The Craft of Political Research

Eighth Edition

W. Phillips Shively
University of Minnesota

Longman
Boston Columbus Indianapolis New York San Francisco
Upper Saddle River Amsterdam Cape Town Dubai London
Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City
Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

To Barbara

Acquisitions Editor: Vikram Mukhija
Editorial Assistant: Toni Magyar
Marketing Manager: Lindsey Prudhomme
Production Manager: Wanda Rockwell
Project Coordination, Text Design, and Electronic Page Makeup: Shiny Rajesh/Integra Software Services, Ltd.
Creative Director: Jayne Conte
Cover Designer: Bruce Kenselaar/Mary Steinert
Cover Illustration/Photo: Corbis Corporation
Printer and Binder: Courier Companies, Inc.

Library of Congress Cataloging-in-Publication Data

Cataloging-in-Publication Data for this title can be obtained from the Library of Congress.

3 4 5 6 7 8 9 10—HAM—13 12 11

Longman
is an imprint of

PEARSON

www.pearsonhighered.com

ISBN-13: 978-0-205-79120-0
ISBN-10: 0-205-79120-4

Brief Contents

Dimensional analysis of this sort allows us to see the interrelations of various approaches to a question and can also give us a rich framework with which to apply a multidimensional concept. It also allows us to compare the importance of the dimensions. Implicit in each later scholar in the discussion sketched above was the idea that his notion of power was deeper and more basic than those that went before. A dimensional analysis gives us a structure within which we can address this question.

Further Discussion

Formal treatment of dimensional analysis was introduced by Allan H. Barton, "The Concept of Property-Space in Social Research" (1955). Philip E. Jacob's "A Multi-dimensional Classification of Atrocity Stories" (1955) furnishes a good example of dimensional analysis in practice.

Some examples from political science are Chapter 11 of Robert Dahl's *Political Oppositions in Western Democracies* (1966), a first-rate analysis of the relevant dimensions for classifying "opposition"; Harry Eckstein's *Pressure Group Politics* (1960), especially pp. 15–40, in which he classifies pressure group activities; and the third chapter of his *Division and Cohesion in Democracy* (1966), an excellent dimensional analysis of "political division"; Hanna Pitkin's *Representation* (1969), a collection of various writings on the concept of representation, among which Pitkin's own essay is particularly insightful; also Pitkin's *The Concept of Representation* (1967); and Giovanni Sartori's *Parties and Party Systems* (1976).

One branch of political philosophy is the "analytic political philosophy" approach, which seeks to study political ideas by a close examination of the meaning of concepts used to describe politics. This approach is reviewed in Richard Bernstein's *Restructuring of Social and Political Theory* (1978) and in Felix Oppenheim, "The Language of Political Inquiry: Problems of Clarification" (1975).

One branch of political philosophy is the Finally, as an exercise, you might consider the conceptual problems involved in the well-worn aphorism of Lord Acton: "Power tends to corrupt and absolute power corrupts absolutely." How would you analyze this conceptually? *Can* it be analyzed?

Problems of Measurement: Accuracy

In this chapter, I explore problems of accurate measurement. These are problems that arise when we try to relate the actual operations in a piece of research—that is, measures of things—to the concepts that form the basis of our underlying theory. Concepts, of course, exist only in the mind. One necessary assumption, if we are to claim that a piece of research has tested a given theory, is that the things measured in the research correspond to the things in the theorist's mind.

This is often a difficult assumption to make. In the preceding chapter, you saw one kind of problem that can stand in the way of it. The political scientist who wanted to measure the amount of interaction between nations found that there was no single satisfactory indicator of "interaction." A number of things—trade, mail exchanges, alliance, and so on—partook of "interaction," but no one of them alone was synonymous with the mental construct.

In the social sciences, only rarely are we able to measure our concepts directly. Consider, for example, the concepts "social class," "respect for the presidency," and "power in the community." Any variables we would choose to measure these concepts correspond only indirectly or in part to our mental constructs. This is the basic problem of measurement in the social sciences.

Consider the concept "social status." Among social scientists there are two popular versions of this concept: "subjective social status," the class that individuals consider themselves as belonging to; and "objective social status," an individual's rank with regard to prestige along social hierarchies such as education, income, and occupation. Neither version of the concept can be measured directly.

In the case of "subjective social status," we cannot measure directly what individuals feel about their status. We know what they report, but their replies to our inquiries may not be what we are looking for. They may not know what they "really" feel, for instance; or they may misunderstand the question and give a misleading answer. Then again, a person may feel differently from one day to the next, in which case the measure of his status will depend on our rather arbitrary choice of day.

In the case of "objective social status," again we cannot measure the variable directly. "Objective status" has something to do with income, education, occupation, and various other hierarchies, some of which we may not know about. None of these provides a sufficient measure in itself. For example, if we tried to use occupation alone as a measure of social status, we would be faced with the difficult question of whether a plumber who made $40,000 a year was really of lower social status than a bank teller who made $25,000 a year. Similarly, if we tried to use income alone as a measure, we would be faced with the problem of what to do with a retired teacher, whose income might be near the poverty level. "Social status" in this case is a concept that is related to a number of measurable things but is related only imperfectly to each of them. The best we can do in measuring it is to combine the various measurable indicators into a pooled measure that is roughly related to the concept "objective social status."

We encounter similar problems in measuring the other concepts I have cited as examples. Like many other variables in political science, these concepts are of considerable interest and use in theories but are by their nature impossible to measure directly. The general problem posed by such variables is presented schematically in Figure 4-1.

As you saw in Chapter 2, in political research we are commonly interested in relating concepts through a theory. This is always true in theory-oriented research, and it is true most of the time in engineering research as well. If we cannot measure directly the concepts we wish to use, we find ourselves in the position depicted in Figure 4-1. We want to say, "Concept A bears a relationship of thus-and-so form to concept B." But all that we can observe is the relationship between measure A and measure B. Whether what we say about the measures is an accurate statement of the relationship between the concepts depends on the unobserved relationships between concept A and measure A and between concept B and measure B. We can only assume what these relationships are. Like the theory itself, these relationships cannot be observed.

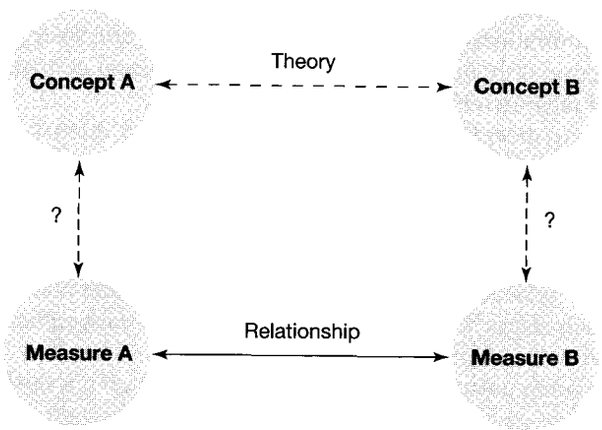

FIGURE 4-1 The Problem of Measurement

As an example, suppose you want to assess the theory tested by Diehl and Kingston, that countries that are increasing their armaments tend to engage in aggressive international policies (see Chapter 1). You might be faced with the following relationships:

1. *Relationship between the concept "increasing armaments" and the measure of "increasing armaments."* You clearly cannot measure increases in a nation's armaments directly; only a national intelligence apparatus has the facilities to do that, and even then, the result is imperfect. Therefore, you might take as your measure the country's reported expenditures on armaments. Now, a country that is not preparing to launch an aggressive military venture would have less reason to lie about an arms buildup than would a country (such as Germany in 1933), that is consciously preparing for aggression. Therefore, the relationship between concept and measure in this case might be: When a country is not building up its armaments, or when it is building them up in order to launch an aggressive action, its reported expenditures on arms will not increase.
2. *Relationship between the concept "increasing armaments" and the concept "aggressive international policies."* Let us assume, for this example, that countries that are increasing their armaments do tend to engage in aggressive international policies.
3. *Relationship between the concept "aggressive international policies" and the measure of "aggressive international policies."* Let us assume, for this example, that we are able to develop a measure that corresponds almost perfectly to the concept "aggressive international policies." (In practice, of course, this would be a difficult variable to measure, and it certainly would be necessary first to analyze the varied dimensions involved in "aggression" and "policies" in order to state more clearly just what was meant by the concept.)

We now find ourselves in the position depicted in Figure 4-2. Here, because of peculiarities in the relationships between the concepts and the measures of these concepts, the relationship you can observe between the measures turns out to be the opposite of the true relationship between the concepts. Worse yet, inasmuch as the two measures and the connection between them are all that you can observe, you would have no way of knowing that this was happening. This is why I have called indirect measurement of concepts *the* problem of measurement.

One solution to the problem might be to measure variables only directly. Some concepts are directly measurable. A few examples are people's votes if an election is nonsecret and you tabulate the result yourself; statements you hear made on the Senate floor; a bomb you see dropped or thrown.[1]

[1]It is reasonable, also, to include here reliable observers' accounts of such events. Even though the measurement of these things is technically indirect, if you accept another observer's account of them, you should be able to achieve a very tight fit between the concept "event happens" and the measure "reliable observer says that event happens."

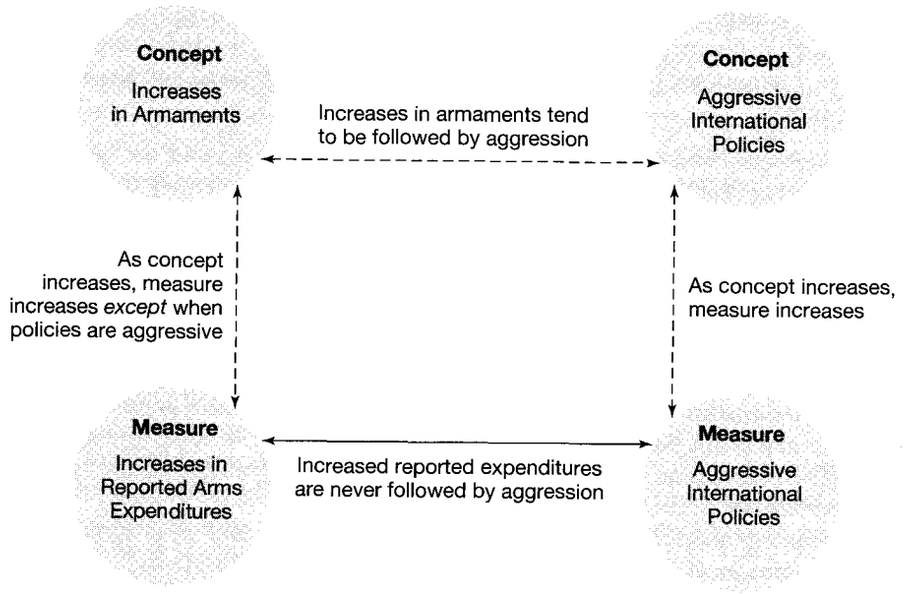

FIGURE 4-2 Example of the Problem of Measurement

The difficulty with such a solution is that concepts that can be measured directly are usually trivial in and of themselves. They are too idiosyncratic to use in general, interesting theories. I would hate to say that particular statements by U.S. senators lead nowhere, but as far as political theory is concerned, it is true. Any given statement can apply only to itself. It takes on a general meaning only if it is placed in a category, so that it can be compared with other statements. For instance, Senator_____'s statement, "The president's policies are bankrupting the people of my state," is not intrinsically of theoretical interest. It can be placed in various categories, however: "statements of opposition to the president," "statements of concern for constituents' needs," "bombastic statements." Placing it in one of these categories allows us to compare it with other senatorial statements and to develop theories about the causes and effects of such statements.

Note, though, that by placing it into a category, we have used the statement as an indirect measure of the concept which the category represents. No given statement is a perfect case of the "bombastic statement" or of the "statement of opposition to the president." Rather, a number of statements *approximate* each category, and we choose to use these statements as indirect measures of the abstract concept we cannot measure.

To sum up the argument thus far: For a concept to be useful in building theories, it usually must be an abstraction, which cannot be measured directly. Further, of those interesting concepts that *are* in principle directly measurable (how individuals vote in an election, for example), many cannot be measured directly for

practical reasons (the elections are held with a secret ballot). Therefore, most of the time we must work with variables that are indirect measures of the concepts in which we are interested. This means that there are interposed, between our (concrete) operations and the (abstract) theory we want to work on, relationships between our concrete concepts and their abstract measures.[2] This is the situation illustrated graphically in Figures 4-1 and 4-2. *The chief problem of measurement is to ensure, as much as possible, that the relationships between concepts and measures are such that the relationship between the measures mirrors the relationship between the concepts.*

Problems we may encounter in trying to achieve this correspondence between measures and concepts fall under two headings: problems of measure *reliability* and problems of measure *validity*.

RELIABILITY

A measure is reliable to the extent that it gives the same result again and again if the measurement is repeated. For example, if people are asked several days in a row whether they are married, and their answers vary from one day to the next, the measure of their marital state is unreliable. If their answers are stable from one time to the next, the measure is reliable.

The analogy of measuring with a yardstick may help make the meaning of reliability clear. If an object is measured a number of times with an ordinary wooden yardstick, it will give approximately the same length each time. If the yardstick were made of an elastic material, its results would not be so reliable. It might say that a chair was 20 inches high one day, 16 the next. Similarly, if it were made of a material that expanded or contracted greatly with changes in temperature, its results would not be reliable. On hot days it would say that the chair was shorter than on cold days. In fact, the choice of wood as a material for yardsticks is in part a response to the problem of reliability in measurement, a problem certainly not confined to the social sciences. Wood is cheap, rigid, and relatively unresponsive to changes in temperature.

There are many sources of unreliability in social science data. The sources vary, depending on what kinds of data are used. Official statistics, for example, may be unreliable because of an unusual number of clerical errors or because of variability in how categories are defined from one time to the next. ("Votes cast in an election," for instance, may mean "all votes, including spoiled ballots" at one time, "all valid votes" at another.) Attitude measures may be unreliable because a question is hard for respondents to understand, and they interpret it one way at one time, another way the next. Or the people entering their responses into the computer may make mistakes.

[2]Technically, these relationships are called *epistemic correlations*. "Correlation" means relationship, and "epistemic" has the same root as "epistemology"—the study of how we know.

As an illustration, let us list the various sources of unreliability that might be involved in tabulating responses even to a simple question about whether a person is married:

1. The question might be phrased badly, so that respondents sometimes interpreted it to mean, "Are you now married?" and sometimes, "Have you ever been married?" It might not be clear how people who were separated from their spouses, but not divorced, should answer.
2. Respondents might be playing games with the interviewer, answering questions randomly.
3. Dishonest interviewers might be playing games with the researcher by filling out the forms themselves instead of going through the trouble of getting the respondents to answer them.
4. Respondents' answers might depend on their mood. Perhaps they would answer "yes" when they had had a good day, or "no" when they had had a bad day.
5. Respondents' answers might depend on the context of the interview. A person might say "no" to an attractive interviewer and "yes" to everyone else.
6. There might be simple clerical errors in copying down the answers, either by the interviewer on the spot or by the person who transcribes the interviewers' copy into a computer.

Admittedly, some of these possibilities are far-fetched. The example itself is a bit strained, inasmuch as straightforward informational items like this one can usually be measured with reasonable reliability. But the same sorts of conditions affect the reliability of less straightforward survey questions, such as "What do you like about candidate X?" "What social class do you consider yourself to be a member of?" and "Do you feel people generally can be trusted?"

A few examples from a classic study of reliability (Asher, 1974) will help give you a sense of the magnitude of this problem, at least in American survey research. Even on attributes that should be relatively easy to measure reliably, such as gender and race, some errors appeared when the same respondents were interviewed at two-year intervals by a well-administered survey. On the average, the reported gender of respondents changed 0.5 percent of the time from one interview to the next, whereas race did not change at all. Characteristics on which it is somewhat easier to be vague or mistaken showed substantial unreliability. For example, the report of respondents' educational background showed *lower* education two years later (which is logically impossible), an average of 13 percent of the time. Presumably, questions that permit a considerable degree of interpretation, such as attitudinal questions, would show even more unreliability.

Reliability as a Characteristic of Concepts

Thus far, I have treated the unreliability of a measure as if it were a result of unpredictability in the relationship between the concept and its measure. An additional source of unreliability in the measure is variability in the "true value" of the concept. In our previous example, perhaps some people got married, or

divorced, from one time to the next. This source of unreliability would be easily distinguishable from others, however, because it would show up as a recognizable pattern of stable answers up to a certain point, followed by changed, but once again stable, answers.

A more interesting case is presented when the true value itself varies randomly. This situation sometimes provides the basis for interesting theories. In one study, Converse (1964) noted that on many standard questions of political policy, people's attitudes appeared to vary randomly across time. He concluded that on certain issues, the mass public simply did not form stable opinions; and he went on to draw interesting comparisons between elite and mass opinion, based on that conclusion.

Note that to reach this conclusion, Converse had to assume that he had effectively eliminated other sources of unreliability, such as interviewer error and confusion about the meaning of questions. Having first eliminated these sources of unpredictability in the relationship between concept and measure, he could then treat the unreliability in his measure as a reflection of unreliability in the concept. Christopher Achen (1975) later challenged Converse's conclusions on just these grounds.

Testing the Reliability of a Measure

Although unreliability may sometimes spur on further theoretical research, as it did in this case, it is usually a barrier we want to eliminate. Careful work is the best way to achieve reasonable reliability—double-checking all clerical work, trying out the questionnaire on a small pilot study in order to catch and correct unclear questions, and so on.

We often wish to know how successfully we have reduced unreliability. A number of tests have been developed to help researchers check the reliability of a measure. I shall describe two of them briefly.

The *test-retest check for reliability* simply consists of repeating the measurement a second time, allowing for a suitable interval of time between the two measurements. If the second measure strongly resembles the first—that is, if the measure is stable over the elapsed time—it is considered relatively reliable. One problem with this test, of course, is that there is no way to distinguish instability that stems from "real" unreliability in the concept being measured, from instability due to problems in the measurement process.

Another test, the *split-half check for reliability*, avoids this problem. It is particularly useful whenever a measure is multidimensional—for instance, a measure of "social status," which is made by combining such items as an individual's income, occupation, education, house size, and neighborhood into a single summary measure; or a measure of "welfare policy expenditures," comprising such disparate items as welfare payments, unemployment relief, hospital subsidies, and school lunch programs.

In the split-half test, the researcher randomly divides these assorted items into two groups and then composes a summary "measure" out of each of the groups.

Because all of the items are taken to be measures of the same thing, the two summary measures should tend to be the same. A measure of how close they are to each other provides a check on how reliable the total summary measure is.

As an example, consider a state-by-state measure of "welfare policy expenditures." It might be that one particular item—disaster relief, for instance—varies greatly from state to state and from one year to the next in any one state. In one year, there might be no natural disasters; in another, there might be floods or a hurricane. That particular item would be a source of unreliability in the overall measure. It also should cause the split-half test to show a relatively low reliability, for its erratic variation would make the score based on the group in which it was included less likely to equal the score based on the group that did not include it.

These two checks for reliability complement each other. The test-retest check is appropriate for any sort of measure that can be replicated. It checks for *all* sources of unreliability, but this often includes changes in the true value of the concept rather than only the instability that is due to the measurement process.

The split-half check is appropriate for measures comprising a group of subitems. It checks only for those sources of unreliability that do not operate over time, inasmuch as all of the subitems presumably are measured at the same time. Accordingly, it can miss some sources of instability in the measurement process, such as the effect of the length of time since payday or of changes in the weather, on respondents' answers to an interview question. But this is actually an important benefit: If we are able to screen out true change over time in the concept, we will have a much better idea of any instability due to the measurement process.

VALIDITY

Reliability has to do with how dependably a measure mirrors its concept. In thinking about reliability, we assumed implicitly that the measure tended to mirror the concept faithfully and that the problem of reliability was simply that this tendency may be a rather loose one. We assumed, in other words, that if the concept were measured a large number of times, the average of those measures would reflect our "ideal" concept. The problem in reliability is that since the measures vary, at any given time one of them could be rather far from the true value of the concept.

A more serious failing of our measurements, however, could result if they lack *validity*. A measure is valid if it actually measures what it purports to measure. That is, if there is in principle a *relationship of equivalence* between a measure and its concept, the measure is valid. A measure cannot be valid if it is not reliable. But it can be reliable and yet still not be valid. If it gives the same result repeatedly, the measure is reliable, but it could distort the concept in the same way each of these times, so that it does not tend to mirror it faithfully. In effect, it would be "reliably invalid."

The relationship between the measure "increases in reported arms expenditures" and the concept "increases in armaments" in Figure 4-2 is an example of invalid measurement. The relationship between the concept and the measure is such that when "increases in armaments" are high, "increases in reported arms expenditures" may be either high or low, depending on the reason for the arms buildup. The measure might be reliable (a country that reported low increases in one year, for instance, should be likely to report low increases the next year also), but it would still be invalid, because the measure does not mirror the concept accurately.

The relationship between validity and reliability can be clarified by introducing the notion of *random error* and *nonrandom error*, which will also be important when we look at measuring the strength of relationships in Chapter 8. Random error is the sort of error we have addressed in discussing reliability. If in the long run, on the average, the measures of a concept tend to be true, we can assume that any error in the measure is random. In measuring education, for example (see p. 46), we encountered a good deal of random error; people tended to report their level of education differently from one time to the next. There was no reason to expect that people were deliberately misrepresenting their educational level, however, so we would expect that in the long run, accidental reporting errors would cancel each other out and the average of many such reports would give a true measure of the concept for a particular group of people.

By contrast, nonrandom error is systematic error that tends in the long run, on the average, to distort a given measure of a concept. Thus, if we asked people whether they had a prison record, it is likely that there would be a good deal of nonrandom error in the measure, as people systematically tried to suppress their prison records. Even in the long run, on the average, this measure would not give an accurate estimate of the true value.

Quite simply, a measure is valid to the extent that it is free of both sorts of error. A measure is reliable to the extent that it is free of random error alone. Thus, reliability is a necessary but not a sufficient condition of validity.

The two sorts of error are presented visually in Figure 4-3. Think of an archer who is trying to produce "valid" shots, that is, shots at the center of a target. The shooting may suffer from either random error (the archer is erratic) or nonrandom error (the archer has some systematic problem—perhaps there is a wind from the left, and the archer has not yet learned to correct for it), or both. These two sorts of error result in the four possible combinations shown in Figure 4-3. The archer achieves "reliability" on both targets B and D but achieves "validity" only on target D.

Some Examples

A few examples are in order. There are many ways that a measure can be invalid. We have already discussed several instances of random error, so we will confine ourselves here to examples of nonrandom error.

One common source of invalid measures is extrapolation from a sample to a population that is not really represented by that sample. Using letters to the editor as an indicator of public opinion would be unwise, for instance, because the people who write such letters are not an accurate cross section of the public as a whole. Their opinions would not be a valid measure of "public opinion."

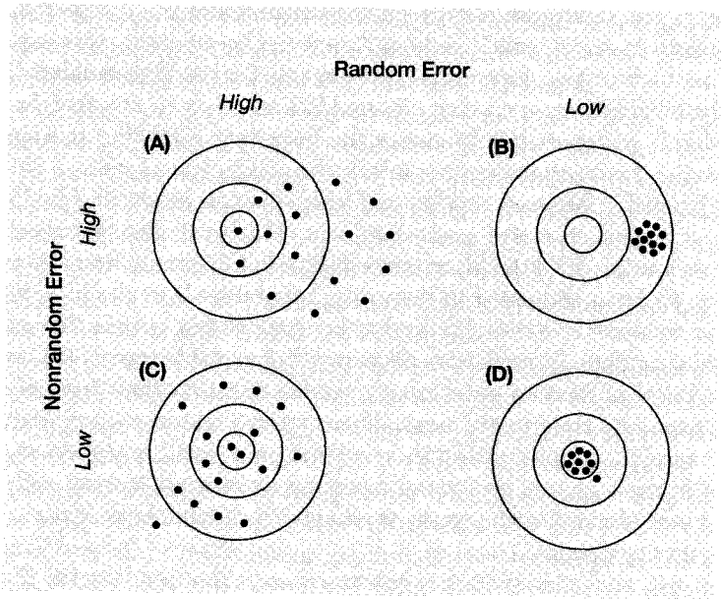

FIGURE 4-3 Random and Nonrandom Error

A comic case of sampling problems from the early days of opinion polls is the *Literary Digest* poll. The *Literary Digest* was a giant magazine in the United States in the early part of the twentieth century. Starting in 1924, the *Digest* ran an ambitious poll in presidential election years. Virtually everyone who owned a car or a telephone was reached by the poll, which was sent out to a mailing list obtained from telephone directories and state automobile registration lists. Only about 20 percent of the sample ballots mailed out were returned, but even at that, the *Digest* had over 2 million responses each time.

The *Digest* sample distorted the U.S. population in two ways. First, it essentially sampled only the upper and middle classes, inasmuch as those who did not have a car or a telephone—at a time when cars and telephones were far less universally owned than today—did not get onto the mailing list. Also, it sampled only those who were interested enough and energetic enough to return the sample ballot. Because only 20 percent of those who received the ballot returned it, this seems to have been a rather select sample.

In 1924, 1928, and again in 1932, the *Digest* poll was very successful, coming within a few percentage points of the actual outcome in each of those elections. By 1932, the poll had become an institution; it was attacked in the *Congressional Record* and featured in *New Yorker* cartoons. Thus it was a shock when the poll's prediction of a landslide victory for Landon in 1936 was gainsaid by FDR's decisive victory. When the *Literary Digest* went out of business the next year, it was thought that the shock and loss of reputation from having called the election so badly was a factor in the magazine's demise.

Apparently, the interested, middle-class sample the *Digest* drew upon did not vote much differently from the rest of the country from 1924 through 1932. Accordingly, its sympathies were a valid measure of the way the country was going to vote in those elections. Between 1932 and 1936, however, Roosevelt initiated the New Deal, which broadened his support among the poor and drove the middle class to the Republicans. After 1936, the sympathies of the middle class were no longer a valid measure of the way the country would vote.

A questionnaire item may also result in an invalid measure when respondents attach a significantly different meaning to the question than was intended by the researcher. It had always been thought, for example, that French farmers were essentially apolitical. When asked in surveys "How interested are you in politics?" they had generally responded, "Not at all." At the same time, it was striking that voting participation was higher among farmers than among most groups in the French population. If they were not interested in politics, why did they vote?

In a study designed to explore this paradox (see p. 29), Sidney Tarrow discovered that the innocent question about political interest had been spreading confusion. French farmers apparently interpreted "interest in politics" to mean commitment to some particular party, with many of them vehemently rejecting political parties. Thus, many farmers who were interested in politics but considered themselves independents responded "Not at all" to this invalid measure of "interest in politics."

Checks for Validity

Taking precautions. Our problem in checking the validity of a measure is similar to the general problem of measurement, depicted in Figure 4-1. We say that a measure is valid if it is a true measure of its concept. *But the general problem of measurement is precisely the fact that usually all we can observe is the measures.* We cannot know what the relationship between a concept and its measure is. How, then, can we assess the validity of the measure?

The answer, of course, is that there is no pat way to do so. Part of the "craft" in the craft of political research is cleverness and care in developing measures that appear likely to be valid. Some techniques are available to help in developing valid measures.

These deceptively simple strategies consist of taking various precautions against invalidity during the construction of a measure. The most important thing is to think through the measurement process carefully and to be on guard against any way in which the relationship between concept and measure might be distorted.

For example, we now know that drawing a sample in certain ways (drawing a random sample, for instance) guards against a fiasco like that which destroyed the *Literary Digest* poll. Also, in determining the final form of a questionnaire that you hope to use in a study, you should ask a few people to answer your questions and then to relate their understanding of the questions themselves. This may alert you to questions that mean something different to your respondents than they mean to you. Such

preliminary testing of one's measures is called a "pilot study." Similarly, in using official documents, you should go thoroughly into the background of the things reported there—how they were developed, what the terms mean, how broadly they are intended to apply, and so on.

These techniques simply require the investigator to think ahead to problems that could occur in the relationship between concept and measure and act either to prevent these or to check to see whether they are present. At the most general level, the strategy I have suggested here requires only that the investigator consider carefully how plausible it is that the measure mirrors the concept.

Test of validity. Thus far, our strategies have not actually provided a test of the validity of the measure. Such a test can be made, although it is necessarily subjective and open to varying interpretations. Let us say that we want to decide whether measure α is a valid measure of concept A. If there is some measure β that we are certain is strongly related to concept A, we can check to see whether measure β is related to measure α. If it is not, *and* if our assumption of a relationship between β and A is true, then α cannot be a valid measure of A.

The study by Tarrow cited earlier provides an example of this logic. Tarrow's initial conclusion that the usual question "How interested are you in politics?" was providing an invalid measure of "political interest" among French farmers came from his observation that farmers had in fact one of the highest levels of electoral participation among French citizens. Because he could not conceive of high electoral participation occurring in the absence of high political interest, he concluded that the conventional measures, which had showed low political interest coinciding with high participation, must not have been measuring political interest validly.[3]

As another example of this kind of test, consider a measure of nations' hostility to each other—based on content analysis of the nations' newspapers (α). If we found that two of the nations went to war against each other (β) yet the measure did not show an accompanying increase in feelings of hostility between the two, we would be suspicious of the validity of the measures.

Such an indirect test of validity is possible only when a researcher is quite certain that β must go along with A. That kind of certainty is uncommon and may not be shared equally by every observer. Thus, this test is not always, or even usually, possible; and it is always rather subjective. But assessing the validity of measures is so important that an indirect test, when it can be used, will greatly strengthen your findings.

The most general test of validity is what is called *face validity*. This is just a fancy term for whether a measure looks right to you. Is it valid "on its face"? After all, you have considerable experience with politics and must judge for yourself whether the measure does what you want it to do. If you think it does (and people who read your work agree with you), it has face validity.

[3]In this example, "political interest" corresponds to A, farmers' responses to the question on political interest correspond to α, and farmers' electoral participation corresponds to β.

IMPACT OF RANDOM AND NONRANDOM ERRORS

It should be obvious that the best measure is one that is fully valid, that is, one that has in it neither random nor nonrandom errors. Because social scientists often operate with measures that include some amount of one or the other sort of error, however, it is worth considering what happens under those circumstances. As it happens, the two sorts of error have different effects on the development of theory.

The effect of nonrandom error is simple and severe. If a measure is systematically invalid, there is no reason for us to expect any correspondence between the relationship we actually observe from our measures and the idealized relationship we wish to investigate.

The effect of unreliability (if the measures are otherwise valid) is more subtle. To the extent that measures are unreliable, the relationship at the measure level will tend to be looser and weaker than the true relationship. It will parallel the true relationship but will appear weaker than is actually the case. If the measures are sufficiently unreliable, the basic relationship can be so weakened that it will appear, from what we can observe, as if there is no relationship at all.

This is illustrated in Table 4-1. Each set of two columns tabulates the closeness of elections in ten congressional districts, as well as their representatives' seniority. It is apparent from the figures in the first set of columns (True Values) that there is a relationship between the two, inasmuch as representatives from safe districts tend to have greater seniority than those from marginal districts. The relationship is also quite strong; seniority increases without exception as the representatives' margins of victory increase.

TABLE 4-1 Safe Districts Related to Seniority, Using Simulated Data

True Values		Less Reliable Measures		Very Unreliable Measures	
Seniority	*Margin of Victory (%)*	*Seniority*	*Margin of Victory (%)*	*Seniority*	*Margin of Victory (%)*
32	18	32.0	21.6	12.8	9.0
24	12	19.2	12.0	12.0	12.6
23	11	11.5	12.1	30.2	8.8
20	11	22.0	4.4	22.4	6.6
14	8	14.0	5.6	21.2	12.0
11	6	11.0	6.0	13.1	3.6
10	6	9.0	10.8	2.5	0.1
6	4	3.6	4.0	13.8	2.8
5	3	4.5	3.6	21.3	1.5
2	1	2.4	0.5	0.2	12.9

In the third and fourth columns, random error, such as might occur from clerical errors or other sources of unreliability, has been added to the original measures; this makes them less reliable. In the fifth and sixth columns, an even greater degree of random error has been added. Note that the relationship becomes weaker in the second set of data (there are more exceptions to the general tendency) and virtually disappears in the last. If we were to test the relationship between safe districts and seniority and had only the data from the last two columns in hand, we would probably conclude that there was no relationship.

IMPORTANCE OF ACCURACY

Inaccuracy in measurement is a critical problem whose potential for mischief does not yet seem well enough understood by political scientists. Because the variables we work with are difficult to measure, we have in many cases come to accept measures that we know full well are inadequate. Much of survey research tends to fall into this category. The loose acceptance of cross-national indicators (for instance, using "newspapers read per 1,000 population" as a measure of the political awareness of various electorates) is another example of this problem.

To the extent that our measures are not valid, what we do with them is irrelevant. This simple fact has tended to be forgotten in a general ethos of making do with poor measures. Fortunately, political scientists are now becoming more aware of the problem of measurement, but its importance must be constantly underscored.

Let me reemphasize the important pitfalls to be careful of in measurement:

1. ***Be sure that the measures you choose fit the relationship among concepts that you wish to examine.*** Often, an interesting question is lost through mistakes in setting up empirical operations to parallel the theoretical question.

   Stephen Ansolabehere, Alan Gerber, and James M. Snyder, Jr. (2002) demonstrated this problem in a paper showing that three decades of research on whether electoral reapportionment affected public policy had been subtly misdirected and had, as a result, reached exactly the wrong conclusion. In the wake of the *Baker* v. *Carr*[4] decision, scholars had looked to see whether taking away the unfair overrepresentation of rural counties had led to an expansion of state spending. The implicit assumption was that rural areas would have preferred to keep spending down, so if they lost influence on policy, spending should have risen. When they found that spending levels after reapportionment were no higher than they had been before reapportionment, scholars concluded that changing the electoral rules had had no impact on public policy. The reason was pretty clear: Under the bad, old system both Republican suburbs and Democratic cities had been disenfranchised, while rural areas, which had benefited, were not all that clearly either Democratic or Republican. So redressing the imbalances was approximately a wash in terms of party strength.

   Ansolabehere and his coauthors pointed out, though, that the true test of whether electoral institutions affected policy was not the effect on the two parties, or on the overall spending level of the state, but whether counties that gained fair representation thereafter got a more equal share of state expenditures. In other words, the true test of whether equality of representation affected policy was not whether the level of spending rose—it was pretty obvious why it did not—but whether it was distributed more equally as a result of more equal representation. When the researchers tested the effect in this way, they found dramatic policy effects from reapportionment. Three decades of research and commentary had missed the point.

2. ***Test your measures for possible inadequacies.*** Even when a measurement problem is not so central as to nullify the results of a study, recurrent nagging inadequacies in the chosen measures may debilitate a theory so that it becomes almost a trivial exercise. Consider a test for the simple theory: "To the extent that they understand politics, if people's need for public services is relatively great, they will be more liberal." A political scientist might operationalize the three variables of this theory in a way such as the following:

   1. "Understanding of politics" indicated by *years of education*. This would seem reasonable; at least, understanding should be fairly closely related to education.
   2. "Need for public services" indicated by the *size of the person's family*. Again, although this is a rough measure, it would seem that the more dependents a person had, the more that person would depend on a variety of public services.
   3. "Liberalism" indicated by *voting Democratic*.

   Now, the empirical analog of the theory becomes: "The more educated a person is, the stronger the relationship between the size of that person's family and the probability that the person will vote Democratic." This statement is ridiculous. I have exaggerated here slightly, but only slightly, the extent to which unimaginative scholars will allow moderate errors of measurement to accumulate in a statement until the statement loses much of its meaning. The cure for this problem is simply to use care and imagination in developing measures.

In this chapter, I have discussed problems in the accuracy of measurement. These problems turn out to be of two basic types, depending on whether they stem from flux in the measure (the problem of reliability) or from a basic lack of correspondence between measure and concept (the problem of nonrandom error). In the next chapter, we look at another aspect of measurement, the question of how precisely a measure should be calibrated. In Chapter 8, we will return to tackle the problem of measuring relationships when random or nonrandom error is present.

[4]*Baker* v. *Carr* 369 US 186 (1962). In this decision, the Supreme Court ruled that it was unconstitutional to have state legislative districts with widely varying populations. Prior to the decision, many states had done no legislative redistricting for half a century, so backward rural areas that had not experienced much growth in their populations were hugely advantaged relative to rapidly growing cities and suburbs. For instance, in Florida, before the Court's decision, Jefferson County, with a population of 9,543, had had one seat in the state senate and one seat in the state house of representatives. Miami's Dade County, population 935,047, had had one seat in the state senate and three seats in the state house of representatives.

Further Discussion

A delightful book with a unique approach to handling certain problems of validity is *Unobtrusive Measures*, by Eugene J. Webb, Donald T. Campbell, Richard D. Schwartz, and Lee Sechrist (1966). The ideas presented in the book are both creative and sound, and the text itself is filled with interesting and highly unusual examples. Frederick Mosteller's article "Errors," in the *International Encyclopedia of the Social Sciences* (1968), is well worth reading. Mosteller details a variety of possible sources of invalidity and unreliability. Herbert Asher (1974) presents some good examples of reliability problems, with a rather technical discussion of ways to handle them. See also Kirk and Miller, *Reliability and Validity in Qualitative Research* (1985).

Several good examples of specific measurement problems are Niemi and Krehbiel (1984), "The Quality of Surveying Responses About Parents and the Family"; Hammond and Fraser (1984), "Studying Presidential Performance in Congress"; Feldman (1983), "The Measurement and Meaning of Trust in Government"; and Converse and Pierce (1985), "Measuring Partisanship."

A useful exercise would be to list as many factors as you can that would lead to random or nonrandom error for each of the following measures: intended vote (in surveys); strength of armed forces; agreement with the president (in congressional voting); tribalism; unemployment; hierarchical control in an agency; and personal income.

Problems of Measurement: Precision

The preceding chapter dealt with problems of the reliability and validity of measures. Those problems concerned the relationship between a measure and the concept that measure is intended to mirror. In this chapter, we deal with the "quality" of the measure itself—how precise it should be, or how finely calibrated, if it is to be useful.

In a study of Norwegian politics, Harry Eckstein felt that it was necessary to apologize for the fact that some measures he would use were subjective intangibles ("warmth in social relations," for instance, and "sense of community") rather than precise numerical quantities.

> [M]any of the indicators used in the text may not be readily recognized as such by contemporary social scientists. By an indicator we usually mean nowadays a precisely ascertainable quantity that stands for some imprecise quantity (as GNP may indicate level of economic development, or as the number of casualties in revolutionary violence may indicate its intensity). I do use such quantities in what follows. More often, however, readily observable "qualities" are used as indicators of not-so-readily observable qualities. This strikes me as both defensible and desirable, for quantitative indicators are not always as "indicative" of what one wants to know as other observations, nor always obtainable. In overemphasizing quantities we sometimes miss the most telling data—in any case, data that may be reliable in their own right or used as checks on the inferences drawn from quantitative data. I conceive of all social behavior as a vast "data bank," only some of which is quantitatively aggregated in yearbooks and the like, and much of the rest of which may speak volumes to our purposes, if used circumspectly. (Eckstein, 1966, footnote pp. 79–80)[1]

[1]From Eckstein, Harry, *Division and Cohesion in Democracy: A Study of Norway* (Copyright 1966 by Princeton University Press). Reprinted by permission of Princeton University Press.