Kerner Report, publications from the Lemberg Center for the Study of Violence, Facts on File, and Congressional Quarterly. Fording collected data for a total of 923 events.

Because Fording also wanted to investigate whether the effect of insurgency on welfare generosity depended on the extent to which insurgent groups have effective access to electoral institutions, he needed to define and measure electoral access. Arguing that effective electoral power depends on both the acquisition of voting rights and the use of population-based districts for the election of state legislatures, he determined for each state for each year in his study whether or not the state legislature was apportioned and whether or not African Americans had access to the ballot. His measurement of ballot access was based on the passage of the Voting Rights Act of 1965 and its implementation in targeted states.

Furthermore, Fording sought to control for need for welfare, as it was possible that expanding welfare rolls merely reflected an increase in need. He measured increase in need by looking at the growth in the percentage of households in a state that were headed by females and below poverty. Fording used this measure rather than the growth in poverty among all households because AFDC was a program for families with dependent children. He also used the growth in the unemployment rate as an indicator of need. Thus, Fording developed clear definitions for the concepts in his research as he explained how he actually planned to measure each concept.

Daron Shaw's research concerned the impact of TV ads and campaign appearances on statewide presidential voting.[10] For this he needed schedules of candidate appearances, candidates' television advertising purchases by media

markets, and trial ballot polling data from each state. Data to measure all three variables had previously been difficult to obtain. Shaw relied on three sources of information about campaign appearances: the *Hotline* (a daily political newsletter published in Falls Church, Virginia), the *Washington Post*, and schedules kept by the campaigns. In measuring candidate appearances, Shaw did not count visits by candidates to their hometowns, places of work, vacation destinations or debate sites unless campaign appearances were reported, nor did he weigh high-visibility appearances more than low-visibility appearances (although he did check to see if this distinction affected his results, which it did not). Data on television advertising came from the campaigns themselves and encompassed all TV advertising purchased by the campaigns between September 1 and election day. Shaw's data on television advertising represented the first time that such data were available for media markets and the states. Shaw relied on polling data from a number of public and private sources including the Republican presidential campaigns, news media consortiums, and the Republican National Committee.

Shaw confronted a number of measurement issues concerning television advertising. One problem was that his data omitted political party and independent spending, which accounted for a substantial portion of media spending, because estimates of this spending were incomplete. A second problem was related to the variability in the cost of advertising across media markets. Simply using the amount of money spent on advertising would not measure the amount of media exposure obtained. Shaw borrowed a measurement scheme called gross rating points (GRPs) used by advertising companies and communication scholars. GRPs provide a measure of the audience reached in media markets independent of markets costs. Statewide GRPs were calculated by multiplying the number of GRPs bought in a market by the percentage of the state's eligible voters in that market, repeating the procedure for all markets in a state, and then summing the results. The third issue Shaw faced was that advertising time is not always purchased in the state for which the ad is targeted. For example, ads aimed at audiences in New Jersey may be purchased in New York or Philadelphia and thus the audience exposure being purchased is not exclusively in New Jersey.

The research conducted by Segal and Cover on the behavior of U.S. Supreme Court justices is a good example of an attempt to overcome a serious measurement problem to test a scientific hypothesis.[11] Recall that Segal and Cover were interested, as many others have been before them, in the extent to which the votes cast by Supreme Court justices are dependent upon their own personal political attitudes. Measuring the justices' votes on the cases decided by the Supreme Court is no problem; the votes are public information. But measuring the personal political attitudes of judges, *independent of their votes* (remember the discussion in Chapter 3 on avoiding tau-

tologies), is a problem. Many of the judges whose behavior is of interest have died, and it is difficult to get living Supreme Court justices to reveal their political attitudes through personal interviews or questionnaires. Furthermore, ideally one would like a measure of attitudes that is comparable across many judges and that measures attitudes related to the cases decided by the Court.

Segal and Cover decided to limit their inquiry to votes on civil liberties cases between 1953 and 1987, so they needed a measure of related political attitudes for the judges serving on the Supreme Court over that same period of time. They decided to infer each judge's attitudes from the newspaper editorials written about them in four major daily newspapers from the time each justice was appointed by the president until the justice's confirmation vote by the Senate. Trained analysts read the editorials and coded each paragraph for whether it asserted that a justice designate was liberal, moderate, or conservative (or if the paragraph was inapplicable) regarding "support for the rights of defendants in criminal cases, women and racial minorities in equality cases, and the individual against the government in privacy and First Amendment cases."[12] They selected the editorials appearing in two liberal papers and in two conservative papers to produce a more accurate measure of judicial attitudes.

Because of practical barriers to ideal measurement, then, Segal and Cover had to rely on a measure of judicial attitudes *as perceived by four newspapers* rather than on a measure of the attitudes themselves. While this approach *may* have resulted in flawed measures, it also permitted the test of an interesting hypothesis about the behavior of Supreme Court justices that had not been tested previously. If the measures that resulted were both accurate and precise then this measurement strategy would permit the empirical verification of an important hypothesis. Without such measurements, the hypothesis would have to have gone untested.

Next, let us consider the research on regime stability described briefly in Chapter 1. Recall that a common hypothesis is that the distribution of income and wealth has an impact on the tendency of populations to protest and rebel and, therefore, on the stability of governmental regimes. More specifically, it is often hypothesized that income/wealth inequality leads to political violence and instability. In many parts of the world the main form of wealth is land ownership. Consequently, researchers have been investigating the relationship between land inequality and civil unrest in places such as Latin America. Obviously the task requires an accurate measure of land inequality in a variety of nation-states.

Measuring the distribution of land ownership and especially the inequality of land ownership is not a simple matter, however. Everyone agrees that equality of land ownership would mean that every adult or family owned the

same amount of land, so that 25 percent of the population owned 25 percent of the land, 75 percent of the population owned 75 percent of the land, and so on. But when the pattern of land ownership departs from this strict equality it is not altogether clear what should be measured. If a few people own most of the land and most people own little or no land, there is land inequality. But should the measure of land inequality focus on unequal land holdings throughout the range of land ownership, or on the land owned by those at both ends of the ownership range (the very rich and the very poor) and the gap between them, or on the number of people who own little or no land, regardless of how land is distributed among those who own land? The answer is not clear.

Because of this uncertainty over how best to measure land inequality, researchers have used a number of different measures. One commonly used measure is the Gini index, which attempts to take into account inequalities throughout the range of land ownership. Manus Midlarsky proposed a different measure that concentrated on the gap between the very rich and the very poor.[13]

There is some evidence that the measure of inequality used affects the conclusions reached about the relationship between land inequality and political violence. Research using the Gini index has generally found a weak relationship between land inequality and political violence and has led researchers to think about the other factors affecting regime stability. Midlarsky's research using the new measure of land inequality, however, reports a much stronger relationship with political violence, at least in Latin America. This is a good example, then, of how measurement decisions may affect substantive conclusions. Our choice of measures, especially of abstract phenomena such as land inequality, is a crucial aspect of the entire research process.

Finally, let us take a look at an innovative attempt to measure what individuals mean by that troubling concept, "democracy."[14] Given the fact that many different theorists over several centuries have written about democracy in very different ways, it is hardly surprising that the concept means different things to different people. In an attempt to understand better what citizens mean by democracy, two researchers collected roughly three hundred statements about democracy from newspapers, magazines, dictionaries, ethnographic studies, and voters' pamphlets. From these they selected sixty-four they thought best represented the domain of the concept. They then asked a group of subjects to score the statements on a scale ranging from +6 (most agree) to –6 (most disagree).

Once the sixty-four statements were scored, the researchers then looked for patterns in their subjects' responses. They discovered four different response patterns that they believe represent different meanings of democracy and developed a "discourse" to capture each one of them. The four response

patterns are called contented republicanism, deferential conservatism, disaffected populism, and private liberalism, and are defined as follows:

> *Contented republicanism.* We live in a democracy, which is fortunate because democracy is without doubt the best form of government. Democracy is a way of life, not just a political system; it is bound up with our freedoms, and, though fallible, can correct its mistakes. . . . Political equality is important and easily achieved and does not require social and economic equality. . . . Politics need not be based on greed or self-interest, for democratic debate can help establish an identity between what's good for me and what's good for society. The importance of this debate means that there should be no restraints on the availability of information; a free press is crucial, and we should not tolerate lying in politics.
>
> *Deferential conservatism.* Politics is only for the few. Not everyone is capable of making good decisions, people don't know what they want, and not everyone can be represented. . . . It is undesirable for people to get any more involved in politics; citizen activism is not a good thing. Nor are the liberal values of a free press, independent judiciary, social justice rooted in basic rights, or the market very attractive. . . . There is no need to fear government; we should rely on elites to govern and hope that they are honest. Such elites will be able to look out for the long-term interest of society, which matters more than short-term economic concerns.
>
> *Disaffected populism.* We do not live in a democracy, as power is in the hands of conservative, corporate elites and a government that represses the people. . . . Over time, democratic control has deteriorated, and the future of democracy is bleak unless people wake up and do something about it. The freedoms that are central to democracy have been curtailed. . . . Ordinary individuals are well motivated; and they should attend to politics, rather than be preoccupied by work and family—though politics cannot, and should not, be all-consuming. . . . One should not necessarily condemn political violence when it is undertaken by the oppressed.
>
> *Private liberalism.* We do not live in an especially democratic society, and not everyone can be represented. But this is no cause for concern, for democracy is of no particular value, given that it can encompass both desirable kinds of government and undesirable forms, such as socialism. . . . Individuals should be free to pursue their own interests, but democracy will not guarantee that freedom. It is the private realm which really matters—work and family are the most important things in life, and one should rely on friends, neighbors, and the market, rather than government. . . . Government, however democratic, has intruded too far into this private realm; government should be small and subject to separation of powers and constitutional restraints.[15]

The discourses overlap some and exhibit imperfect similarities to historical/philosophical traditions, but they also reveal a good deal of variation and complexity in the meaning of democracy among the public.

The cases we have discussed here are good examples of researchers' attempts to measure important political phenomena (behaviors or attributes) in the real world. Whether the phenomenon in question was public policy decisions, judges' political attitudes, land inequality, campaign appearances and advertising by presidential candidates, mass insurgency, or democracy, the researchers devised measurement strategies that could detect and measure the presence and amount of the concept in question. These observations were then generally used as the basis for an empirical test of the researchers' hypotheses.

To be useful in providing scientific explanations for political behavior, measurements of political phenomena must correspond closely to the original meaning of a researcher's concepts. They must also provide the researcher with enough information to make valuable comparisons and contrasts. Hence the quality of measurements is judged in terms of both their *accuracy* and their *precision*.

The Accuracy of Measurements

Since we are going to use our measurements to test whether or not our explanations for political phenomena are valid, those measurements must be as accurate as possible. Inaccurate measurements may lead to erroneous conclusions since they will interfere with our ability to observe the actual relationship between two or more variables.

Suppose, for example, that you have hypothesized that "courses taught by political scientists are more worthwhile than courses taught by psychologists." During registration time you see a description for a course in the political science department that looks worthwhile, so you sign up for it. Your friend, on the other hand, signs up for a course in the psychology department. Two weeks after classes start you meet your friend in the campus dining hall and swap evaluations of your courses. Your political science course has turned out to be a dreadful experience, but your friend seems to be enjoying the psychology course.

Does your experience that semester (actually an empirical observation of two cases seemingly relevant to the hypothesis) mean that your hypothesis is wrong? Not necessarily. Suppose the course you had signed up for was not being taught by a political scientist but rather by a historian on loan to the political science department. This would mean that you had inaccurately measured the nature of the course you were taking by assuming that all courses taught in the political science department were taught by political scientists.

Thus you were being led to what might, in fact, have been an erroneous conclusion about the quality of courses taught by political scientists. Because of the inaccuracy of your measurement, the comparison you made with your friend turns out to be irrelevant to the hypothesis with which you began.

There are two major threats to the accuracy of measurements. Measures may be inaccurate because they are *unreliable* or because they are *invalid.*

Reliability

Reliability "concerns the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials. . . . The more consistent the results given by repeated measurements, the higher the reliability of the measuring procedure; conversely, the less consistent the results, the lower the reliability."[16]

Suppose, for example, you are given the responsibility of counting a stack of 1,000 paper ballots for some public office. The first time you count them, you obtain a particular result. But as you were counting the ballots you might have been interrupted, two or more ballots might have stuck together, some might have been blown onto the floor, or you might have written the totals down incorrectly. As a precaution, then, you count them five more times and get four other people to count them as well. The similarity of the results of all ten counts would be an indication of the reliability of the measure.

Or suppose you design a series of questions to measure how cynical people are and ask a group of people those questions. If a few days later you ask the same questions of the same group of people, the correspondence between the two measures would indicate the reliability of that particular measure of cynicism (assuming that the amount of cynicism has not changed). Similarly, suppose you wanted to test the hypothesis that the *New York Times* is more critical of the federal government than the *Wall Street Journal.* This would require you to measure the level of criticism found in articles in the two papers. You would need to develop criteria or instructions for identifying or measuring criticism. The reliability of your measuring scheme could be assessed by having two people read all the articles, independently rate the level of criticism in them according to your instructions, and then compare their results. Reliability would be demonstrated if both people reached similar conclusions regarding the content of the articles in question.

The reliability of political science measures can be calculated in a number of ways. The **test-retest method** involves applying the same "test" to the same observations after a period of time and then comparing the results of the different measurements. For example, if a series of questions measuring liberalism is asked of a group of respondents on two different days, a comparison of their scores at both times could be used as an indication of the reliability of the measure of liberalism. We frequently engage in test-retest behav-

ior in our everyday lives. How often have you stepped on the bathroom scales twice or more in a matter of seconds?

The test-retest method of measuring reliability may be both difficult and problematic since one must measure the phenomenon at two different points. It is possible that two different results may be obtained because what is being measured has actually changed, not because the measure is unreliable. For example, if your bathroom scales give you two different weights within a few seconds, the scales are unreliable as your weight can not have changed. However, if you weigh yourself once a week for a month and find that you get different results each time, are the scales unreliable or has your weight changed? A further problem with the test-retest check for reliability is that the administration of the first measure may affect the second measure's results. For instance, the difference between Scholastic Aptitude Test scores the first and second times that individuals take the test may not be assumed to be a measure of the reliability of the test since test takers might alter systematically their behavior the second time as a result of taking the test the first time.

The **alternative-form method** of measuring reliability also involves measuring the same attribute more than once, but uses two different measures of the same concept rather than the same measure. For example, a researcher could devise two different sets of questions to measure the concept of liberalism, ask the two sets of questions of the same respondents at two different times, and compare the respondents' scores. Using two different forms of the measure prevents the second scores from being influenced by the first measure, but it still requires being able to measure the phenomenon twice and, depending on the length of time between the two measurements, what is being measured may change.

Going back to our bathroom scale example, if you weigh yourself on your home scales, go the the gym, weigh yourself again and get the same number, you may conclude that the scales are reliable. But, what if you get two different numbers? Assuming your weight has not changed, what is the problem? If you go back home immediately and step back on your bathroom scales and find that they give you a measurement that is different from the first, you could conclude that your scales have a faulty mechanism, are inconsistent, and therefore, unreliable. However, what if your bathroom scales give you the same weight as the first time? They would appear to be reliable. Maybe the gym scales are unreliable. You could test this out by going back to the gym and reweighing yourself. If the gym scales give a different reading than the first time, then they are unreliable. But, what if the gym scales give consistent readings? Clearly one (or both) of the scales is inaccurate and you have a measurement problem that needs to be resolved and it involves more than unreliability. Before we address this situation, let us mention one more test for reliability.

The **split-halves method** of measuring reliability involves two measures of the same concept, with both measures applied at the same time. The results of the two measures are then compared. This method avoids the problem of the change in the concept being measured. The split-halves method is often used when there is a multi-item measure that can be split into two equivalent halves. For example, one may devise a measure of liberalism consisting of the responses to ten questions on a public opinion survey. Half of these questions could be selected to represent one measure of liberalism and the other half selected to represent a second measure of liberalism. If individual scores on the two measures of liberalism are similar, then the ten-item measure may be said to be reliable by the split-halves approach.

The test-retest, alternative-form, and split-halves methods provide a basis for calculating the similarity of results of two or more applications of the same or equivalent measure. The less consistent the results are, the less reliable the measure is. Reliability of the measures used by political scientists is a serious problem. Survey researchers are often concerned about the reliability of the answers they receive. For example, respondents' answers to survey questions often vary considerably when given at two different times.[17] If respondents are not concentrating or taking the survey seriously, the answers they provide may as well have been pulled out of a hat.

Now, let us return to the problem of a measure that yields consistent results as in the case of the two scales that differed consistently in how much you weighed. Each of the scales appears to be reliable (they are not giving you different weights at random), but one or both of them is giving you a wrong measurement. That is, it is not giving you your correct weight. This problem is the type of problem one confronts in trying to assess whether or not one's measures are valid.

Validity

Essentially, a valid measure is one that measures what it is supposed to measure. Unlike reliability, which depends on whether repeated applications of the same or equivalent measure yield the same result, **validity** involves the correspondence between the measure and the concept it is thought to measure.

Let us consider first some examples of invalid measures. Suppose a researcher hypothesizes that the larger a city's police force is, the less crime that city will have. This requires the measurement of crime rates in different cities. Now also assume that some police departments systematically overrepresent the number of crimes in their cities to persuade public officials that crime is a serious problem and that the local police need more resources. Some police departments in other cities may systematically underreport crime in order to make the city appear safe. If the researcher relied on offi-

cial, reported measures of crime, the measures would be invalid because they did not correspond closely to the actual amount of crime in some cities.

Or suppose you hypothesize that the more productive a scholar a faculty member is, the better a teacher he or she is. This requires measuring different faculty members' teaching abilities. The task might be accomplished by asking students to grade their instructors on teaching. But do students know a good teacher when they see one? Is it possible that students will give good grades to instructors who are personable, humorous, approachable, or easy graders, rather than to instructors who teach them something? If so, such a measurement strategy might well produce an invalid measure of good teaching.

Finally, there are many studies looking into the factors that affect voter turnout. These studies require an accurate measurement of voter turnout. One way of measuring voter turnout is to ask people if they voted in the last election. However, given the social desirability of voting in the United States, will all the people who did not vote in the previous election admit that they did not vote to an interviewer? More people might say that they voted than actually did, resulting in an invalid measure of voter turnout. In fact, this is what usually happens. Voter surveys commonly overestimate turnout by several percentage points.[18]

A measure's validity is more difficult to demonstrate empirically than its reliability because validity involves the relationship between the measurement of a concept and the actual presence or amount of the concept itself. Information regarding the correspondence is seldom abundant. Nonetheless, there are a number of ways of evaluating the validity of any particular measure.

Face validity may be asserted (not empirically demonstrated) when the measurement instrument appears to measure the concept it is supposed to measure. To assess the face validity of a measure, we need to know the meaning of the concept being measured and whether the information being collected is "germane to that concept."[19] For example, suppose you want to measure an individual's political ideology, whether someone is conservative, moderate, or liberal. It would not be a good idea to use an individual's responses to a question on party identification such as whether a person usually thinks of himself or herself as a Democrat, Republican, or Independent. It would be a mistake to assume that all Democrats are liberal, Republicans conservative, and Independents moderate. Similarly, some have argued that the results of many standard IQ tests measure intelligence *and* exposure to middle-class white culture, thus making the test results a less valid measure of intelligence.

In general, measures lack face validity when there are good reasons to question the correspondence of the measure to the concept in question. In other words, assessing face validity is essentially a matter of judgment. If

there is no consensus about the meaning of the concept to be measured, the face validity of one's measure is bound to be problematic.

A second kind of validity test, **content validity**, is similar to face validity. This test (actually a logical argument rather than a test) involves determining the full domain or meaning of a particular concept and then making sure that measures of all portions of this domain are included in the measurement technique. For example, suppose you wanted to design a measure of the extent to which a nation's political system is democratic. As noted earlier, democracy means many things to many people. Ross Burkhart and Michael Lewis-Beck in their study of the relationship between economic development and democracy relied on data compiled by Raymond D. Gastil. Gastil's measure of democracy included two dimensions, political rights and civil liberties. His checklists for each dimension consisted of eleven items.[20] Political scientists are often interested in concepts with multiple dimensions or complex domains and spend quite a bit of time discussing and justifying the content of their measures. Unfortunately, many political science concepts are so abstract and ill-defined that there is little agreement about their domain. This makes content validity less useful to political scientists than to other researchers.

A third way to evaluate the validity of a measure is by empirically demonstrating **construct validity**. When a measure of a concept is related to a measure of another concept with which the original concept is thought to be related, construct validity is demonstrated. In other words, a researcher may specify, on theoretical grounds, that two concepts ought to be related (say, political efficacy with political participation, or education with income). The researcher then develops a measure of each of the concepts and examines the relationship between them. If the measures are related, then one measure has construct validity for the other measure. If the measures are unrelated, there is an absence of construct validity. In that case the theoretical relationship is in error, one or more of the measures is not an accurate representation of the concept, or the procedure used to test the relationship is inappropriate. The absence of a hypothesized relationship does not tell us for certain that the measure is invalid, but the presence of a relationship gives us some assurance of a measure's validity.

A good example of an attempt to demonstrate construct validity may be found in the Educational Testing Service's (ETS) booklet describing the Graduate Record Exam (GRE), a standardized test required for admission to most graduate schools. Since GRE test scores are supposed to measure a person's aptitude for graduate study, presumably construct validity could be demonstrated if the scores did, in fact, accurately predict the person's performance in graduate school. Over the years ETS has tested the relationships between GRE scores and first-year graduate school grade-point average. The results,

shown in Table 4-1, appear to indicate that GRE scores are not very strong predictors of this measure of graduate school performance and therefore do not have construct validity. In fact, there has been much discussion in recent years about this very issue and the role that GRE scores should play in admissions decisions. But this is a good example of a situation where the absence of a strong relationship does not necessarily mean the measure lacks construct validity. Because persons with low GRE scores are generally not admitted to graduate school, we lack performance measures for them. Thus the people for whom we can test the relationship between scores and performance may be of similar ability and may not exhibit meaningful variation in their graduate school performance. Hence the test scores may be valid indicators of ability and may in fact show a stronger relationship to performance for a less selective sample of test takers (one that would include people who were not admitted to graduate school). The lack of a relationship in Table 4-1 undercuts claims of test score validity, but it does not necessarily disprove such claims.

A fourth way to demonstrate validity is through **interitem association**. This is the type of validity test most often used by political scientists. It relies

TABLE 4-1

Construct Validity of Graduate Record Examination Test Scores

Average Estimated Correlations of GRE, General Test Scores and Undergraduate Grade Point Average with Graduate First-Year Grade Point Average by Department Type

Type of Department	Number of Departments	Number of Examinees	Predictors					
			V	Q	A	U	VQA	VQAU
All Departments	1,038	12,013	.30	.29	.28	.37	.34	.46
Natural Sciences	384	4,420	.28	.27	.26	.36	.31	.44
Engineering	87	1,066	.27	.22	.24	.38	.30	.44
Social Sciences	352	4,211	.33	.32	.30	.38	.37	.48
Humanities & Arts	115	1,219	.30	.33	.27	.37	.34	.46
Education	86	901	.31	.30	.29	.35	.36	.47
Business	14	196	.28	.28	.25	.39	.31	.47

V = GRE Verbal, Q = GRE Quantitative, A = GRE Analytical, U = Undergraduate grade point average.
The departments included in these analyses participated in the GRE Validity Study Service between 1986 and 1990. A minimum of 10 departments and 100 examinees in any departmental grouping were required for inclusion in the table.

Source: Table 7, GRE® 2000–2001 Guide to the Use of Scores, 2000, 24. Reprinted by permission of Educational Testing Services, the copyright owner.

Note: The numbers in this table are product-moment correlations—numbers that can vary between −1.0 and +1.0 and that indicate the extent to which one variable is associated with another. The closer the correlation in to ±1, the stronger the relationship between the two variables; the closer the correlation is to 0.0, the weaker the relationship. Since the correlations between VQA and graduate first-year GPA in this table are between .30 and .40, the relationships are not very strong. Notice also that undergraduate GPA is also not a very strong predictor of graduate first-year GPA, but that together GRE scores and undergraduate GPA improve predictions.

on the similarity of outcomes of more than one measure of a concept to demonstrate the validity of the entire measurement scheme.

Let us return to the researcher who wants to develop a valid measure of liberalism. First the researcher might measure people's attitudes toward (1) school vouchers, (2) welfare, (3) protection of the rights of the accused, (4) military spending, (5) affirmative action, (6) social security benefit levels, (7) abortion, and (8) a progressive income tax. Then the researcher could determine how the responses to each question relate to the responses to each of the other questions. The validity of the measurement scheme would be demonstrated if there were strong relationships among people's responses across the eight questions.

The results of such interitem association tests are often displayed in a **correlation matrix** (Table 4-2). Such a display shows how strongly related each of the items in the measurement scheme is to all of the other items. In the hypothetical data shown in Table 4-2, we can see that people's responses to six of the eight measures were strongly related to each other while responses to the questions on protection of the rights of the accused and school vouchers were not part of the general pattern. Thus the researcher would probably conclude that the first six items all measure a dimension of liberalism and that, taken together, they are a valid measurement of liberalism.

TABLE 4-2
Interitem Association Validity Test of a Measure of Liberalism

	Welfare	Military Spending	Abortion	Social Security	Affirmative Action	Income Tax	School Vouchers	Rights of Accused
Welfare	x							
Military Spending	.56	x						
Abortion	.71	.60	x					
Social Security	.80	.51	.83	x				
Affirmative Action	.63	.38	.59	.69	x			
Income Tax	.48	.67	.75	.39	.51	x		
School Vouchers	.28	.08	.19	.03	.30	−.07	x	
Rights of Accused	−.01	.14	−.12	.10	.23	.18	.45	x

Note: Hypothetical data. The figures in this table are product-moment correlations, explained in the note to Table 4-1. A high correlation indicates a strong relationship between how people answered the two different questions designed to measure liberalism. The figures in the last two rows are considerably closer to 0.0 than are the other entries, indicating that people's answers to the school vouchers and rights of the accused questions did not follow the same pattern as their answers to the other questions. Therefore, it looks like school vouchers and rights of the accused are not part of the measure of liberalism accomplished with the other questions.

Such a procedure was used by Ada Finifter and Ellen Mickiewicz in their study of popular attitudes toward political change in the Soviet Union.[21] They designed six survey questions to measure attitudes toward the pace of political change, the perceived locus of responsibility (individual or collective) for individuals' social well-being, the acceptability of differences in individual incomes and standards of living, the acceptability of unconventional forms of political expression, the importance of free speech, and the level of support for competitive elections. The questions are as follows:

1. *(Rapid vs. slow change) "Some people think that to solve our most pressing problems it is necessary to make decisive and rapid changes, since any delay threatens to make things worse. Others, on the other hand, think that changes should be cautious and slow, since you can never be sure that they won't cause more harm than good. Which of these points of view are you more likely to agree with?*

   "Below are some widespread but contradictory statements relating to problems of the development of our society. Which would you be most likely to agree with?"

2. *(Individual vs. state responsibility) "The state and government should be mainly responsible for the success and well-being of people" or "People should look out for themselves, decide for themselves what to do for success in life."*

3. *(Income differences) "The state should provide an opportunity for everyone to earn as much as he can, even if it leads to essential differences in people's standard of living and income" or "The state should do everything to reduce differences in people's standard of living and income, even if they won't try to work harder and earn more."*

4. *(Protest vs. traditional methods) "Strikes, spontaneous demonstrations, political meetings and other forms of social protest are completely acceptable methods of mass conduct and an effective means for solving social problems" or "These forms of protest are undesirable for society; they should be avoided in favor of more peaceful, traditional and organized methods of solving social conflicts."*

5. *(Free speech vs. order) "To improve things in our country people should be given the opportunity to say what they want, even if it can lead to public disorder" or "Keeping the peace in society should be the main effort, even if it requires limiting freedom of expression."*

6. *(Competitive elections) "In the coming elections for local soviets, should we elect deputies from among several candidates, as we mostly did in the spring elections? or "Is it better to avoid the conflicts these elections generated and go back to the old system of voting?"*[22]

Finifter and Mickiewicz found a pattern in Soviet citizen responses to five of the six questions. Attitudes toward the locus of responsibility for individual well-being were only weakly related to the other five questions about political change. Hence the researchers were able to combine the answers to five of the measures into an attitude scale of "support for political change" as a result of the observed interitem associations among them.

Content and face validity are difficult to assess when there is a lack of agreement on the meaning of a concept, and construct validity, which requires a well-developed theoretical perspective, usually yields a less-than-definitive result. The interitem association test requires multiple measures of the same concept. Although these validity "tests" provide important evidence, none of them is likely to support an unequivocal decision concerning the validity of particular measures.

Problems with Reliability and Validity in Political Science Measurement

An example of research performed at the Survey Research Center at the University of Michigan illustrates the numerous threats to the reliability and validity of political science measures. In 1980 the Center conducted interviews with a national sample of eligible voters and measured their income levels with the following question:

> "Please look at this page and tell me the letter of the income group that includes the income of *all members of your family living here in 1979 before taxes.* This figure should include salaries, wages, pensions, dividends, interest, and all other income."

Respondents were given the following choices:

A. None or less than $2,000
B. $2,000–$2,999
C. $3,000–$3,999
D. $4,000–$4,999
E. $5,000–$5,999
F. $6,000–$6,999
G. $7,000–$7,999
H. $8,000–$8,999
J. $9,000–$9,999
K. $10,000–$10,999
M. $11,000–$11,999
N. $12,000–$12,999
P. $13,000–$13,999
Q. $14,000–$14,999
R. $15,000–$16,999
S. $17,000–$19,999
T. $20,000–$22,999
U. $23,000–$24,999
V. $25,000–$29,999
W. $30,000–$34,999
X. $35,000–$49,999
Y. $50,000 and over

Both the reliability and the validity of this method of measuring income are questionable. Threats to the reliability of the measure include the following:

1. Respondents may not know how much money they make and therefore incorrectly guess their income.
2. Respondents may also not know how much money other family members make and guess incorrectly.
3. Respondents may know how much they make but carelessly select the wrong categories.
4. Interviewers may circle the wrong categories when listening to the selections of the respondents.
5. Data entry personnel may touch the wrong numbers when entering the answers into the computer.
6. Dishonest interviewers may incorrectly guess the income of a respondent who does not complete the interview.
7. Respondents may not know which family members to include in the income total; some respondents may include only a few family members while others may include even distant relations.
8. Respondents whose income is on the border between two categories may not know which one to pick. Some pick the higher category, some the lower one.

Each of these problems may introduce some error into the measurement of income, resulting in inaccurate measures that are too high for some respondents and too low for others. Therefore, if this measure were applied to the same people at two different times we could expect the results to vary.

In addition to these threats to reliability, there are numerous threats to the validity of this measure:

1. Respondents may have illegal income they do not want to reveal and, therefore, may systematically underestimate their income.
2. Respondents may try to impress the interviewer, or themselves, by systematically overestimating their income.
3. Respondents may systematically underestimate their before-tax income if they believe too much money is being withheld from their paychecks.

This long list of problems with both the reliability and the validity of this fairly straightforward measure of a relatively concrete concept is worrisome. Imagine how much more difficult it is to develop reliable and valid measures when the concept is abstract (for example, intelligence, self-esteem, or liberalism) and the measurement scheme is more complicated.

The reliability and validity of the measures used by political scientists are seldom demonstrated to everyone's satisfaction. Most measures of political phenomena are neither completely invalid or valid nor thoroughly unreliable

or reliable, but rather are partially accurate. Therefore, researchers generally present the rationale and evidence that are available in support of their measures and attempt to persuade their audience that their measures are at least as accurate as alternative measures would be. Nonetheless, a skeptical stance on the part of the reader toward the reliability and validity of political science measures is often warranted.

Reliability and validity are not the same thing. A measure may be reliable without being valid. One may devise a series of questions to measure liberalism, for example, which yields the same result for the same people every time but which misidentifies individuals. A valid measure, on the other hand, will also be reliable since if it accurately measures the concept in question then it should do so consistently across measurements. It is more important, then, to demonstrate validity than reliability, but reliability is usually more easily and precisely tested.

The Precision of Measurements

Measurements should be not only accurate but also precise; that is, measurements should contain as much information as possible about the attribute or behavior being measured. The more precise our measures, the more complete and informative can be our test of the relationships between two or more variables.

Suppose, for example, that we wanted to measure the height of political candidates to see if taller candidates usually win elections. Height could be measured in many different ways. We could have two categories of the variable height, tall and short, and assign different candidates to the two categories based on whether they were of above-average or below-average height. Or we could compare the heights of candidates running for the same office and measure which candidate was the tallest, which the next tallest, and so on. Or we could take a tape measure and measure each candidate's height in inches and record that measure. Clearly, the last method of measurement captures the most information about each candidate's height and is, therefore, the most precise measure of the attribute.

When we consider the precision of our measurements, we refer to the **level of measurement**. The level of measurement involves the type of information that we think our measurements contain and the type of comparisons that can be made across a number of observations on the same variable. The level of measurement also refers to the claim we are willing to make when we assign numbers to our measurements.

There are four different levels of measurement: nominal, ordinal, interval, and ratio. Very few concepts used in political science research inherently

require a particular level of measurement, so the level used in any specific research project is a function of the imagination and resources of the researcher and the decisions made when the method of measuring each of the variables is developed.

A **nominal measurement** is involved whenever the values assigned to a variable represent only different categories or classifications for that variable. In such a case, no category is more or less than another category, simply different. For example, suppose we measure the religion of individuals by asking them to indicate whether they are Protestant, Catholic, Jewish, or something else. Since the four categories or values for the variable "religion" are simply different, the measurement is at a nominal level.

Nominal level measures ought to consist of categories that are exhaustive and mutually exclusive; that is, the categories should include all of the possibilities for the measure, and they should be differentiated in such a way that a case will fit into one and only one category. For example, the categories of the measure of religion are exhaustive (because of the "something else" category) as well as mutually exclusive (since presumably an individual cannot be of more than one religion). If we attempted to measure "types of political systems" with the categories democratic, socialist, authoritarian, undeveloped, traditional, capitalist, and monarchical, however, the categories would be neither exhaustive nor mutually exclusive. (In which one category would Japan, Great Britain, and India belong?) The difficulty of deciding the category into which many countries should be put would hinder the very measurement process the variable was intended to further.

An **ordinal measurement** assumes that more or less of a variable can be measured and that a comparison can be made on which observations have more or less of a particular attribute. For example, we could create an ordinal measure of formal education completed with the following categories: "eighth grade or less," "some high school," "high school graduate," "some college," "college degree or more." Notice that we are not concerned here with the exact difference between the categories of education, but only with whether one category is more or less than another. Or suppose we ask individuals three questions designed to measure social trust, and we believe that an individual who answers all three questions a certain way has more social trust than a person who answers two of the questions a certain way, and this person has more social trust than a person who answers one of the questions a certain way. We could assign a score of 3 to the first group, 2 to the second group, 1 to the first group, and 0 to those who did not answer any of the questions in a socially trusting manner. In this case, the higher the number, the more social trust an individual has. With an ordinal measure it does not

matter whether we assign to the four categories the numbers 0, 1, 2, 3; 5, 6, 7, 8; 10, 100, 1000, 10005; or 100, 101, 107, 111. The intervals between the numbers have no meaning; all that matters is that the higher numbers represent more of the attribute than do the lower numbers.

With an **interval measurement** the intervals between the categories or values assigned to the observations have meaning. For interval measures, the value of a particular observation is important not just in terms of whether it is larger or smaller than another value (as in ordinal measures), but also in terms of how much larger or smaller it is. For example, suppose we record the year in which certain events occurred. If we have three observations—1950, 1960, and 1970—we know that the event in 1950 occurred ten years before the one in 1960 and twenty years before the one in 1970. We also know that the difference between the 1950 and 1970 observations is twice the difference between the 1950 and 1960 or 1960 and 1970 observations. One-unit change (the interval) all along this measurement is identical in meaning: the passage of one year's time. This is not necessarily the case for the measure of social trust discussed earlier, since we are not certain in that example whether the difference between a score of 1 and a score of 2 is identical with the difference between a score of 2 and a score of 3.

Another characteristic of an interval level of measurement that distinguishes it from the next level of measurement (ratio) is that the zero point is arbitrarily assigned and does not represent the absence of the attribute being measured. For example, many time and temperature scales have arbitrary zero points. Thus, the year 0 A.D. does not indicate the beginning of time—if this were true, there would be no B.C. dates. Nor does 0°C indicate the absence of heat; rather, it indicates the temperature at which water freezes. For this reason, with interval level measurements we cannot calculate ratios; that is, we cannot say that 60°F is twice as warm as 30°F because it does not represent twice as much warmth.

The final level of measurement is a **ratio measurement.** This type of measurement involves the full mathematical properties of numbers. That is, the values of the categories order the categories, tell something about the intervals between the categories, and state precisely the relative amounts of the variable that the categories represent. If, for example, a researcher is willing to claim that an observation with ten units of a variable possesses exactly twice as much of that attribute as an observation with five units of that variable, then a ratio level measurement exists.

The key to making this assumption is that a value of zero on the variable actually represents the absence of that variable. Because ratio measures have a true zero point, it makes sense to say that one measurement is [×] times another. It makes sense to say a sixty-year-old person is twice the age of a thirty-

year-old person (60/30 = 2), while it does not make sense to say that 60° C is twice as warm as 30° C.[23]

Identifying the level of measurement of variables is important since it affects the data analysis techniques that can be used and the conclusions that can be drawn about the relationships between variables. However, the decision is not always a straightforward one, and there is often uncertainty and disagreement among researchers concerning these decisions. Very few phenomena inherently require one particular level of measurement. Often a phenomenon can be measured with any level of measurement, depending upon the particular technique designed by the researcher and the claims that the researcher is willing to make about the resulting measure.

Political science researchers have measured many concepts at the ratio level. People's ages, unemployment rates, percent vote for a particular candidate, and crime rates are all examples of measures that contain a zero point and represent the full mathematical properties of the numbers used. However, more political science research has probably relied upon nominal and ordinal level measures than interval or ratio level measures. This has restricted the types of hypotheses and analysis techniques that political scientists have been willing and able to use.

Researchers usually try to devise as high a level of measurement for their concepts as possible (nominal being the lowest level of measurement and ratio the highest). With a higher level of measurement, more advanced data analysis techniques can be used and more precise statements about the relationships between variables can be made. Consequently, one might start with a nominal level measure and think of a way to turn it into an ordinal or interval level measure. For example, a researcher investigating the effect of campaign spending on election outcomes could devise an ordinal level measure that simply distinguished between those candidates who spent more or less than their opponent. However, more information would be preserved if a ratio level variable measuring how much more (or less) a candidate spent than the opposition were devised. Similarly, researchers measuring attitudes or personality traits also often construct a scale or index from nominal level measures that permits at least ordinal level comparisons between observations.

Multi-item Measures

Many of the measures considered so far in this chapter have consisted of a single item. Fording's measures of mass insurgency as the number of incidents and welfare generosity as the annual growth in state AFDC rates, Page and Shapiro's measures of public opinion change, and Midlarsky's measure of political violence are all based on single measures of each phenomenon in

question. Often, however, researchers need to devise measures of more complicated phenomena that have more than one facet or dimension. Liberalism, democracy, access to electoral institutions and even land inequality, for example, are complex phenomena that may be measured in many different ways. In this situation, researchers often develop a measurement strategy that allows them to capture numerous aspects of a complex phenomenon while representing the existence of that phenomenon in particular cases with a single representative value. Usually this involves the construction of a multi-item index or scale representing the several dimensions of a complex phenomenon. These multi-item measures are useful because they enhance the accuracy of a measure, simplify a researcher's data by reducing it to a more manageable size, and increase the level of measurement of a phenomenon. In the remainder of this section we will describe several common types of indexes and scales.

Indexes

An **index** is a method of accumulating scores on individual items to form a composite measure of a complex phenomenon. An index is constructed by assigning a range of possible scores for a number of items, determining the score for each item for each observation, and then combining the scores for each observation across all of the items. The resulting summary score is the representative measurement of the phenomenon.

A researcher interested in measuring how much freedom there is in different countries, for example, might construct an index of political freedom by devising a list of items germane to the concept, determining where individual countries score on each of the items, and then adding these scores together to get a summary measure. In Table 4-3 such a hypothetical index is used to measure the amount of freedom in countries A through E.

The index in Table 4-3 is a simple, additive one; that is, each of the items counts equally toward the calculation of the index score, and the total score is the summation of the individual item scores. However, indexes may be constructed with more complicated aggregation procedures and by counting some items as more important than others. In the preceding example a researcher might consider some indicators of freedom more important than others and wish to have them contribute more to the calculation of the final index score. This could be done either by weighting (multiplying) some item scores by a number indicating their importance or by assigning a higher score than 1 for those attributes considered more important.

Fording's measure of access to electoral institutions included two items: whether or not a state's legislature was apportioned (representatives elected on a one person–one vote principle from equal-sized electoral districts) and

TABLE 4-3
Hypothetical Index for Measuring Freedom in Countries

	Country A	Country B	Country C	Country D	Country E
Does the country possess:					
Privately owned newspapers	1	0	0	0	1
Legal right to form political parties	1	1	0	0	0
Contested elections for significant public offices	1	1	0	0	0
Voting rights extended to most of the adult population	1	1	0	1	0
Limitations on government's ability to incarcerate citizens	1	0	0	0	1
Index Score	5	3	0	1	2

Note: Hypothetical data. The score is 1 if the answer is yes, 0 if no.

whether or not the Voting Rights Act of 1965 had been implemented within the state. Each item was scored 0 or 1 and the results of the items were multiplied. Thus only states whose legislative bodies were apportioned and complied with the Voting Rights Act received a score of 1.

Indexes are often used with public opinion surveys to measure political attitudes. This is because attitudes are complex phenomena and we usually do not know enough about them to devise single-item measures. So we often ask several questions of people about a single attitude and aggregate the answers to represent the attitude. A researcher might measure attitudes toward abortion, for example, by asking respondents to choose one of five possible responses—strongly agree, agree, undecided, disagree, and strongly disagree—to the following three statements: (1) Abortions should be permitted in the first three months of pregnancy; (2) Abortions should be permitted if the woman's life is in danger; (3) Abortions should be permitted whenever a woman wants one.

An index of attitudes toward abortion could be computed by assigning numerical values to each response (such as 1 for strongly agree, 2 for agree, 3 for undecided, and so on) and then adding the values of a respondent's answers to these three questions. (The researcher would have to decide what to do when a respondent did not answer one or more of the questions.) The lowest possible score would be a 3, indicating the most extreme pro-abortion attitude, and the highest possible score would be a 15, indicating the most ex-

treme anti-abortion attitude. Scores in between would indicate varying degrees of approval of abortion.

Finifter and Mickiewicz, the researchers who attempted to measure attitudes toward political change in the former Soviet Union, developed this type of index of attitudes. Once they had decided that there was a pattern in the responses to five of their questionnaire items, they then assigned a score of +1 to proreform answers and a –1 to status quo (opposite) answers and summed each individual's answers to the five questions. What resulted were index scores representing individual answers to all five questions that ranged in value from +5 to –5. This single index score of attitudes toward political change was then used in further analysis.

Another example of a multi-item index appears in a study of attitudes toward feminism in Europe.[24] To determine the extent and distribution of attitudinal support for feminism across European society, Lee Ann Banaszak and Eric Plutzer constructed a measure of profeminism attitudes. Respondents were asked six questions about various aspects of a feminist belief system (e.g., achieving equality between women and men in their work and careers, fighting against people who would like to keep women in a subordinate role, achieving gender equality in responsibilities for child care) and were given scores ranging from 0 to 3 for the degree of agreement with each profeminist statement. Responses across the six items were then summed to yield a profeminism index score that varied from 0 to 18.

Indexes are typically fairly simple ways of producing single representative scores of complicated phenomena such as political attitudes. They are probably more accurate than most single-item measures, but they may also be flawed in important ways. Aggregating scores across several items assumes, for example, that each item is equally important to the summary measure of the concept and that the items used faithfully encompass the domain of the concept. Although individual item scores can be weighted to change their contribution to the summary measure, there is often little information upon which to base a weighting scheme.

Several standard indexes are often used in political science research. The FBI crime index and the consumer price index, for example, have been used by many researchers. Although simple summation indexes are generally more accurate than single-item measures of complicated phenomena would be, it is often unclear how valid they are or what level of measurement they represent.

Scales

Although indexes are often an improvement over single-item measures, there is also an element of arbitrariness in their construction. Both the selection of particular items making up the index and the way in which the scores on in-

dividual items are aggregated are based on the judgment of the researcher. Scales are also multi-item measures, but the selection and combination of items in them is more systematically accomplished than is usually the case for indexes. Over the years several different kinds of multi-item scales have been used frequently in political science research. We will discuss three of them: Likert scales, Guttman scales, and the semantic differential.

A **Likert scale** score is calculated from the scores obtained on individual items. Each item generally asks a respondent to indicate a degree of agreement or disagreement with the item, as with the abortion questions discussed earlier. A Likert scale differs from an index, however, in that once the scores on each of the items are obtained, only some of the items are selected for inclusion in the calculation of the final score. Those items that allow a researcher to distinguish most readily those scoring high on an attribute from those scoring low will be retained, and a new scale score will be calculated based only on those items.

For example, recall the researcher interested in measuring the liberalism of a group of respondents. Since definitions of liberalism vary, the researcher cannot be sure how many aspects of liberalism need to be measured. With Likert scaling the researcher would begin with a large group of questions thought to express various aspects of liberalism that respondents would be asked to agree or disagree with. A provisional Likert scale for liberalism, then, might look like this:

	Strongly Disagree (1)	*Disagree (2)*	*Undecided (3)*	*Agree (4)*	*Strongly Agree (5)*
The government should ensure that no one lives in poverty.	___	___	___	___	___
Military spending should be reduced.	___	___	___	___	___
It is more important to take care of people's needs than it is to balance the federal budget.	___	___	___	___	___
Social Security benefits should not be cut.	___	___	___	___	___
The government should spend money to improve housing and transportation in urban areas.	___	___	___	___	___

	Strongly Disagree (1)	*Disagree (2)*	*Undecided (3)*	*Agree (4)*	*Strongly Agree (5)*
Wealthy people should pay taxes at a much higher rate than poor people.	___	___	___	___	___
Busing should be used to integrate public schools.	___	___	___	___	___
The rights of persons accused of a crime must be vigorously protected.	___	___	___	___	___

In practice, a set of questions like this would be scattered throughout a questionnaire so that respondents do not see them as related. Some of the questions might also be worded in the opposite way (that is, so an "agree" response is a conservative response) to ensure genuine answers.

The respondents' answers to these eight questions would be summed to produce a provisional score. The scores in this case can range from 8 to 40. Then the responses of the most liberal and the most conservative people to each question would be compared; any questions with similar answers from the disparate respondents would be eliminated—such questions would not distinguish liberals from conservatives. A new summary scale score for all the respondents would be calculated from the questions that remained.

Likert scales are improvements over multi-item indexes because the items that make up the multi-item measure are selected in part based on the behavior of the respondents rather than on the judgment of the researcher. Likert scales suffer two of the other defects of indexes, however: the researcher cannot be sure that all of the dimensions of a concept have been measured, and the relative importance of each item is still arbitrarily determined.

The **Guttman scale** also employs a series of items to produce a scale score for respondents. Unlike the Likert scale, however, a Guttman scale is designed to present respondents with a range of attitude choices that are increasingly difficult to agree with; that is, the items composing the scale range from those easy to agree with to those difficult to agree with. Respondents who agree with one of the "more difficult" attitude items will also generally agree with the "less difficult" ones. (Guttman scales have also been used to measure attributes other than attitudes. Their main application has been in the area of attitude research, however, so an example of that type is used here.)

Let us return to the researcher interested in measuring attitudes toward abortion. He or she might devise a series of items ranging from "easy to agree with" to "difficult to agree with." Such an approach might be represented by the following items.

Do you agree or disagree that abortions should be permitted:

1. When the life of the woman is in danger.
2. In the case of incest or rape.
3. When the fetus appears to be unhealthy.
4. When the father does not want to have a baby.
5. When the woman cannot afford to have a baby.
6. Whenever the woman wants one.

This array of items seems likely to result in responses consistent with Guttman scaling. A respondent agreeing with any one of the items is likely to also agree with those items numbered lower than that one. This would result in the "stepwise" pattern of responses characteristic of a Guttman scale.

Suppose six respondents answered this series of questions, as shown in Table 4-4. Generally speaking, the pattern of responses is as expected; those who agreed with the "most difficult" questions were also likely to agree with the "less difficult" ones. However, the responses of three people (2, 4, and 5) to the question about the father's preferences do not fit the pattern. Consequently, the question about the father does not seem to fit the pattern and would be removed from the scale. Once that has been done, the stepwise pattern becomes clear.

With real data, it is unlikely that every respondent would give answers that fit the pattern perfectly. For example, in Table 4-4 respondent 6 gave an

TABLE 4-4
Guttman Scale of Attitudes toward Abortion

Respondent	Life of Woman	Incest or Rape	Unhealthy Fetus	Father	Afford	Anytime	No. of Agree Answers	Revised Scale Score
1	A	A	A	A	A	A	6	5
2	A	A	A	D	A	D	4	4
3	A	A	A	D	D	D	3	3
4	A	A	D	A	D	D	3	2
5	A	D	D	A	D	D	2	1
6	D	A	D	D	D	D	1	0

Note: Hypothetical data. A = Agree, D = Disagree.

"agree" response to the question about incest or rape. This response is unexpected and does not fit the pattern. Therefore, we would be making an error if we assigned a scale score of "0" to respondent 6. There are statistical procedures to calculate how well the data fit the scale pattern. When the data fit the scale pattern well (number of errors is small), researchers assume that the scale is an appropriate measure and that the respondent's "error" may be "corrected" (in this case, either the "agree" in the case of incest or rape or the "disagree" in the case of the life of the woman). There are standard procedures to follow to determine how to correct the data to make it conform to the scale pattern. We emphasize, however, that this is done only if the changes are few.

Guttman scales differ from Likert scales in that generally only one set of responses will yield a particular scale score. That is, to get a score of 3 on the abortion scale a particular pattern of responses (or something very close to it) is necessary. In the case of a Likert scale, however, many different patterns of responses can yield the same scale score. A Guttman scale is also much more difficult to achieve than a Likert scale since the items must have been ordered and be perceived by the respondents as representing increasingly more difficult responses to the same attitude.

Both Likert and Guttman scales have shortcomings in their level of measurement. The level of measurement produced by Likert scales is, at best, ordinal (since we do not know what the relative importance is of each item and so we cannot be sure that a "5" answer on one item is the same as a "5" answer on another), and the level of measurement produced by Guttman scales is usually assumed to be ordinal.

Another method of producing multi-item summary measures is a technique called the **semantic differential**. This technique presents respondents with a series of adjective pairs to bring out the ways in which people respond to some particular object. These responses may then be used to understand the dimensions or attributes of that object and/or to compare evaluations across objects.

Suppose, for example, that you were a political consultant preparing for the reelection campaign of an incumbent U.S. senator. You would probably be interested in the public's attitude toward your client so that you could identify his or her major strengths and weaknesses. If you were uncertain about the attitudes that people had toward your candidate you might use the semantic differential to explore those attitudes.

In the typical semantic differential application respondents are presented with adjective pairs (opposites) with seven response categories available for each pair, and they are asked to evaluate some object in terms of each adjective pair. For example, respondents might be asked to reveal their feelings toward a political candidate in the following way:

Listed below are several pairs of words that could be used to describe Senator X. Between the words in each pair are several blanks. Please put an X in the blank between each pair that best describes how you feel about Senator X.

Senator X

honest	___	___	___	___	___	___	___	dishonest
smart	___	___	___	___	___	___	___	dumb
sincere	___	___	___	___	___	___	___	insincere
superficial	___	___	___	___	___	___	___	profound
good	___	___	___	___	___	___	___	bad
serious	___	___	___	___	___	___	___	humorous
idealistic	___	___	___	___	___	___	___	realistic
strong	___	___	___	___	___	___	___	weak
pleasant	___	___	___	___	___	___	___	unpleasant
helpful	___	___	___	___	___	___	___	unhelpful
powerful	___	___	___	___	___	___	___	powerless
active	___	___	___	___	___	___	___	inactive
young	___	___	___	___	___	___	___	old
nice	___	___	___	___	___	___	___	awful

Research on the use of the semantic differential has discovered that there are often three primary underlying dimensions for attitudes toward most objects: an evaluative dimension (favorable versus unfavorable), a potency dimension (strong versus weak), and an activity dimension (active versus passive). Typically, responses to a set of adjective pairs are analyzed in terms of these three dimensions, allowing a researcher to infer the respondents' attitudes toward one or more objects of interest. In the example above, the adjective pairs honest-dishonest, good-bad, pleasant-unpleasant, nice-awful, helpful-unhelpful, sincere-insincere, and smart-dumb would probably capture the evaluative dimension while strong-weak, powerful-powerless, serious-humorous, and superficial-profound would represent the potency dimension. Young-old and active-inactive would measure activity. Asking citizens to indicate their perceptions about a candidate in this way would probably allow a political consultant to determine the perceived evaluation, potency, and activity of a candidate more accurately than would any single questionnaire item.

Factor Analysis

The procedures described so far for constructing multi-item measures are fairly straightforward. Sometimes, however, researchers attempt to construct measures of abstract, complicated phenomena where they are uncertain

about the domain to be measured. **Factor analysis** is a statistical technique that may be used to uncover patterns across a number of measures. It is especially useful when a researcher has a large number of measures and when there is uncertainty about how the measures are interrelated. Factor analysis is often used in attitudinal research to construct a limited number of attitude scales, and corresponding scale scores, out of a much larger number of questionnaire items.

The factor analysis procedure involves calculating a statistic that measures the relationships between every pair of measures and then looking for groups of measures that are closely related to each other. These groups of closely related measures are said to *load* on a *factor,* which measures a particular aspect or dimension of the phenomenon.

Let us take a look at an example. Banaszak and Plutzer, who were interested in studying feminism among men and women in Europe, included six questions meant to measure attitudes toward feminism on their questionnaire. They did not know, however, whether the six questions measured six different attitudes, one attitude, or something in between; nor did they know if all six questions were equally effective measures of feminism. A factor analysis of people's responses to the questions revealed that all six measures were, in fact, highly interrelated; that the six measures constituted one, and only one, factor, or attitude; and that each question was an equally good measure of feminism. This information led to the decision to construct a single attitude scale utilizing the answers to all six questions.

A more complicated and abstract use of factor analysis may be found in research explaining levels of governmental welfare spending during the period between 1960 and 1982 in eighteen capitalist democracies.[25] In this study, researchers Alexander Hicks and Duane Swank measured a number of attributes of the eighteen nation-states they studied, including partisan control of the government, the amount of electoral competition and voter turnout, working class organization and interest representation, the amount of governmental centralization and bureaucratization, and the existence of welfare-related policy precedents. By employing a factor analysis, the researchers were able to identify three dimensions of a nation-state's political institutions and process. One dimension, called "left corporatism," reflects union strength, class mobilization, and left-leaning governmental control. Another dimension, called "state centralization," captures various aspects of governmental centralization. The third dimension, called "bureaucratic patrimonialism," measures such things as the extent of bureaucratic power, resistance to mass enfranchisement, and class rigidity. The researchers constructed three scale scores made up of several measures to capture each of these dimensions and demonstrate their relationships with governmental welfare spending.

Factor analysis is just one of many techniques that have been developed to explore the dimensionality of measures and to construct multi-item scales. The readings listed at the end of this chapter include some resources for students who are especially interested in this aspect of variable measurement.

Through indexes and scales, researchers attempt to enhance both the accuracy and the precision of their measures. Although these multi-item measures have received most use in attitude research, they are often useful in other endeavors as well. Both indexes and scales require researchers to make decisions regarding the selection of individual items and the way in which the scores on those items will be combined to produce more useful measures of political phenomena.

Conclusion

To a large extent, a research project is only as good as the measurements that are developed and used in it. Inaccurate measurements will interfere with the testing of scientific explanations for political phenomena and may lead to erroneous conclusions. Imprecise measurements will limit the extent of the comparisons that can be made between observations and the precision of the knowledge that results from empirical research.

Despite the importance of good measurement, political science researchers often find that their measurement schemes are of uncertain accuracy and precision. Abstract concepts are difficult to measure in a valid way, and the practical constraints of time and money often jeopardize the reliability and precision of measurements. The quality of a researcher's measurements makes an important contribution to the results of his or her empirical research and should not be lightly or routinely sacrificed.

Sometimes the accuracy of measurements may be enhanced through the use of multi-item measures. With indexes and scales, researchers select multiple indicators of a phenomenon, assign scores to each of these indicators, and combine those scores into a summary measure. While these methods have been used most frequently in attitude research, they can also be used in other situations to improve the accuracy and precision of single-item measures.

Notes

1. Stephen C. Poe and C. Neal Tate, "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis," *American Political Science Review* 88 (December 1994): 853–872.
2. Bruce Bueno de Mesquita, Randolph M. Siverson, and Gary Woller, "War and the Fate of Regimes: A Comparative Analysis," *American Political Science Review* 86 (September 1992): 638–646.
3. Benjamin I. Page and Robert Y. Shapiro, "Effects of Public Opinion on Policy," *American Political Science Review* 77 (March 1983): 175–190.
4. Jeffrey A. Segal and Albert D. Cover, "Ideological Values and the Votes of U.S. Supreme Court Justices," *American Political Science Review* 83 (June 1989): 557–565.
5. B. Dan Wood and Richard W. Waterman, "The Dynamics of Political Control of the Bureaucracy," *American Political Science Review* 85 (September 1991): 801–828.
6. Richard C. Fording, "The Conditional Effect of Violence as a Political Tactic: Mass Insurgency, Welfare Generosity, and Electoral Context in the American States," *American Journal of Political Science* 41 (January 1997): 1–29.
7. Stephen Ansolabehere, Shanto Iyengar, Adam Simon, and Nicholas Valentino, "Does Attack Advertising Demobilize the Electorate?" *American Political Science Review* 88 (December 1994): 829–838.
8. Page and Shapiro, "Effects of Public Opinion."
9. Fording, "The Conditional Effect of Violence as a Political Tactic," 11.
10. Daron R. Shaw, "The Effect of TV Ads and Candidate Appearances on Statewide Presidential Votes, 1988–96," *American Political Science Review* 93 (June 1999): 345–362.
11. Segal and Cover, "Ideological Values."
12. Ibid., 559.
13. Manus I. Midlarsky, "Rulers and the Ruled: Patterned Inequality and the Onset of Mass Political Violence," *American Political Science Review* 82 (June 1988): 491–509.
14. John S. Dryzek and Jeffrey Berejikian, "Reconstructive Democratic Theory," *American Political Science Review* 87 (March 1993): 48–60.
15. Ibid.
16. Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment*, Series on Quantitative Applications in the Social Sciences, No. 07-001, Sage University Papers (Beverly Hills, Calif.: Sage, 1979).
17. Philip E. Converse, "The Nature of Belief Systems in Mass Publics," in David E. Apter, ed., *Ideology and Discontent* (New York: Free Press, 1964); D.M. Vaillancourt, "Stability of Children's Survey Responses," *Public Opinion Quarterly* 37 (fall 1973): 373–387; J. Miller McPherson, Susan Welch, and Cal Clark, "The Stability and Reliability of Political Efficacy: Using Path Analysis to Test Alternative Models," *American Political Science Review* 71 (June 1977): 509–521; and Philip E. Converse and Gregory B. Markus, "The New CPS Election Study Panel," *American Political Science Review* 73 (March 1979): 32–49.
18. Raymond E. Wolfinger and Steven J. Rosenstone, *Who Votes?* (New Haven: Yale University Press, 1980), Appendix A.
19. Kenneth D. Bailey, *Methods of Social Research* (New York: Free Press, 1978), 58.
20. Ross E. Burkhart and Michael S. Lewis-Beck, "Comparative Democracy: The Economic Development Thesis," *American Political Science Review* 88 (December 1994): Appendix A.

21. Ada W. Finifter and Ellen Mickiewicz, "Redefining the Political System of the USSR: Mass Support for Political Change," *American Political Science Review* 86 (December 1992): 857–874.
22. Ibid.
23. The distinction between an interval and a ratio level measure is not always clear, and some political science texts do not distinguish between them. Interval level measures in political science are rather rare; ratio level measures (money spent, age, number of children, years living in the same location, for example) are more common.
24. Lee Ann Banaszak and Eric Plutzer, "The Social Bases of Feminism in the European Community," *Public Opinion Quarterly* 57 (spring 1993): 29–53.
25. Alexander M. Hicks and Duane H. Swank, "Politics, Institutions and Welfare Spending in Industrialized Democracies, 1960–82," *American Political Science Review* 86 (September 1992): 658–674.

Terms introduced

ALTERNATIVE-FORM METHOD. A method of calculating reliability by repeating different but equivalent measures at two or more points in time.
CONSTRUCT VALIDITY. Validity demonstrated for a measure by showing that it is related to the measure of another concept.
CONTENT VALIDITY. Validity demonstrated by ensuring that the full domain of a concept is measured.
CORRELATION MATRIX. A table showing the relationships among a number of discrete measures.
FACE VALIDITY. Validity asserted by arguing that a measure corresponds closely to the concept it is designed to measure.
FACTOR ANALYSIS. A statistical technique useful in the construction of multiple-item scales to measure abstract concepts.
GUTTMAN SCALE. A multi-item measure in which respondents are presented with increasingly difficult measures of approval for an attitude.
INDEX. A multi-item measure in which individual scores on a set of items are combined to form a summary measure.
INTERITEM ASSOCIATION. A test of the extent to which the scores of several items, each thought to measure the same concept, are the same. Results are displayed in a correlation matrix.
INTERVAL MEASUREMENT. A measure for which a one-unit difference in scores is the same throughout the range of the measure.
LEVEL OF MEASUREMENT. An indication of what is meant by assigning scores or numerals to empirical observations.
LIKERT SCALE. A multi-item measure in which the items are selected based on their ability to discriminate between those scoring high and those scoring low on the measure.

Measurement. The process by which phenomena are observed systematically and represented by scores or numerals.
Nominal measurement. A measure for which different scores represent different, but not ordered, categories.
Operational definition. The rules by which a concept is measured and scores assigned.
Ordinal measurement. A measure for which the scores represent ordered categories that are not necessarily equidistant from each other.
Ratio measurement. A measure for which the scores possess the full mathematical properties of the numbers assigned.
Reliability. The extent to which a measure yields the same results on repeated trials.
Semantic differential. A technique for measuring attitudes toward an object in which respondents are presented with a series of opposite adjective pairs.
Split-halves method. A method of calculating reliability by comparing the results of two equivalent measures made at the same time.
Test-retest method. A method of calculating reliability by repeating the same measure at two or more points in time.
Validity. The correspondence between a measure and the concept it is supposed to measure.

Exercises

1. Read the article by David H. Folz, "Municipal Recycling Performance: A Public Sector Environmental Success Story," *Public Administration Review* 59 (July–August 1999): 336–345.
   a. How was the importance of problems in municipal recycling measured? What is the level of measurement for this variable?
   b. How was change in recycling participation measured? What is its level of measurement?
   c. How were recycling program costs measured? What is the level of measurement? What problems did Folz encounter in trying to measure program costs?
2. Refer to the article by Jon S.T. Quah, "Corruption in Asian Countries: Can It Be Minimized?" *Public Administration Review* 59 (November–December 1999): 482–494. Review the discussion of the measurement of political corruption on page 484. What exactly is being measured? How is it measured? And what level of measurement are the measures?

3. Read the article by Jeff Yates and Andrew Whitford, "Presidential Power and the United States Supreme Court," *Political Research Quarterly* 51 (June 1998): 539–550. How are the variables for presidential approval, judicial appointment, and policy area of cases measured and what are their levels of measurement?
4. Read the article by Stephen C. Poe and C. Neal Tate, "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis," *American Political Science Review* 88 (December 1994): 853–872. The authors do not measure the complete domain of the concept of human rights. What components do they leave out and why do they limit their definition of human rights abuses to those which violate the "integrity of the person"? Poe and Tate also discuss their measurement of democracy. What are the limitations of the two measures of democracy that they use? Are these reliability or validity problems?
5. What would be the level of measurement of the following measures?
   a. Current marital status (married, divorced, widowed, never married)
   b. College class (freshman, sophomore, junior, senior)
   c. Percent of news time devoted to election coverage during the evening news broadcast
   d. Attitudes toward government spending on the environment (too little, just about right, too much)
   e. Partisan control of Congress (Republican, split, Democratic)
   f. Republican control of Congress (neither chamber, one chamber only, both chambers)
   g. Sales tax rates in each of the fifty American states
   h. Month in which presidential primary is held in a state
   i. Percent of eligible voters registered to vote
   j. Whether or not a state requires annual safety inspection of automobiles

Suggested Readings

Carmines, Edward G., and Richard A. Zeller. *Reliability and Validity Assessment.* Series on Quantitative Applications in the Social Sciences. No. 07-001, Sage University Papers. Beverly Hills, Calif.: Sage, 1979.

DeVellis, Robert F. *Scale Development.* Newbury Park, Calif.: Sage, 1991.

Hatry, Harry P. *Performance Measurement: Getting Results.* Washington, D.C.: Urban Institute Press, 1999.

Kerlinger, Fred N. *Behavioral Research.* New York: Holt, Rinehart & Winston, 1979.

Kim, Jae-On, and Charles W. Mueller. *Introduction to Factor Analysis.* Newbury Park, Calif.: Sage, 1978.

Lodge, Milton. *Magnitude Scaling.* Newbury Park, Calif.: Sage, 1983.

Maranell, Gary M. *Scaling: A Sourcebook for Behavioral Scientists.* 4th ed. Hawthorne, N.Y.: Longman, 1983.

Rabinowitz, George. "Nonmetric Multidimensional Scaling and Individual Difference Scaling." In William D. Berry and Michael S. Lewis-Beck, eds. *New Tools for Social Scientists,* 77–107. Beverly Hills, Calif.: Sage, 1986.

Robinson, John P., Jerrold G. Rusk, and Kendra B. Head. *Measures of Political Attitudes.* Ann Arbor, Mich.: Institute for Social Research, 1969.

Rubin, Herbert J. *Applied Social Research.* Columbus, Ohio: Merrill, 1983.

CHAPTER

Resea

In the previ
search hypo
the variable
chapter we
dent and dej
conclusions
sions are wl

In genera
tends to fulf
will be mad
how the ob
cedures to l
is to test hy
accomplishe

A researc
in the proce
model of pr
causal relat
the researc
whether th
tion or to di

Developi
search ques
design may
original an
search desi
research d

Many fa
purpose of
ploratory, d
a research
may test th
impossible