have been collected in a structured mass survey. As we've seen, in less-structured methods such as field research, the identification and specification of relevant concepts is inseparable from the ongoing process of observation.

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

Criteria of Measurement Quality

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

Precision and Accuracy

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

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

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

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

Reliability

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

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

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

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

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

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

Here's another strategy for assessing morale. Suppose we check the company records to see how many grievances have been filed with the union during some fixed period. Presumably this would be an indicator of morale: the more grievances, the lower the morale. This measurement strategy would appear to be more reliable: Counting up the grievances over and over, we should keep arriving at the same number.

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

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

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

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

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

Test-Retest Method

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

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

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

Sacks and his colleagues report the following:

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

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

Split-Half Method

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

Using Established Measures

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

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

Reliability of Research Workers

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

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

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

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

Validity

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

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

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

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

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

Is very religious

Supports equality of men and women

Supports far-right militia groups

Is concerned about the environment

Some possible validators would be, respectively, attends church, votes for women candidates, belongs to the NRA, and belongs to the Sierra Club. Sometimes it is difficult to find behavioral criteria that can be taken to validate measures as directly as in such examples. In those instances, however, we can often approximate such criteria by applying a different test We can consider how the variable in question ought, theoretically, to relate to other variables. **Construct validity** is based on the logical relationships among variables.

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

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

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

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

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

Who Decides What's Valid?

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

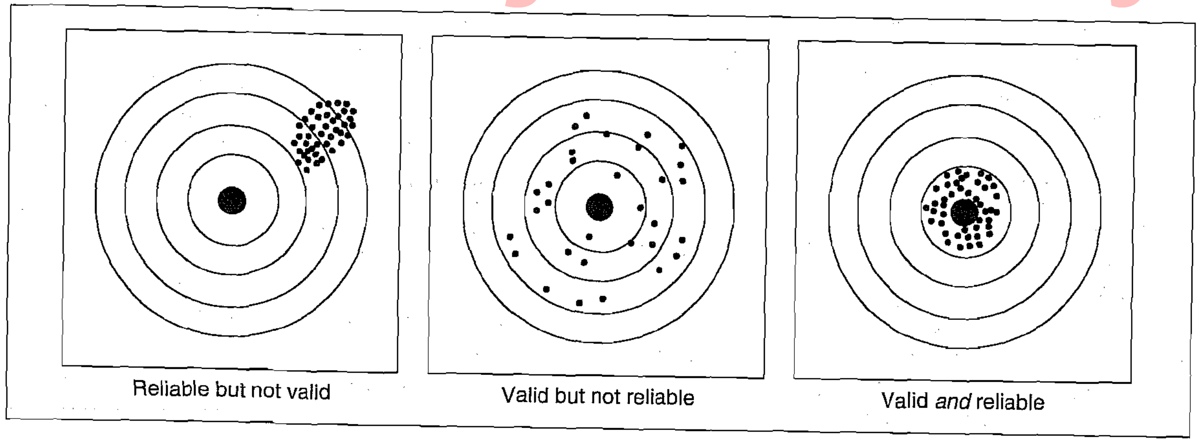
Social researchers sometimes criticize themselves and one another for implicitly assuming they are somewhat superior to those they study. For example, researchers often seek to uncover motivations that the social actors themselves are unaware of. You think you bought that new Burpo-Blasto because of its high performance and good looks, but *we* know you're really trying to achieve a higher social status.

This implicit sense of superiority would fit comfortably with a totally positivistic approach (the biologist feels superior to the frog on the lab table), but it clashes with the more humanistic and typically qualitative approach taken by many social scientists. We'll explore this issue more deeply in Chapter 10.

In seeking to understand the way ordinary people make sense of their worlds, ethnomethodologists have urged all social scientists to pay more respect to these natural social processes of conceptualization and shared meaning. At the very least, behavior that may seem irrational from the scientist's paradigm may make logical sense when viewed through the actor's paradigm.

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

FIGURE 5-2
An Analogy to Validity and Reliability



Reliable but not valid

Valid but not reliable

Valid *and* reliable

Tension between Reliability and Validity

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

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

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

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

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

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

MAIN POINTS

- Conceptions are mental images we use as summary devices for bringing together observations and experiences that seem to have something in common. We use terms or labels to reference these conceptions.

- Concepts are constructs; they represent the agreed-upon meanings we assign to terms. Our concepts don't exist in the real world, so they can't be measured directly, but it's possible to measure the things that our concepts summarize.
- Conceptualization is the process of specifying observations and measurements that give concepts definite meaning for the purposes of a research study.
- Conceptualization includes specifying the indicators of a concept and describing its dimensions. Operational definitions specify how variables relevant to a concept will be measured.
- Precise definitions are even more important in descriptive than in explanatory studies. The degree of precision needed varies with the type and purpose of a study.
- Operationalization is an extension of conceptualization that specifies the exact procedures that will be used to measure the attributes of variables.
- Operationalization involves a series of interrelated choices: specifying the range of variation that is appropriate for the purposes of a study, determining how precisely to measure variables, accounting for relevant dimensions of variables, clearly defining the attributes of variables and their relationships, and deciding on an appropriate level of measurement.
- Researchers must choose from four levels of measures that capture increasing amounts of information: nominal, ordinal, interval, and ratio. The most appropriate level depends on the purpose of the measurement.
- A given variable can sometimes be measured at different levels. When in doubt, researchers should use the highest level of measurement appropriate to that variable so they can capture the greatest amount of information.
- Operationalization begins in the design phase of a study and continues through all phases of the research project, including the analysis of data.
- Criteria of the quality of measures include precision, accuracy, reliability, and validity.
- Whereas reliability means getting consistent results from the same measure, validity refers to getting results that accurately reflect the concept being measured.
- Researchers can test or improve the reliability of measures through the test-retest method, the split-half method, the use of established measures, and the examination of work performed by research workers.
- The yardsticks for assessing a measure's validity include face validity, criterion-related validity, construct validity, and content validity.
- Creating specific, reliable measures often seems to diminish the richness of meaning our general concepts have. This problem is inevitable. The best solution is to use several different measures, tapping the different aspects of a concept.

KEY TERMS

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

REVIEW QUESTIONS AND EXERCISES

1. Pick a social science concept such as liberalism or alienation, then specify that concept so that it could be studied in a research project. Be sure to specify the indicators you'll use as well as the dimensions you wish to include in and exclude from your conceptualization.
2. Locate a research report in a book or journal article. Identify the key variable studied by the researcher(s) and describe how the variable was operationalized for measurement.
3. What level of measurement—nominal, ordinal, interval, or ratio—describes each of the following variables:
   a. Race (white, African American, Asian, and so on)
   b. Order of finish in a race (first, second, third, and so on)
   c. Number of children in families
   d. Populations of nations
   e. Attitudes toward nuclear energy (strongly approve, approve, disapprove, strongly disapprove)
   f. Region of birth (Northeast, Midwest, and so on)
   g. Political orientation (very liberal, somewhat liberal, somewhat conservative, very conservative)
4. In a newspaper or magazine, find an instance of invalid and/or unreliable measurement. Justify your choice.
5. Go to Holocaust Studies: Prejudice (http://www.socialstudies.com/c/ZeCwFuEspbb41/Pages/holo.html) and browse through the materials described there. Make a list of the various dimensions of prejudice that you find there.

ADDITIONAL READINGS

Bohrnstedt, George W. 1983. "Measurement." Pp. 70–121 in *Handbook of Survey Research*, edited by Peter H. Rossi, James D. Wright, and Andy B. Anderson. New York: Academic Press. This essay offers the logical and statistical grounding of reliability and validity in measurement.

Grimes, Michael D. 1991. *Class in Twentieth-Century American Sociology: An Analysis of Theories and Measurement Strategies*. New York: Praeger. This book provides an excellent, long-term view of conceptualization as the author examines a variety of theoretical views of social class and the measurement techniques appropriate to those theories.

Lazarsfeld, Paul F., and Morris Rosenberg, eds. 1955. *The Language of Social Research*, Section I. New York: Free Press of Glencoe. An excellent and diverse classic collection of descriptions of specific measurements in past social research. These 14 articles present useful and readable accounts of actual measurement operations performed by social researchers, as well as more conceptual discussions of measurement in general.

Miller, Delbert. 1991. *Handbook of Research Design and Social Measurement*. Newbury Park, CA: Sage. A powerful reference work. This book, especially Part 6, cites and describes a wide variety of operational measures used in earlier social research. In several cases, the questionnaire formats used are presented. Though the quality of these illustrations is uneven, they provide excellent examples of possible variations.

Silverman, David. 1993. *Interpreting Qualitative Data: Methods for Analyzing Talk, Text, and Interaction*, Chapter 7. Newbury Park, CA: Sage. This chapter deals with the issues of validity and reliability specifically in regard to qualitative research.

U.S. Department of Health and Human Services. 1992. *Survey Measurement of Drug Use. Washington, DC: Government Printing Office.* An extensive review of techniques devised and used for measuring various kinds of drug use.

SOCIOLOGY WEB SITE

See the Wadsworth Sociology Resource Center, Virtual Society, for additional links, Internet exercises by chapter, quizzes by chapter, and Microcase-related materials:

http://www.sociology.wadsworth.com

INFOTRAC COLLEGE EDITION

SEARCH WORD SUMMARY

Go to the Wadsworth Sociology Resource Center, Virtual Society, to find a list of search words for each chapter. Using the search words, go to InfoTrac College Edition, an online library of over 900 journals where you can do online research and find readings related to your studies. To aid in your search and to gain useful tips, see the Student Guide to InfoTrac College Edition on the Virtual Society Web site:

http://www.sociology.wadsworth.com