CHAPTER 3

Defining Variables

▼ **PROLOGUE** ▼

Suppose I am a sociologist who wishes to study the level of bigotry in a designated group of people. Short of asking each one, "Are you a bigot?" which is likely to be answered in the negative, I would need to come up with a series of questions, for example, which would tap into the degree and type of bias—religious, racial, ethnic, and so on—that I might encounter. I might want to use a system to score the responses in such a way as to ultimately give each respondent a bigotry score. In addition, I would want to be sure my questions are actually measuring bigotry rather than some other phenomenon. The techniques presented below will assist me in designing my study.

INTRODUCTION

In this chapter, we are dealing with the way in which we develop systems of measurement for the variables we are studying. We begin by determining how we will make a measurement and what specific criteria we will use for assigning our subjects or respondents to specific categories of each variable. Attention is given to the creation of numerical scales or indices of opinions or attitudes and how we determine the validity and reliability of such scales. Selected examples of variable measurement are also presented.

GATHERING THE DATA

Let us assume that a researcher has identified one or more hypotheses to be tested in a study. The selection of the hypothesis also generally locks a researcher into studying a particular unit of analysis. If the generalizations are about individual behavior or attitudes, we normally choose human subjects as our units of analysis. If the hypothesis refers to characteristics of cities, cities or metropolitan areas become the units of analysis, and so on. In each hypothesis, there are usually two variables to be studied, a dependent variable whose variation the researcher is trying to explain or predict and an independent variable that hypothetically accounts for the change in the dependent variable. To study these variables, we must be able to measure the characteristic or amount possessed by each of our units of analysis. Before we can make measurements, however, we must determine exactly what we want to measure and how we are going to take these measurements.

Suppose we have chosen individuals as our units of analysis and we intend to administer a questionnaire to each of our subjects. Also suppose, as is often the case, that we intend to begin our survey by asking each respondent for some basic background information. We would not want the subject's name if we wanted an anonymous survey, but we might like to know such things as the subject's age, sex, marital status, religion, ethnic background, and so on. These social characteristics may be related to many of the variables in our study. Such background information is called **demographic data.** If one of our hypotheses is that liberalism and age are inversely related, one of our variables—age—will be among the demographic data in the early part of our questionnaire. For our purposes, what do we mean by age, and how do we propose to measure it? Do we want the respondent's age at the time he or she fills out the questionnaire? Suppose the survey is to be conducted over several weeks and to several groups of people. We might select to have the respondent indicate his or her age as of some specific date; for example, "How old were you as of February 1st of this year?" Let us

assume that we are satisfied with the respondent's age at the time the survey document is filled out. Are we satisfied to know the respondent's age only in years? This is usually the case, but there are instances when we might opt for more specific information. In studying children of elementary school age, for instance, we might want the age in years plus months if we have reason to believe that, for example, a 7-year-old child may respond to certain items quite differently from a 7½-year-old child.

Demographic data Background information that gives the social characteristics of a subject.

Suppose for our study that age in years only is sufficient. How shall we get our age data? With an adult respondent, we may simply use the following format:

Age: _____ years old

The respondent just fills in the blank with the appropriate year.

Most of the time, this is adequate for social research, but imagine a situation in which we have reason to suspect that the respondent may misrepresent his or her age. We might want to obtain the age from documentation provided by the respondent, such as a birth certificate. What if someone said he was 18 years old, but his birth certificate indicates that he is only 17½ years old? The age we record depends on what we have decided in advance. If we had decided to accept whatever age the respondent gave, then this person will be listed as 18 years old. If we wanted the age as indicated on the birth certificate, we would record 17½.

Likewise, in studying voting behavior, we often find instances of people claiming they had voted in a particular election when they had not. (After all, we learn that voting is a civic duty.) In this case, the respondent's answer to the question of having voted in that election may be a far less accurate operational definition than one requiring the researcher to confirm the answer by examining public voting records.

OPERATIONAL DEFINITIONS

In making such decisions, we are formulating a **working** definition or an **operational definition.** For demographic data from an adult sample, we are usually satisfied to operationalize these concepts by accepting whatever response the subject provides. We would list the subject as

18 years old because that was what the subject said, and we assume that he or she is telling the truth. The operational definition is thus a measurement definition. It defines how we are going to measure someone or something to determine the subject's score on a variable.

Working or operational definition A definition of the way that someone or something will be measured to determine the subject's score on a variable.

The idea behind the operational definition is that once formulated and applied, there would be no disagreement as to the respondent's score or category assignment. In a particular room, some occupants might find the temperature too hot, whereas others are comfortable. Because there is disagreement among the occupants, we cannot characterize the room temperature as being either too hot or comfortable. Suppose, though, that we agree in advance to measure room temperature with a thermometer and operationally define "too hot" to be any temperature equal to or greater than 78 °F. If the thermometer reads 77 °F, we consider the room to be comfortable even though several occupants feel it to be too hot; if the thermometer reads 78 °F, we consider the room to be too hot even though several occupants consider the room to be comfortable. Thus, the operational definition, by virtue of its arbitrary specificity, eliminates for our purposes any disagreement as to whether or not the room is too hot. The disagreement comes in advance of our measurement when we decide arbitrarily that 78 °F is our cutoff point.

When we move from demographic concepts to other social or political variables, the problems of operationalization may become more difficult. All of these must be addressed before we can continue our study.

In the case of research involving human subjects, we are likely to face conflicts between attributes (what we say we are), attitudes (the way we actually feel), and behaviors (what we actually do). Suppose ideology (liberal to conservative) is our variable. We could ask the respondent for a self-assignment to an ideological attribute as follows:

Do you consider yourself to be: (please indicate)

_____ a liberal?
_____ a moderate?
_____ a conservative?

Suppose the respondent checks liberal. We then ask a series of questions designed to tap attitudes that would reflect ideology, such as attitudes

on abortion, aid to antidictatorial insurgencies in Latin America, censorship of "adult" magazines, and so on. Suppose the same respondent who said he or she was a liberal then gives consistently conservative responses to these attitude questions. Assuming that we have used attitude questions that reflect current major differences between liberals and conservatives so that our questions are valid, we obviously have a conflict between the respondent's self-assigned attribute (liberal) and that person's political attitudes (conservative). In designing our study, we need to know what will be most germane and useful to us—the attribute or the attitude.

A similar conflict between an attribute and a behavior could occur. Take a respondent who, when asked his or her political party identification, says Democrat. We then discover that in the last five elections, the same respondent consistently voted for the Republican candidate. What will be most useful for us in our study, to assign that subject by attribute (Democrat) or by behavior (Republican)? Subjects are rarely as consistent as we would like them to be, particularly when they do not perceive the topic that interests the researcher as having much direct importance in their own daily lives. Because we must live with such inconsistencies, we as researchers must make decisions about what we are trying to find out and what we will do with the information. If, for instance, our goal is to predict a respondent's vote in the next election, that person's past voting behavior is likely to be a better predictor of future voting than is the self-assigned attribute of party identification.

A related problem in forming our operational definition is that before we operationalize, we must have consensus on at least the major parts of our **conceptual** definition. The conceptual definition is the more general definition of that concept such as one would find in a textbook or dictionary. As an extreme example, note the term *democracy.* As we currently use it in the West, a democracy is a political system that governs based on a popular consent determined by free elections. By our standards, prior to reunification, West Germany was more democratic than East Germany. Yet, the formal name for East Germany was the German Democratic Republic, and at least to a Marxist ideologue, East Germany was democratic in that the representatives of the workers and peasants, through the Communist Party, controlled the government. Clearly, we have two very different views and definitions of the word *democracy.* Any operational definitions that stem from the Western concept of a democracy will be far different from the operational definitions based on Communist interpretations.

Conceptual definition A general definition of a concept such as one would find in a textbook or dictionary.

The above case is extreme. More commonly, there are agreements as to the general, conceptual definition but disagreements as to what aspects of that conceptual definition compose the essence of the concept essential to the operational definition. An example is the attempt to operationally define a concept such as *freedom.* Is a particular country free (its citizens possess freedom) and, if so, how free? Suppose we begin by looking up the dictionary definitions of *freedom* and selecting the portion of those definitions most germane to political freedom.

> *Freedom:* Possession of civil rights; immunity from arbitrary exercise of authority.[1]

There are two general parts to the definition: (1) civil rights and (2) exercise of authority. Should our operational definition be based on one of these? Which one? Or should we use both?

Suppose we decide to include possession of civil rights. What is a civil right, and which rights should we include in the operational definition? Civil rights are rights granted to an individual based on citizenship or national residency. We might begin with the "four freedoms" in the First Amendment to the U.S. Constitution:

- Freedom of religion
- Freedom of speech
- Freedom of the press
- Freedom of assembly

To this list we could add other civil rights gleaned from the U.S. Constitution's Bill of Rights:

- The right to bear arms
- Freedom from unreasonable searches and seizures
- The right to a jury trial
- Freedom from double jeopardy
- Freedom from cruel and unusual punishment

If we examine other documents such as the UN Charter or other bills of rights, we could add additional items such as the rights of certain linguistic groups to have their language used as an official national language or the rights of citizens to a specified economic standard of living.

What we include in our operational definition will reflect our individual values and levels of knowledge. Once we agree on what to include, further clarification must be undertaken to tighten our definitions. Suppose we had decided to use the four freedoms of religion, speech, press, and assembly.

We still have to clarify what these mean. In the U.S. Bill of Rights, for instance, freedom of religion really referred to the government's not making laws *establishing* a particular religion. In modern times, many nations have "established" religions, even though they are, by our definition, democracies (examine the status of the Church of England in the United Kingdom). The real issue for us to examine is not whether there are official religions in a country but whether adherents to the other religions are restricted in their freedom of worship or in other civil rights.

A second consideration is that all freedoms are limited even in the most democratic of countries. For example, your religion may believe in ritual human sacrifice, but that does not mean that the state allows you to practice that ritual. Likewise, freedom of speech is limited. Recall Justice Oliver Wendell Holmes's dictum that freedom of speech does not give one the right to shout "Fire!" in a crowded theater. We limit freedom of the press through libel laws and anti-pornography legislation. We limit freedom of assembly by requiring permits to hold public meetings. Therefore, our operational definition cannot be so tight as to disallow these kinds of limitations.

A final but crucial problem in forming operational definitions is whether there exist available data that will enable us to code each country in terms of the specific civil liberties chosen for inclusion in our operational definition. Is there any source of data available to us that would enable us to determine, say, the existence and level of freedom of assembly in each country? Economic and social statistics are available from several sources, but do they contain the information we need? In the case of our civil rights scores, we may have to rely on the opinions of experts who are asked to score each country for which they possess expertise in terms of the freedoms we have included. Some examples of operationalizing such variables will be discussed later.

INDEX AND SCALE CONSTRUCTION

For attitudinal variables, the operational definition usually is based on a subject's response to one or more questions designed to tap the variable being studied. In a previous example, we determined one's attitude toward abortion using a Likert-type response set.

Statement:	Abortion should be illegal.
Response:	Strongly Agree
	Agree
	Unsure
	Disagree
	Strongly Disagree

We could code each response as an ordinal ranking from (1) *strongly agree* to (2) *agree* and so on to (5) *strongly disagree,* thus creating a rank ordering on opposition to abortion. By simply reversing the rankings, (5) *strongly agree* to (1) *strongly disagree,* we would have a rank ordering on support for abortion rather than opposition to abortion as originally ranked.

A variation on this idea is a *ladder question.*

Image a ladder on which those most opposed to abortion stand on top run and those least opposed stand on the bottom rung. Where on the ladder would you place yourself.	⌊___⌋	1 Most opposed
	⌊___⌋	2
	⌊___⌋	3 Unsure
	⌊___⌋	4
	⌊___⌋	5 Least opposed

The respondent self-selects his or her place on the ladder, and the researcher codes that response by indicating the number (rank) of the rung chosen.

A second variation is a *feeling thermometer.* Instead of a ladder, the subject sees a picture of a thermometer ranging, for instance, from 0° to 100°. The accompanying statement asks the respondent to self-assign his or her own "temperature," with 100° most opposed, 50° unsure, and 0° least opposed.

Such questions may suffice to measure attitudes along single issues. A problem arises when what we are measuring is a compound variable made up of many differing attitudes. Suppose we want to measure an individual's *social conservatism.* While in its broadest sense, conservatism relates to mistrust of change, in the social context, we associate conservatives as taking certain positions on issues. Instead of asking the respondent to simply indicate whether he or she is conservative, we might better tap the issue by asking a series of questions, each designed to tap a separate aspect or dimension of conservatism. The issues chosen must be carefully selected to be meaningful in the current social and political context because attitudes change over time. Forty years ago, many, if not most, conservatives opposed mandatory desegregation of racially separate schools in the U.S. South. Today, few conservatives would be opposed.

Suppose we decided on five **items** (questions) that we considered good differentiators of conservatives from liberals in contemporary U.S. politics. The respondent would provide a Likert-type (*strongly agree* through *strongly disagree*) response to each of the following items.

1. Abortion should be illegal.
2. "Family Values" should be taught in schools.
3. Full funding for the Defense Department is needed for national security.

4. Educational and welfare issues should be primarily handled by the states or localities, not by the federal government.
5. The controlling or outlawing of handguns by the government is wrong.

As worded above, we would expect a very conservative individual to respond "strongly agree" to most of the five items, whereas a strongly liberal, least conservative individual would "strongly disagree" with most of the items. Thus, we could assign points for each item's response and then sum the points for each item, giving us an **index** of social conservatism.

Items The various components (e.g., abortion, family values, etc.) used to generate a scale or index.

Index A range of scores measuring some phenomenon. In this example, the higher one's score, the more politically conservative he or she is.

Suppose we score each question's response as follows:

Strongly Agree	20 points
Agree	15 points
Unsure	10 points
Disagree	5 points
Strongly Disagree	0 points

If a respondent gave the most conservative response, *strongly agree,* to each of the five questions, that respondent would score 100 ($20 \times 5 = 100$) on the index. The person strongly disagreeing with each item would receive a total score of zero ($0 \times 5 = 0$) on the index. One who is completely unsure, assumed to be in the middle on all five items, would score 50 ($10 \times 5 = 50$).

These scores on our social conservatism index (or social conservatism *scale*) are treated, for purposes of data manipulation, as an *interval* level of measurement.

We would do several other things to refine our index before using it for actual research purposes. First, we initially set up our questions so that the most conservative response to each item was *strongly agree,* but in doing so we may have introduced bias into the response set. After several questions, the conservative respondent might automatically answer *strongly agree* or *agree* without carefully reading the question. To avoid this, we "reverse" some of the questions so that at times the most conservative response

would be *strongly disagree* instead of *strongly agree.* In such instances, the most conservative response will still receive 20 points, even though it was *strongly disagree* rather than *strongly agree.* In the following example, we reword two of the items, show the possible responses, and indicate (in parentheses) the number of points we will assign. In the actual questionnaire, the number of points for each response should not appear in print, but the ones coding the scores later on would use the point values to determine the final index score for each subject.

Directions: Circle the response to each of the following questions that most closely reflects your own opinion.

1. Abortions should continue to be legal.

Strongly Agree	*Agree*	*Unsure*	*Disagree*	*Strongly Disagree*
(0)	(5)	(10)	(15)	(20)

2. "Family Values" should be taught in schools.

Strongly Agree	*Agree*	*Unsure*	*Disagree*	*Strongly Disagree*
(20)	(15)	(10)	(5)	(0)

3. Full funding for the Defense Department is needed for national security.

Strongly Agree	*Agree*	*Unsure*	*Disagree*	*Strongly Disagree*
(20)	(15)	(10)	(5)	(0)

4. Educational and welfare issues should be primarily handled by the federal government, not the states.

Strongly Agree	*Agree*	*Unsure*	*Disagree*	*Strongly Disagree*
(0)	(5)	(10)	(15)	(20)

5. The controlling or outlawing of handguns by the government is wrong.

Strongly Agree	*Agree*	*Unsure*	*Disagree*	*Strongly Disagree*
(20)	(15)	(10)	(5)	(0)

The very conservative respondent will answer *strongly agree* to items 2, 3, and 5 (for a total of 60 points) and answer *strongly disagree* to items 1 and 4 (for an additional 40 points). The grand total will still be 100 points for that individual.

BOX 3.1
Interval-Level Scores From Ordinal-Level Data

These scores on our social conservatism index (or social conservatism scale) are treated as an interval level of measurement for purposes of data manipulation, even though the Likert response set for each item is really only ordinal. We arbitrarily assigned the point spread and arbitrarily assumed that the difference between each adjacent response would be worth 5 points (*strongly agree:* 20 points – *agree:* 15 points equals a differential of 5 points). We have no evidence to verify that these points reflect the true amount of difference between the two responses. We did violate some mathematical assumptions in creating an interval level of measurement index out of ordinal components, but as previously indicated, this is common practice in the social and behavioral sciences. While our index was developed from only five questions, most such indices contain many more items than five. The more items we add, the more possible options of opinion we add to our index, and the closer our index gets to being truly interval-level data.

VALIDITY

Once our questionnaire is reordered, we would *pretest* it on a group of subjects, administering it once and possibly readministering it to the same group several weeks later. During this pretesting phase, we would be seeking to refine the scale by determining two things—the validity and reliability of our questionnaire as a measurement device.

Validity is the extent to which the concept one wishes to measure is actually being measured by a particular scale or index. Does the scale measure the concept it claims to measure? Is it congruent to the generally accepted definitions of the concept? For instance, if occupational income alone is being used as a measure of poverty, those with low incomes will be considered to be poor. In most instances, the measure is valid, but what about the millionaire who does not need to work and therefore has no income? This individual is not poor by anyone's definition. Thus, work-related income is not necessarily a valid index of poverty.

Validity The extent to which the concept one wishes to measure is actually being measured by a particular scale or index.

There are several strategies for determining a measure's validity. The first two—face validity and content validity—rely on the internal logic of the measure. **Face validity** is the extent to which the measure is subjectively viewed by knowledgeable individuals as covering the concept. For instance, my conservatism scale developed earlier in this chapter seems valid to me. Each of the five items seems to tap a relevant distinction between more and less conservative people. If I showed the scale to others with knowledge of the subject matter and they confirmed that each item measured conservatism, I could say that the measure had face validity. If there was controversy about some item, say, the abortion question, I would have to ask if in reality the abortion stand was a valid aspect of conservatism.

Face validity The extent to which the measure is subjectively viewed by knowledgeable individuals as covering the concept.

Content validity is related to face validity, being based on logic and expertise. It asks whether the measure covers all the generally accepted meanings of the concept. What if I showed the conservatism index to my judges, and they responded that each item had face validity but that the scale was incomplete? Several of my experts say, "What about communism? How can you measure conservatism without asking the respondent about communism?" If we concur that this item must be included in the scale to give it content validity, then I would need to add a statement such as this: "Worldwide aid for anticommunist insurgents should be increased."

Content validity The extent to which the measure covers all the generally accepted meanings of the concept.

Two other types of validity are less subjective and more empirical. They are known as criterion validity and construct validity.

Criterion validity is based on our measure's ability to predict some criterion external to it. The criterion could be in the present and currently predictable (*concurrent validity*), or it could be in the future (*predictive validity*). For instance, suppose we have designed a scale for determining whether an individual would be good in a management position with a firm. We can look at those who later became managers and compare their performance evaluations with their scale scores. If the index has criterion

validity, those scoring high on it would also be expected to perform well as managers. If some aptitude test claims to measure mathematical aptitude, we would expect those receiving high scores to also earn higher grades in math. If the opposite situation should result, high scores and low grades, or if those with both high and low aptitude scores performed equally well in class, then the aptitude test would be a poor predictor of performance and would lack criterion validity.

Criterion validity The extent to which the measure is able to predict some criterion external to it.

Construct validity has to do with the ability of the scale to measure variables that are theoretically related to the variable that the scale purports to measure.

Construct validity The ability of the scale to measure variables that are theoretically related to the variable that the scale purports to measure.

Imagine that you have developed a scale to measure overall life satisfaction. The higher the score on the index, the greater is the person's life satisfaction. To establish construct validity for the scale, ask what characteristics are likely to be related to overall life satisfaction. For instance, a satisfied individual would be less likely to be a heavy drinker or a spouse or child abuser. Is this the case with those scoring high on your life satisfaction scale? If these or other theoretical attributes are associated with life satisfaction, we should be able to empirically test the relationship between one's score on your scale and alcohol consumption or incidents of abuse. If these associations are indeed found to be the case, then your measure is likely to be a valid index of life satisfaction. You have established construct validity.

Both the criterion validity and the construct validity may be measured using techniques similar to the association and correlation measures presented in later chapters of this text.

RELIABILITY

For a measure to be **reliable,** it must be free of measurement errors. That is, (a) the overall score should correspond to the scores of its

components, a type of internal consistency, and (b) if the measure is taken over intervals of time, the scores of individuals should remain consistent over time as well. Think of an observed score as differing from the true score due to errors in measurement. That is, the observed score equals the true score plus or minus some measurement error. Ideally, true reliability is attained when measurement error is eliminated. More realistically, a score is reliable when we have minimized the impact of measurement error as much as possible.

Reliability The likelihood that the scale is actually measuring what it supposed to measure.

Split-half reliability is one way to measure internal consistency. To see if the items are all measuring the same concept, we split our overall scale into two scales, each containing half the original items. Suppose our original index was a 20-item scale designed to predict whether a teenager was prone to juvenile delinquency. We break the 20-item scale into two 10-item scales either by putting the odd-numbered items in one group and the even-numbered items in the other or by assigning 10 of the items at random to one group and putting the remaining 10 items in the second. Then we compare scores by subscale. A person appearing prone to delinquency on one subscale should also appear prone to delinquency on the other. If this is the case, we may assume that the original 20-item scale is reliable in terms of internal consistency.

Split-half reliability A measure of internal consistency that splits an overall scale into two scales, each containing half the original items.

The second kind of reliability is **test-retest reliability,** also known as reliability over time. Reliability in this context has to do with an individual's consistency in responding the same way to a specific item over time. Suppose we were to administer the conservatism questionnaire twice to the same group of people and compare each set of responses. If the responses remain about the same over time, that scale is considered reliable. If responses change, the scale may not be reliable. The cause for unreliability may lie in the fact that one or more questions were vague or confusingly worded. As a result, the reader's interpretation at the second reading may have differed from the initial interpretation of the same item. For example,

suppose the question on the conservatism measure about regulating handguns showed that many people opposed handgun legislation the first time they filled out the questionnaire but showed changes in their response to support legislation the second time they filled it out. Or the responses may have changed from favoring to supporting the legislation. Under normal circumstances, we would consider the item unreliable and delete it from the final questionnaire, concluding that gun control attitude is not a consistent and reliable indicator of political conservatism.

Test-retest reliability A measure that determines an individual's consistency in responding the same way to a specific item over time.

Before concluding unreliability, be sure that no intervening event occurred between the first and second administration of the questionnaire that would cause a consistent one-directional shift of opinion. For instance, every time there is an attempted or successful assassination of a popular public figure, attitudes favoring gun control legislation increase. In such a case, the consistency of the response changes suggests that the item may still be a reliable indicator of social conservatism.

To be a good measure, a scale or index must be both valid and reliable. This is often a difficult order to fill given the fact that many social science concepts are difficult to define and, once defined, are subject to measurement and other human error. In addition, as Babbie (1989)[2] has pointed out, there is a certain tension between validity and reliability. Validity seeks to be inclusive, extending a measure to cover all of the meanings and nuances of the concept in question. Reliability tends to exclude nuances and multiple aspects of a variable so as to focus on what can be specifically scored. One solution is to create several measures for the same concept and see if they produce similar results.

CONCLUSION

Understanding the nature of operational definitions and formulating actual operational definitions are among the hardest tasks for students to master. While in many courses, our task is to *broaden* the scope of a definition to include more and more nuances and examples, here our task is to *narrow* that definition, making it ever more specific. In many ways, this task parallels the formulation of specific legal definitions. When members of a jury determine facts and thus the guilt or innocence of the accused, they base

their determination in part on the judge's instructions, which include the legal definitions of the charged crimes. For example: What constitutes murder? How do first-degree murder, second-degree murder, and manslaughter differ under current state law? What facts must be proven for a jury to conclude a guilty verdict? The legal definitions given to the jury are as specific as possible, and the jury members then determine if the facts presented to them match those required by the definitions.

An expert coder, like a juror, can take the researcher's operational definition and conclude on a case-by-case basis whether the definition is met. For example, based on the operational definition supplied, the coder can determine whether or not country x is economically developed. But suppose we have no coder. How do we then decide whether country x is an economically developed country? One way is to make use of one or more variables as stand-ins for, or indicators of, economic development. We would pick specific variables, each of which may tap only part of the concept of economic development, such as percentage of the population in agriculture, radios per 1,000 population, and so on. Finding such data and selecting valid and reliable indicators of the concept we want to measure are often not easy.

Furthermore, an improper operational definition will lead to improper statistical results because the statistics will only be as good as the data. In popular terminology, this is the GIGO principle: "Garbage in; garbage out." Following are four general situations that lead to misleading operational definitions.

1. Ideological Assumptions. An aspect of the operational definition may be a debatable ideological assumption. For instance, there is an organization that rates each country on its adherence to principles of human rights, particularly its treatment of prisoners. In its rating system, capital punishment is considered to be an indicator of *reduced* human rights. Several countries, including the United States, get reduced ratings because they have capital punishment.

2. Situational Factors. Situations specific to the subject lead to misleading conclusions. For example, a researcher studying levels of freedom in various countries gives a certain country a low score because it is practicing censorship. The researcher does not take into account the fact that the country was at war at the time of the study. Thus, censorship of militarily sensitive subjects had been instituted, whereas in peacetime there would have been

no censorship. Another example would be a recent immigrant with a low IQ score. The reason for the score being low was that the IQ test administered was not in that person's native language. Thus, it was the testing situation, not the person's intelligence, that led to the low score.

3. Key Word Inconsistency. Respondents identify incorrectly with popular terms. For example, students were asked to assign themselves to one of three categories: liberal, moderate, or conservative. Later, they responded to items dealing with policy issues normally thought to differentiate liberals from conservatives. Several who had identified themselves as conservatives responded to the specific policy items with clearly liberal preferences.

4. Poor Predictability. The operational definition has a poor track record in predicting what it claims to predict. For instance, many high school students in the United States take standardized aptitude tests to determine their probable performance in college. These test scores are often used as criteria for college admission and for qualification for varsity sports. Yet actual studies of the relationship between aptitude test scores and first-year college grade point averages (GPAs) show that only about 6% of the variation in grade point average can be accounted for by such aptitude test scores. (It has been argued, though, that this is because not all who take the tests actually attend college—only those who score high enough. That may have been the case in the past, but today almost everyone can get admitted to some 2- or 4-year college in the United States. It would be interesting to see if the predictability of GPAs from such scores is now going up.)

In all of the above examples, weaknesses in the operational definitions could lead to misleading statistical results. One must guard against such pitfalls. The ancient Greek dictum of "Know thyself!" could well be expanded to say, "Know thyself . . . and thy subject!"

EXERCISES

Exercise 3.1

Assume that you are developing a written questionnaire. Develop questions and categories of response (or scoring instructions) that together form the operational definitions of the following concepts:

1. Age
2. Religion
3. Marital status
4. Party identity
5. Attitude on environmental problems
6. Attitude on rights of homosexuals
7. Attitude on compulsory national service (military or nonmilitary)
8. Attitude on tolerance toward racial, religious, or linguistic minorities
9. Attitude on tolerance of sexually related publications
10. Attitude on tolerance of cigarette smoking by others

Exercise 3.2

Assume that you are developing indices in which countries are the units of analysis. What factors would you consider in developing scales for each of the following? How might you weight these factors?

1. Political tolerance
2. Harshness of criminal penalties
3. Freedom of religion
4. Disability awareness
5. Public safety
6. Public health

NOTES

1. As defined in *The American Heritage Dictionary of the English Language,* New College Edition (Boston: Houghton Mifflin, 1981), p. 524.

2. Earl Babbie, *The Practice of Social Research,* 5th ed. (Belmont, CA: Wadsworth, 1989), pp. 125–6.

▼ KEY CONCEPTS ▼

central tendency
mean/arithmetic mean
summation symbol (capital sigma: $\sum$)
$\bar{y}$ and so on
median (*Md*.)
median position (*Md*. *Pos*.)
array
cumulative frequency (*cf*)
mode
modal class/modal category
x-axis
f-axis
origin (of a graph)
frequency polygon
histogram
smooth curve
continuous variable
modality
unimodal, bimodal, and trimodal frequency distributions
skewness
symmetric frequency distribution
positively skewed/skewed to the right
negatively skewed/skewed to the left
stem and leaf displays
boxplots/box and whisker plots
fractiles
quartiles
deciles
percentiles