journal and find examples of at least three different units of analysis. Identify each unit of analysis, and present a quotation from the journal in which that unit of analysis is reported.

Additional Readings

Bart, Pauline, and Frankel, Linda, ***The Student Sociologist's Handbook*** (Morristown, N.J.: General Learning Press, 1976). A handy little reference book to assist you in getting started on a research project. Written from the standpoint of a student term paper, this volume gives a particularly good guide to the periodical literature of the social sciences that's waiting for you in your campus library.

Hammond, Phillip (ed.), ***Sociologist at Work*** (New York: Basic Books, 1964). A collection of candid research biographies written by several eminent social science researchers, discussing the studies that made them eminent. A variety of research motivations and designs are illustrated in these honest reports of how the research actually came about and unfolded. Take two chapters every four hours to relieve the discomfort of believing that social science research is routine and dull.

Hunt, Morton, ***Profiles of Social Research: The Scientific Study of Human Interactions*** (New York: Basic Books, 1985). An engaging and informative series of project biographies: James Coleman's study of segregated schools is presented, as well as several other major projects that illustrate the elements of social research in actual practice.

Miller, Delbert, ***Handbook of Research Design and Social Measurement*** (New York: Longman, 1983). A useful reference book for introducing or reviewing numerous issues involved in design and measurement. In addition, the book contains a wealth of practical information relating to foundations, journals, and professional associations.

Stouffer, Samuel, ***Social Research to Test Ideas*** (New York: Free Press of Glencoe, 1962). A stimulating and downright inspirational posthumous collection of research articles by one of the giants of social research. In these reports, you will see how an ingenious man formulates an idea, designs the perfect study for testing it, is prevented from conducting the study, and then devises another feasible method for testing the same idea. Especially enlightening are Paul Lazarsfeld's introduction and Chapter 6, in which Stouffer reports on the effects of the Great Depression on the family.

Answers to Units of Analysis

Exercise (pages 95–96)

1. individuals (men and women, black and white people)
2. groups (incorporated U.S. cities)
3. groups (TM organizations)
4. groups (nursing staffs)
5. groups (establishments)
6. individuals (women and men farmers)
7. groups (neighborhoods)
8. individuals (black-Americans)
9. organizations (service and production organizations)
10. artifacts (job titles)

5 Conceptualization and Measurement

What You'll Learn in This Chapter

In this chapter, you'll discover that most of the words used in everyday language communicate vague, unspecified meanings. In science, it's essential to specify exactly what we mean (and don't mean) by the terms we use.

Introduction

This chapter is the first of three dealing with the process of moving from vague ideas about what you want to study to being able to recognize it and measure it in the real world. In this chapter we deal with the general issue of *conceptualization*, which sets up a foundation for the discussions of *operationalization* in Chapter 6. The issues raised in Chapters 5 and 6 will be concluded in Chapter 7, which deals with more complex types of measurements.

I want to begin the chapter with a frontal attack on the hidden concern people sometimes have about whether it's possible to measure the stuff of life: love, hate, prejudice, radicalism, alienation, and things like that. The answer is yes, but it will take a few pages for me to make that point. Once you see that we can measure anything that exists, we'll turn to the steps involved in doing that.

Measuring Anything that Exists

It seems altogether possible to me that you may have some reservations about the ability of science to measure the really important aspects of human social existence. You may have read research reports dealing with something like liberalism or religion or prejudice, and you may have been dissatisfied with the way the researchers measured whatever they were studying. You may have felt they were too superficial, that they missed the aspects that really matter most. Maybe they measured *religiosity* as the number of times a person went to church, or maybe they measured *liberalism* by how people voted in a single election. Your dissatisfaction would surely have been increased if you found yourself being misclassified by the measurement system. People often have that experience.

Or, you may have looked up the definition of a word like *compassionate* in the dictionary and found the definition wanting. You may have heard yourself muttering, "There's more to it than that." In fact, whenever you look up the definition of something you already understand well, you can probably see ways people might misunderstand the term if they had only that definition to go on.

Earlier in this book, I said that one of the two pillars of science is *observation*. Because this word can suggest a rather casual, passive activity, scientists often use *measurement* instead, meaning careful, deliberate observations of the real world for the purpose of describing objects and events in terms of the attributes composing a variable. If the variable under study were *political party affiliation*, we might consult the list of registered voters to note whether the people we were studying were registered as Democrats or Republicans. In this fashion, we would have measured their political party affiliation.

Although measurement would seem to present a special problem for social science, this section of the chapter makes the point that ***we can measure anything that exists.*** There are no exceptions. If it exists, we can measure it.

How Do You Know?

To demonstrate to you that social scientists can measure anything that exists, I'd like you to imagine that we are discussing the matter. I'll write the script, but feel free to make substitutions for your side of the dialogue as you see fit.

ME: Social scientists can measure anything that exists.

YOU: Hah! Betcha can't.

ME: Tell me something that exists, and I'll tell you how to measure it.

YOU: Okay, let's see you measure prejudice.

ME: Good choice. Now, I'm not willing to waste our time trying to measure something that doesn't exist. So, tell me if it exists.

YOU: Yes, of course it exists. Everybody knows that.

ME: How do you *know* prejudice exists?

YOU: *Everybody* knows that.

ME: Everybody used to think the world was flat, too. I want to know how *you* know prejudice really exists.

YOU: I've seen it in action.

ME: What have you seen that proves prejudice exists?

YOU: Well, a businessman told me that he'd never hire a woman for an executive position because he thought all women were flighty and irrational. How's that?

ME: Great! That sounds like prejudice to me, so I guess we can assume prejudice exists. I am now prepared to measure prejudice. Ready?

YOU: Ready.

ME: You and I will circulate quietly through the business community, talking to businessmen about hiring. Whenever a businessman tells us that he would never hire a woman for an executive position because he thinks all women are flighty and irrational, we'll count that as a case of prejudice. Whenever we are not told that, we'll count the conversation as a case of nonprejudice. When we finish, we'll be able to classify all the businessmen we've talked to as either prejudiced or nonprejudiced.

YOU: Wait a minute! That's not a very good measure of prejudice. We're going to miss a lot of prejudice that way. All we'll measure is blatant prejudice against women in hiring.

ME: I see what you mean. But your comment also means that the situation you described before proves only that blatant prejudice against women in hiring exists. We'd better reconsider whether *prejudice* exists. Does it?

YOU: Of course it does. I was just giving you one example of prejudice. There are hundreds of other examples.

ME: Give me one that proves prejudice exists.

YOU: Okay, try this for size. I was in a bar the other night, and two guys—one white and one black—were arguing about politics. Finally, the white guy got so angry, he started using ugly racist language and yelled, "All you people ought to be sent back to Africa." Is that prejudiced enough for you?

ME: Suits me. That would seem to prove that prejudice exists, so I'm ready again to measure prejudice. This will be more fun. You and I will split up and start touring bars every night. We'll keep our ears open and listen for a white person using ugly racial epithets and saying, "All you people—"

YOU: Hold it! I see where this is headed, and that's not going to do it either. A person who said that would be prejudiced, but we're going to classify a lot of prejudiced people as nonprejudiced just because they don't happen to get carried away and talk dirty.

ME: All of which brings me back to my original question. Does prejudice really exist, or have you been just stringing me along?

YOU: Yes, it exists!

ME: Well, I'm not sure any longer. You persuaded me that businessmen who discriminate against women in hiring exist, because you saw that, and I believe you. You persuaded me that there are people who call black people ugly names and say they should all go back to Africa. But I'm not so sure *prejudice* exists. I'd sure like to track it down so I can show you that I can measure it. To be honest, though, I'm beginning to doubt that it really exists. I mean, have you ever seen a prejudice? What color are they? How much do they weigh? Where are they located?

you: What on earth are you talking about?

The point of this dialogue, as you may have guessed, is to demonstrate that *prejudice doesn't exist*. We don't know what a prejudice looks like, how big it is, or what color. None of us has ever touched a prejudice or ridden in one. But we do talk a lot about prejudice. Here's how that came about.

As you and I wandered down the road of life, we observed a lot of things and knew they were real through our observations. We heard about a lot of other things that other people said they observed, and those other things seemed to have existed. Someone reported seeing a lynching and described the whole thing in great detail.

With additional experience, we noticed something more. We noticed that people who participate in lynchings are also quite likely to call black people ugly names. A lot of them, moreover, seemed to want women to "stay in their place." Eventually, we began to get the feeling that there was a certain kind of person running around the world that had those several tendencies. When we discussed the people we'd met, it was sometimes appropriate to identify someone in terms of those tendencies. We used to say a person was "one of those who participate in lynchings, call black people ugly names, and wouldn't hire a woman for an executive position." After a while, however, it got pretty clumsy to say all of that, and you had a bright idea: "Let's use the word *prejudiced* as a shorthand notation for people like that. We can use the term even if they don't do all those things – as long as they're pretty much like that."

Being basically agreeable and interested in efficiency, I agreed to go along with the system. That's where *prejudice* came from. It never really existed. We never saw it. We just made it up as a shortcut for talking behind people's backs. Ultimately, *prejudice* is merely a *term* we have agreed to use in communication: a name we use to represent a whole collection of apparently related phenomena that we've each observed in the course of life. Each of us developed his or her own mental image of what the set of real phenomena we've observed represent in general and what they have in common.

When I say the word *prejudice*, I know it evokes a mental image in your mind, just as it evokes a mental image for me. It's as though we have file drawers in our minds containing thousands of sheets of paper, and each sheet of paper has a label in the upper right-hand corner. One sheet of paper in your file drawer has the term *prejudice* on it, and I have one too. On your sheet are all the things you were told about prejudice and everything you've observed that seemed to be an example of it. My sheet has what I was told about *prejudice* plus all the things I've observed that seemed to be examples of it.

Conceptions and Concepts

The technical term for those mental images, those sheets of paper in our mental file drawers, is *conception*. Each sheet of paper is a conception. Now, those mental images cannot be communicated directly. There is no way I can directly reveal to you what's written on mine. So we use the *terms* written in the upper right-hand corner as a way of communicating about our conceptions and the things we observe that are related to those conceptions.

Let's suppose that I'm going to meet someone named Pat whom you already know. I ask you what Pat is like. Now suppose that you have seen Pat help lost children find their parents and put a tiny bird back in its nest. Pat got you to take turkeys to poor families on Thanksgiving and to visit a children's hospital on Christmas. You've seen Pat weep in a movie about a mother overcoming adversities to save and protect her child. As you search through your mental file drawer, you may find all or most of those phenomena recorded on a single sheet labeled *compassionate* in the upper right-hand corner. You look over the other entries on the page, and you find they seem to provide an accurate description of Pat. So, you say, "Pat is compassionate."

Now I leaf through my own mental file drawer until I find a sheet marked *compassionate*. I then look over the things written on my sheet, and say, "Oh, that's nice." I now feel I know what Pat is like, but my expectations in that regard reflect the entries on *my* file sheet, not yours. Later, when I meet Pat, I may find that my own, personal experiences correspond to the entries I have on my *compassionate* file sheet, and I'll say you were sure right. Or, my observations of Pat may contradict the things I have on my file sheet, and I'll tell you that I don't think Pat is very compassionate. If the latter happens, we may begin to compare notes.

You say, "I once saw Pat weep in a movie about a mother overcoming adversity to save and protect her child." I look at my *compassionate* sheet and can't find anything like that. Looking elsewhere in my file, I locate that sort of phenomenon on a sheet labeled *sentimental*. I retort, "That's not compassion. That's just sentimentality."

To further strengthen my case, I tell you that I saw Pat refuse to give money to an organization dedicated to saving the whales from extinction. "That represents a lack of compassion," I argue. You search through your files and find saving the whales on a sheet marked *environmental activism*, and you say so. Eventually, we set about comparing the entries we have on our respective sheets labeled *compassionate*. We may discover that we have quite different mental images represented by that term.

In the big picture, language and communication only work to the extent that you and I have considerable overlap in the kinds of entries we have on our corresponding mental file sheets. The similarities we have on those sheets represent the agreements existing in the society we both occupy. When we were growing up, we were both told approximately the same thing when we were first introduced to a particular term. Dictionaries formalize the agreements our society has about such terms. Each of us, then, shapes his or her mental images to correspond with those agreements, but because all of us have different experiences and observations, no two people end up with exactly the same set of entries on any sheet in their file systems.

Returning to the assertion made at the outset of this chapter, we *can* measure anything that is real. We can measure, for example, whether Pat actually puts the little bird back in its nest, visits the hospital on Christmas, weeps at the movie, or refuses to contribute to saving the whales. All of those things exist, so we can measure them. But is Pat really compassionate? We can't answer that question, we can't measure compassion in that sense, because compassion doesn't exist the way those things I just described exist.

Compassion as a *term* does exist. We can measure the number of letters it contains and agree that there are ten. We can agree that it has three syllables and that it begins with the letter C. In short, we can measure those aspects of it that are real.

Some aspects of our conceptions are real also. Whether you *have* a mental image associated with the term *compassion* is real. When an elementary school teacher asks a class how many know what *compassion* means, those who raise their hands can be counted. The presence of particular entries on the sheets bearing a given label is also real, and that can be measured. We could measure how many people do or do not associate giving money to save the whales with their conception of compassion. About the only thing we cannot measure is what compassion really means, because compassion isn't real. Compassion exists only in the form of the agreements we have about how to use the term in communicating about things that are real.

In this context, Abraham Kaplan (1964) distinguishes three classes of things that scientists measure. The first class is *direct observables*: those things we can observe rather simply and directly, like the color of an apple or the check mark made in a questionnaire. *Indirect observables* require "relatively more subtle, complex, or indirect observations" (1964:55). We note a person's check mark beside *female* in a questionnaire and have indirectly observed that person's sex. History books or minutes of corporate board meetings provide indirect observations of past social actions. Finally, *constructs* are theoretical creations based on observations but which cannot be observed directly or indirectly. IQ is a good example. It is constructed mathematically from observations of the answers given to a large number of questions on an IQ test. The other composite measures we'll discuss in Chapter 7 are other examples of constructs.

Kaplan (1964:49) defines *concept* as a "family of conceptions." A concept is, as Kaplan notes, a construct. The concept of *compassion*, then, is a construct created from your conception of it, my conception of it, and the conceptions of all those who have ever used the term. It cannot be observed directly or indirectly, because it doesn't exist. We made it up.

Conceptualization

Day-to-day communication usually occurs through a system of vague and general agreements about the use of terms. Usually, people do not understand exactly what we wish to communicate, but they get the general drift of our meaning. Although you and I do not agree completely about the use of the term *compassionate*, I'm probably safe in assuming that Pat won't pull the wings off flies. A wide range of misunderstandings and conflict—from the interpersonal to the international—is the price we pay for our imprecision, but somehow we muddle through. Science, however, aims at more than muddling, and it cannot operate in a context of such imprecision.

Conceptualization is the process through which we specify precisely what we will mean when we use particular terms. Suppose we want to find out, for example, whether women are more compassionate than men. I suspect most of us assume that is the case, but it might be interesting to find out if it's really so. We can't meaningfully study the question, let alone agree on the answer, without some precise working agreements about the meaning of the term. They are working agreements in the sense that they allow us to work on the question. We don't need to agree or even pretend to agree that a particular specification might be worth using.

Indicators and Dimensions

The end product of this conceptualization process is the specification of a set of *indicators* of what we have in mind, indicating the presence or absence of the concept we are studying. Thus, we may agree to use visiting children's hospitals at Christmas as an indicator of compassion. Putting little birds back in their nests may be agreed on as another indicator, and so forth. If the unit of analysis for our study were the individual person, we could then observe the presence or absence of each indicator for each person under study. Going beyond that, we could add up the number of indicators of compassion observed for each individual. We might agree on ten specific indicators, for example, and find six present in our study of Pat, three for John, nine for Mary, and so forth.

Returning to our original question, we might calculate that the women we studied had an average of 6.5 indicators of compassion, and the men studied had an average of 3.2. We might therefore conclude on the basis of that group difference that women are, on the whole, more compassionate than men. Usually, it's not that simple.

Very often, when we take our concepts seriously and set about specifying what we mean by them, we discover disagreements and inconsistencies. Not only do you and I disagree, but each of us is likely to find a good deal of muddiness within our own individual mental images. If you take a moment to look at what *you* mean by compassion, you'll probably find that your image contains several *kinds* of compassion. The entries on your file sheet can be combined into groups and subgroups, and you'll even find several different strategies for making the combinations. For example, you might group the entries into feelings and actions.

The technical term for such groupings is dimension: a specifiable aspect or facet of a concept. Thus, we might speak of the "feeling dimension" of compassion and the "action dimension" of compassion. In a different grouping scheme, we might distinguish "compassion for humans" from "compassion for animals." Or, compassion might center on helping people be and have what we want for them or what *they* want for themselves. Still differently, we might distinguish "compassion as forgiveness" from "compassion as pity."

Thus, it would be possible for us to subdivide the concept of compassion according to several sets of dimensions. Specifying dimensions and identifying the various indicators for each of those dimensions are both parts of conceptualization.

Specifying the different dimensions of a concept often paves the way for a more sophisticated understanding of what we are studying. We might observe, for example, that women are more compassionate in terms of feelings, and men are more compassionate in terms of actions—or vice versa. Noting that this was the case, we would not be able to say whether men or women are really more compassionate. Our research, in fact, would have shown that there is no single answer to the question.

The Interchangeability of Indicators

Recall for a moment the Chapter 3 discussion of the *interchangeability of indexes*. In the present context, Lazarsfeld's earlier point suggests that we may be able to answer a general question such as whether men or women are the more compassionate—even when we cannot agree on the ultimate or even the best way of measuring it.

Suppose, for the moment, that you and I have compiled a list of 100 indicators of the concept *compassion* and its various dimensions. Suppose further that we disagree widely on which indicators give the clearest evidence of compassion or its absence. If we pretty much agree on some indicators, we could focus our attention on those, and we would probably agree on the answer they provided. But suppose we don't really agree on any of the possible indicators. It is still possible for us to reach an agreement on whether men or women are the more compassionate.

If we disagree totally on the value of the indicators, one solution would be to study all of them. Now, suppose that women turn out to be more compassionate than men on all 100 indicators—on all the indicators you favor and on all of mine. Then we would be able to agree that women are more compassionate than men even though we still disagree on what compassion means in general.

The interchangeability of indicators means that if several different indicators all represent, to some degree, the same concept, then all of them will behave the same way that the concept would behave if it were real and could be observed. Thus, if women are generally more compassionate than men, we should be able to observe that differ-

ence by using any reasonable measure of compassion.

You now have the fundamental logic of conceptualization and measurement. The discussions that follow in this chapter and the next one are mainly refinements and extensions of what I've just presented. Before turning to more technical elaborations on the main framework, however, it may be useful to cover more general topics.

First, I know the previous discussions may not fit exactly with your previous understanding of the meaning of such terms as *prejudice* and *compassion*. We tend to operate in daily life as though such terms have real, ultimate meanings. In the next subsection, then, I want to comment briefly on how we came to that understanding.

Second, concerned lest this whole discussion might create a picture of anarchy in the meanings of words, I will describe some of the ways in which scientists have organized the confusion so as to provide standards, consistency, and commonality in the meaning of terms. You should come away from this latter discussion with a recaptured sense of order—but one based on a conscious understanding rather than on a casual acceptance of common usage.

The Confusion over Definitions and Reality

Reviewing briefly, our concepts are derived from the mental images (conceptions) that summarize collections of seemingly related observations and experiences. Although the observations and experiences are real, our concepts are only mental creations. The terms associated with concepts are merely devices created for purposes of filing and communication. The word *prejudice* is an example. Ultimately, that word is only a collection of letters and has no intrinsic meaning. We could have as easily and meaningfully created the word *slanderice* to serve the same purpose.

Very often, however, we fall into the trap of believing that terms have real meanings. That danger seems to grow stronger when we begin to take terms seriously and attempt to use them precisely. And the danger is all the greater in the presence of experts who appear to know more than you do about what the terms really mean. It's very easy to yield to the authority of experts in such a situation.

Once we have assumed that terms have real meanings, we begin the tortured task of discovering what those real meanings are and what constitutes a genuine measurement of them. Figure 5-1 illustrates the history of this process. We make up conceptual summaries of real observations because the summaries are convenient. They prove to be so convenient, however, that we begin to think they are real. **The process of regarding as real things that are not is called reification, and the reification of concepts in day-to-day life is very common. . .**

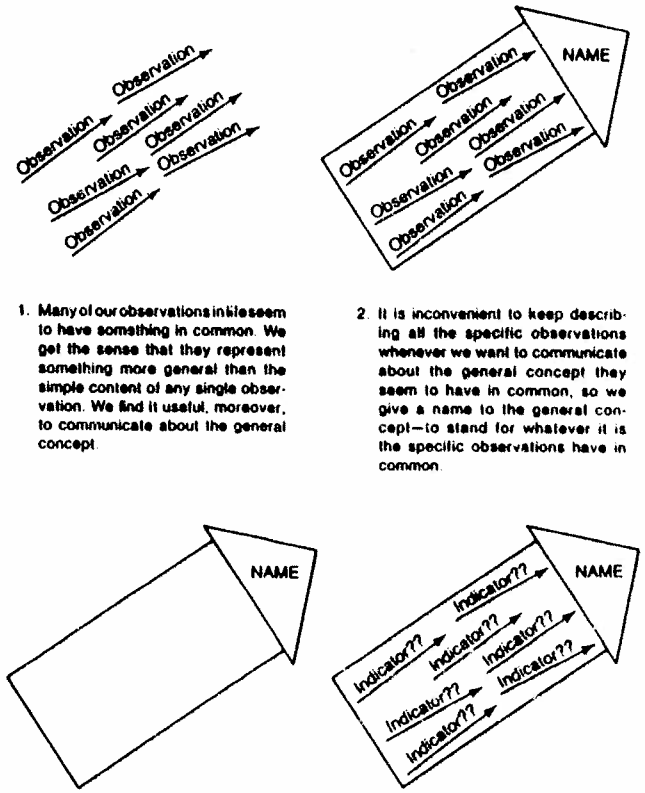
Creating Conceptual Order

The design and execution of social research requires a clearing away of the confusion over concepts and reality. To this end, logicians and scientists have found it useful to distinguish three kinds of definitions: *real*, *nominal*, and *operational*. The first of these reflects the reification of terms, and as Carl G. Hempel has cautioned,

> A "real" definition, according to traditional logic, is not a stipulation determining the meaning of some expression but a statement of the "essential nature" or the "essential attributes" of some entity. The notion of essential nature, however, is so vague as to render this characterization useless for the purposes of rigorous inquiry
>
> (1952:6)

The specification of concepts in scientific inquiry depends on nominal and operational definitions. A ***nominal*** definition is one that

Figure 5-1 The Process of Conceptual Entrapment



1. Many of our observations in life seem to have something in common. We get the sense that they represent something more general than the simple content of any single observation. We find it useful, moreover, to communicate about the general concept.
2. It is inconvenient to keep describing all the specific observations whenever we want to communicate about the general concept they seem to have in common, so we give a name to the general concept—to stand for whatever it is the specific observations have in common.
3. As we communicate about the general concept, using its term, we begin to think that the concept is *something* that really exists, not just a summary reference for several concrete observations in the world.
4. The belief that the concept itself is real results in irony. We now begin discussing and debating whether specific observations are "really" sufficient indicators of the concept.

is *assigned* to a term. In the midst of disagreement and confusion over what a term really means, the scientist specifies a working definition for the purposes of the inquiry. Wishing to examine socioeconomic status (SES) in a study, for example, we may simply specify that we are going to treat it as a combination of income and educational attainment. In that definitional decision, we rule out many other possible aspects of SES: occupational status, money in the bank, property, lineage, life-style, and so forth.

The specification of nominal definitions focuses our observational strategy, but it does not allow us to observe. As a next step we must specify exactly what we are going to observe, how we will do it, and what interpretations we are going to place on various possible observations. All of these further specifications make up what is called the **operational definition of the concept—a definition that spells out precisely how the concept will be measured.** Strictly speaking, an operational definition is a description of the "operations" that will be undertaken in measuring a concept.

Pursuing the case of SES, we might decide to ask the people we are studying two questions:

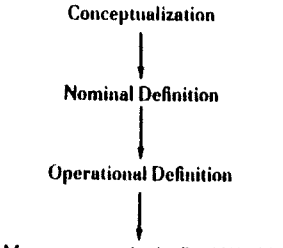
1. What was your total family income during the past twelve months?

2. What is the highest level of school you completed?

Here, we would probably want to specify a system for categorizing the answers people give us. For income, we might use categories such as "under $5,000" or "$5,000 to $10,000." Educational attainment might be similarly grouped in categories. Finally, we would specify the manner in which a person's responses to these two questions would be combined in creating a measure of SES. Chapter 7, on index and scale construction, will present some of the methods for doing that.

Ultimately, we would have created a working and workable definition of SES. Others might disagree with our conceptualization and operationalization, but the definition would have one essential scientific virtue: It would be absolutely specific and unambiguous. Even if someone disagreed with our definition, that person would have a good idea how to interpret our research results, because what we meant by the term *SES*—reflected in our analyses and conclusions—would be clear.

Here is a diagram showing the progression of measurement steps from our vague sense of what a term means to specific measurements in a scientific study:



A Conceptualization Example

I want to bring the preceding discussions together now through a brief history of a social scientific concept. You may recall from Chapter 2 that Edward Ransford, in his study of the Watts riots, was particularly interested in the part played by feelings of "powerlessness." Social scientists sometimes use the term *anomie* in this context. This term was first introduced into social science by Emile Durkheim, the great French sociologist, in his classic 1897 study *Suicide*.

Using only government publications on suicide rates in different regions and countries, Durkheim wrote a work of analytical genius. To determine the effects of religion on suicide, he compared the suicide rates of predominantly Protestant countries with predominantly Catholic ones, Protestant regions of Catholic countries with Catholic regions of Protestant countries, and so forth. To determine the possible effects of the weather, he compared suicide rates in northern and southern countries and regions, and he examined the different suicide rates across the months and seasons of the year. Thus, he was able to draw conclusions about a supremely individualistic and personal act without having any data about the individuals engaging in it.

At a more general level, Durkheim suggested that suicide also reflected the extent to which a society's agreements were clear and stable. Noting that times of social upheaval and change often present the individual with grave uncertainties about what is expected of him or her, Durkheim suggested that such uncertainties cause confusion, anxiety, and even self-destruction. To describe this societal condition of normlessness, Durkheim chose the term *anomie*. It is worth noting that Durkheim did not make this word up out of thin air. Used in both German and French, it meant, literally, *without law*, and the English term *anomy* had been used for at least three centuries before Durkheim to mean *disregard for divine law*. Still, Durkheim created anomie as a social scientific concept.

In the years that have followed the publication of *Suicide*, social scientists have found anomie a useful concept, and many have expanded on Durkheim's use. Robert Merton, in a classic article entitled "Social Structure and Anomie" (1938), concluded that anomie results from a disparity between the goals and means prescribed by a society. Monetary success, for example, is a widely shared goal in our society, yet not all individuals have the resources to achieve it through acceptable means. An emphasis on the goal itself, Merton suggested, produces normlessness, because those denied the traditional avenues to wealth go about getting it through illegitimate means. Merton's discussion, then, could be considered a further conceptualization of the concept of anomie.

Although Durkheim originally intended the concept of anomie to be a characteristic of societies, as did Merton after him, other social scientists have used it to describe individuals. (To clarify this distinction, some scholars have chosen to use the term *anomie* in its original, societal meaning and to use *anomia* in reference to the individual characteristic.) In a given society, then, some individuals experience anomia, and others do not. Elwin Powell, writing 20 years after Merton, provided the following conceptualization of anomia (though using the term *anomie*) as a characteristic of individuals:

> When the ends of action become contradictory, inaccessible or insignificant, a condition of anomie arises. Characterized by a general loss of orientation and accompanied by feelings of "emptiness" and apathy, anomie can be simply conceived as meaninglessness. (1958:132)

Powell went on to suggest there were two distinct kinds of anomia and to examine how the two rose out of different occupational experiences to result, sometimes, in suicide. In his study, however, Powell did not measure anomia per se; he studied the relationship between suicide and occupation, making inferences about the two kinds of anomia. Thus, the study did not provide an operational definition of anomia, only a further conceptualization.

Many other researchers have offered operational definitions, but one name stands out over all the others. Two years before Powell's article appeared, Leo Srole (1956)

The Origins of *Anomia*

by Leo Srole
Center for Geriatrics and Gerontology, Columbia University

My career-long fixation on anomie began with reading Durkheim's *Le Suicide* as a Harvard undergraduate. Later, as a graduate student at Chicago, I studied under two Durkheimian anthropologists: William Lloyd Warner and Alfred Radcliffe-Brown. Radcliffe-Brown had carried on a lively correspondence with Durkheim, making me a collateral "descendant" of the great French sociologist.

For me, the early impact of Durkheim's work on suicide was mixed but permanent. On the one hand, I had serious reservations about his strenuous, ingenious, and often awkward efforts to force the crude, bureaucratic records on suicide rates to fit with his unidirectional sociological determinism. On the other hand, I was moved by Durkheim's unswerving preoccupation with the moral force of the interpersonal ties that bind us to our time, place, and past, and also his insights about the lethal consequences that can follow from shrinkage and decay in those ties.

My interest in anomie received an eyewitness jolt at the finale of World War II, when I served with the United Nations Relief and Rehabilitation Administration, helping to rebuild a war-torn Europe. At the Nazi concentration camp of Dachau, I saw firsthand the depths of dehumanization that macro-social forces, such as those that engaged Durkheim, could produce in individuals like Hitler, Eichmann, and the others serving their dictates at all levels in the Nazi death factories.

Returning from my UNRRA post, I felt most urgently that the time was long overdue to come to an understanding of the dynamics underlying disintegrated social bonds. We needed to work expeditiously, deemphasizing proliferation of macro-level theory in favor of a direct exploratory encounter with individuals, using newly developed state-of-the-art survey research methodology. Such research, I also felt, should focus on a broader spectrum of behavioral pathologies than suicide.

My initial investigations were a diverse effort. In 1950, for example, I was able to interview a sample of 401 bus riders in Springfield, Mass. Four years later, the Midtown Manhattan Mental Health Study provided a much larger population reach. These and other field projects gave me scope to expand and refine my measurements of that quality in individuals which reflected the macro-social quality Durkheim had called *anomie*.

While I began by using Durkheim's term in my own work, I soon decided that it was necessary to limit the use of that concept to its macro-social meaning and to sharply segregate it from its individual manifestations. For the latter purpose, the cognate but hitherto obsolete Greek term, *anomia*, readily suggested itself.

I first published the *anomia* construct in a 1956 article in the *American Sociological Review*,* describing ways of operationalizing it, and presenting the results of its initial field application research. By 1982, the Science Citation Index and Social Science Citation Index had listed some 400 publications in political science, psychology, social work, and sociology journals here and abroad that had cited use of that article's instruments or findings, warranting the American Institute for Scientific Information to designate it a "citation classic."

*Leo Srole, "Social Integration and Certain Corollaries: An Exploratory Study," *American Sociological Review*, Vol. 21, 709–16, 1956.

11

published a set of questionnaire items that he said provided a good measure of anomia as experienced by individuals. It consists of five statements that subjects are asked to agree or disagree with.

1. In spite of what some people say, the lot of the average man is getting worse.
2. It's hardly fair to bring children into the world with the way things look for the future.
3. Nowadays a person has to live pretty much for today and let tomorrow take care of itself.
4. These days a person doesn't really know who he can count on.
5. There's little use writing to public officials because they aren't really interested in the problems of the average man.

(1956:713)

In the decades that followed its publication, the Srole scale has become a research staple for social scientists. You are likely to find this particular operationalization of anomie used in many of the research projects you'll find reported in academic journals. (Professor Srole touches on this in the accompanying box, "The Origins of *Anomia*," in which he discusses the history of his research in this area.)

I've presented this greatly abbreviated history of anomie and anomia as social scientific concepts for several reasons. First, they illustrate the process through which general concepts become operationalized measurements, though I wouldn't want you to think that the issue of anomie/anomia has been resolved once and for all. Scholars will surely continue to reconceptualize and reoperationalize them for years to come, continually seeking more useful measures.

I've ended the story with the Srole scale, however, because it illustrates another important point. Letting conceptualization and operationalization be open-ended does not necessarily produce anarchy and chaos as you might expect. Order emerges. There are several elements in this order. First, although you *could* define anomia any way you chose—in terms of, say, shoe size—you are likely to define it in ways not too different from other people's mental images. If you were to use a really offbeat definition, people would probably ignore you.

Second, as researchers discover the *utility* of a particular conceptualization and operationalization of a concept, they are likely to adopt it, and standardized definitions of concepts appear. Besides the Srole scale, examples include IQ tests and a whole host of demographic and economic measures developed by the Bureau of the Census. Using such established measures has two advantages: They have been extensively pretested and debugged, and studies using the same scales can be compared. If you and I do separate studies of two different groups, and if each of us uses the Srole scale, we will be able to compare our two groups on the basis of anomia.

Thus, social scientists can measure anything that's real, and we can even do a pretty good job of measuring things that aren't. Granting that such concepts as socioeconomic status, prejudice, compassion, and anomia aren't real ultimately, we've now seen that social scientists are able to create order

in handling them. It is an order based on utility, however, and not on ultimate truth.

The remainder of this chapter is devoted to some of the considerations and alternatives involved in the creation of *useful* definitions and measurements. First, we're going to look at the relationship between definitions and research purposes; then the chapter concludes with an examination of some criteria used in determining the *quality* of the measurements we create.

Definitions and Research Purposes

Recall from Chapter 4 that two of the general purposes of research are *description* and *explanation*. The distinction between them has important implications for the process of definition and measurement. If you have formed the opinion that description is a simpler task than explanation, you will be surprised to learn that definitions are more problematic for descriptive research than for explanatory research. This point will be discussed more fully in Part 4, but it is important that you have a basic understanding of why it is so before we turn to other aspects of measurement.

The importance of definitions for descriptive research should be clear. If our task is to describe and report the unemployment rate in a city, our definition of *being unemployed* is critical. That definition will depend on our definition of another term: the *labor force*. If it seems patently absurd to regard a three-year-old child as being unemployed, it is because such a child is not considered a member of the labor force. Thus, we might follow the U. S. Census Bureau's convention and exclude all persons under 14 years of age from the labor force.

This convention alone, however, would not give us a satisfactory definition, because it would count as unemployed such people as high school students, the retired, the disabled, and homemakers. We might follow the census convention further by defining the labor force as "all persons 14 years of age and over who are employed, looking for work, or waiting to be called back to a job from which they have been laid off or furloughed." Unemployed persons, then, would be those members of the labor force who are not employed. If a student, homemaker, or retired person is not looking for work, such a person would not be included in the labor force.

But what does "looking for work" mean? Must a person register with the state employment service or go from door to door asking for employment? Or would it be sufficient to want a job or be open to an offer of employment? Conventionally, "looking for work" is defined operationally as saying yes in response to an interviewer's asking "Have you been looking for a job during the past seven days?" (Seven days is the time period most often specified, but for some research purposes it might make more sense to shorten or lengthen it.)

I have spelled out these considerations in some detail so that you will realize that the conclusion of a descriptive study about the unemployment rate, for example, depends directly on how each issue is resolved. Increasing the period of time during which people are counted as looking for work would have the effect of adding more unemployed persons to the labor force as defined, thereby increasing the reported unemployment rate. If we follow another convention and speak of the *civilian* labor force and the *civilian* unemployment rate, we are excluding military personnel; that, too, increases the reported unemployment rate, because military personnel would be employed – *by definition*.

Thus the descriptive statement that the unemployment rate in a city is 3 percent, or 9 percent, or whatever it might be, depends directly on the operational definitions used. If that seems clear in this example, it is because there are a number of accepted conventions relating to the labor force and unemployment. Consider how difficult it would be to get agreement about the definitions needed to make the descriptive statement "45 percent of the students are politically conservative." This percentage, like the unemployment rate, would depend directly on your definition of what is being measured. A different definition might result in the conclusion "5 percent of the student body are politically conservative."

Ironically, definitions are less problematic in the case of explanatory research. Let's suppose we are interested in explaining political conservatism. Why are some people conservative and others not? More specifically, let's suppose we are interested in whether old people are generally more conservative than young people. What if you and I have 25 different operational definitions of *conservative*, and we can't agree on which definition is the best one? As we've already seen, this is not necessarily an insurmountable obstacle to our research. Suppose, for example, that we found old people more conservative than young people in terms of *all 25 definitions*! (Recall the earlier discussion of compassion in men and women.) Suppose we found old people more conservative than young people by *every* reasonable definition of conservatism we could think of. It wouldn't matter what our definition was. We would conclude that old people are generally more conservative than young people – even though we couldn't agree about what a conservative really was.

In practice, explanatory research seldom results in findings quite as unambiguous as this example suggests; nonetheless, the general pattern is quite common in actual research. There are consistent patterns of relationships in human social life, and they result in consistent research findings. The important point here, however, is that such consistency does not appear in a descriptive situation. Changing definitions almost inevitably result in different descriptive conclusions.

The box "The Importance of Variable Names" explores this issue in connection with the variable "citizen participation."

Criteria for 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 is possible to measure them anyway. I want to 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.

To begin, measurements can be made with varying degrees of *precision*, representing the fineness of distinctions made between attributes composing a variable. The description of a woman as "43 years old" is more precise than "in her forties." Saying "11½ inches long" is a more precise description than "about a foot long."

As a general rule, precise measurements are superior to imprecise ones, as common sense would dictate. There are no conditions under which imprecise measurements would be intrinsically superior to precise ones. Precision is not always necessary or desirable, however. If your research purpose is such that knowing a woman to be in her forties is sufficient, 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

The Importance of Variable Names

by Patricia Fisher
Graduate School of Planning, University of Tennessee

Operationalization is one of those things that's easier said than done. It is quite simple to explain to someone the purpose and importance of operational definitions for variables, and even to describe how operationalization typically takes place. However, until you've tried to operationalize a rather complex variable, you may not appreciate some of the subtle difficulties involved. Of considerable importance to the operationalization effort is the particular name that you have chosen for a variable. Let's consider an example from the field of Urban Planning.

A variable of interest to planners is citizen participation. Planners are convinced that participation in the planning process by citizens is important to the success of plan implementation. Citizen participation is an aid to planners' understanding of the real and perceived needs of a community, and such involvement by citizens tends to enhance their cooperation with and support for planning efforts. Although many different conceptual definitions might be offered by different planners, there would be little misunderstanding over what is *meant* by citizen participation. The name of the variable seems adequate.

However, if we asked different planners to provide very simple operational measures for citizen participation, we are likely to find a variety among their responses that does generate confusion. One planner might keep a tally of attendance by private citizens at city commission and other local government meetings; another might maintain a record of the different topics addressed by private citizens at similar meetings; while a third might record the number of local government meeting attendees, letters and phone calls received by the mayor and other public officials, and meetings held by special interest groups during a particular time period. As skilled researchers, we can readily see that each planner would be measuring (in a very simplistic fashion) a different *dimension* of citizen participation: extent of citizen participation, issues prompting citizen participation, and form of citizen participation. Therefore, the original *naming* of our variable, citizen participation, which was quite satisfactory from a conceptual point of view, proved inadequate for purposes of operationalization.

[redacted]

Variable names quite often evolve from an iterative process of forming a conceptual definition, then an operational definition, then renaming the concept to better match what can or will be measured. This looping process continues (our example above illustrates only one iteration), resulting in a gradual refinement of the variable name and its measurement until a reasonable fit is obtained. Sometimes the concept of the variable that you end up with is a bit different from the original one that you started with, but at least you are measuring what you are talking about, if only because you are talking about what you are measuring!

in Stowe, Vermont," is more precise than "born in New England"—but suppose the person in question was actually born in Boston. The less precise description, in this instance, would have been 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, would yield the same result each time. Suppose, for example, that I asked you to estimate how much I weigh. You look me over carefully and guess that I weigh 165 pounds. (Thank you.) Now let's suppose I ask you to estimate the weights of 30 or 40 other people, and while you're engrossed in that, I slip back into line wearing a clever disguise. When my turn comes again, you guess 180 pounds. Gotcha! That little exercise would have demonstrated that having you estimate people's weights was not a very reliable technique.

Suppose, however, that I had loaned you my bathroom scale to use in weighing people. No matter how clever my disguise, you would presumably announce the same weight for me both times, indicating that the scale provided a more reliable measure of weight than guessing.

Reliability, however, does not ensure accuracy any more than precision ensures it. 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 is called *bias*, and it is discussed in more detail in Chapter 8 on sampling. For now, just be warned that reliability does not ensure accuracy.

Let's suppose that we are interested in studying morale among factory workers in two different kinds of factories. In one set of factories, workers do very 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 spend more time observing 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 produced the higher morale.

Now let's look at some of the possible problems of reliability inherent in this method. First, how you and I are feeling when we do the observing is likely to color what we see. We may misinterpret what we see. We may see workers kidding each other and think they are having an argument. Or, maybe we'll 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 of time. Presumably that would be an indicator of morale: the more grievances, the lower the

morale. This measurement strategy would appear to be more reliable: We could count up the grievances over and over, and we should keep arriving at the same number.

If you find yourself saying "Wait a minute" over the second measurement strategy, you're worrying about validity, not reliability. Let's complete the discussion of reliability, and then we'll handle validity.

Reliability problems crop up in many forms in social research. Survey researchers have known for a long time that different interviewers get different answers from respondents as a result of their own attitudes and demeanor. If we were to conduct a study of 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. Different coders would 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 Bureau of the Census. You and I would not code all those occupations into 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 that are totally irrelevant to them. (Are you satisfied with China's current relationship with Albania?) And sometimes we ask questions that are 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.

How do you create reliable measures? There are a number of techniques; first in asking people for information—if your research design calls for that—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. The danger in these instances is that people *will* give you answers—reliable or not. People will tell you how they feel about China's relationship with Albania even if they haven't the foggiest idea what that relationship is.

Even when respondents are *able* to answer a question, their responses may lack reliability. Consider the dilemma of political reporters on the eve of the 1984 Democratic convention:

> To hear NBC tell it, Mondale has 415 potential delegate votes lined up; Gary Hart, 236, and Jesse Jackson, 30, with 165 uncommitted.
>
> According to CBS, though, Mondale has 752; Hart, 450; Jackson, 88; and 448 are uncommitted.
>
> ABC, meanwhile, counts 755 for Mondale; 444 for Hart; 72 for Jackson; 1 for John Glenn (who has withdrawn); and 355 uncommitted.
>
> The Associated Press and United Press International have similarly varying vote totals. No two counts agree. (The Chronicle uses the Associated Press totals, which show Mondale with 692 delegates; Hart, 422; Jackson, 76; and 160 uncommitted.)
>
> *San Francisco Chronicle*, 3/30/84, p. 12

An important reason for the discrepancies, the article suggested, was that "many of these people have announced whom they support, and the others have been polled by the people keeping count. But they do not always give everyone the same answer to the same question."

The problem of reliability is a basic one in social science measurement, and researchers have developed a number of techniques for dealing with it.

Test-Retest Method Sometimes it is appropriate to make the same measurement more than once. If there is no reason to expect the information sought to change, then you should expect the same response both times. If answers vary, however, that may indicate the measurement method is, to the extent

of that variation, unreliable. Here's an illustration.

In their research on Health Hazard Appraisal (HHA), a part of preventive medicine, Jeffrey Sacks and his colleagues (1980) wanted to determine the risks associated with various background and life-style 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 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 reports (1980:730): "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."

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 if the physician-counselors could have had nearly the impact on their patients as their patients' memories did. Thus, the data-collection method was not especially reliable.

Split-Half Method As a general rule, it is always a good idea 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* (see Chapter 8) 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 sets should correspond in the way they classify the respondents to the study. If the two sets of items measure people differently, that, again, points to a problem in the reliability of how you are measuring the variable.

Using Established Measures Another way to handle the problem of 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.

It is important to recognize that the heavy use of measures does not guarantee their reliability. In 1986, for example, both the *Scholastic Aptitude Tests* and the *Minnesota Multiphasic Personality Inventory (MMPI)*—accepted as established standards in their respective domains—were being fundamentally overhauled to reflect changes in society.

Research-Worker Reliability It is also possible for measurement unreliability to be generated by research workers: interviewers and coders, for example. There are several solutions. 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 are worried that newspaper editorials or occupations may not be classified reliably, why not have each independently coded by several coders? Those that generate disagreement should be evaluated more carefully and resolved.

Finally, clarity, specificity, training, and practice will avoid a great deal of unreliability and grief. If you and I were to spend some time reaching a clear agreement on how we were going to evaluate editorial positions on an issue—discussing the various positions that might be represented and reading through several together—we'd probably be able to 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, the term *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, I've already suggested some of the ways in which researchers deal with this issue.

First, there's something called face validity. Particular empirical measures may or may not jibe with our common agreements and our individual mental images associated with a particular concept. You and I might quarrel about the adequacy of measuring worker morale by counting the number of grievances filed with the union, but we'd surely agree that the number of grievances has *something* to do with morale. 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 any face validity.

Second, I've already pointed to many of the more concrete agreements researchers have reached in the case of some concepts. The Bureau of the Census, 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 those concepts.

Edward Carmines and Richard Zeller (1979) discuss three types of validity: *criterion-related validity*, *construct validity*, and *content validity*.

Criterion-related validity is sometimes called *predictive validity* and is based on some external criterion. For example, the validity of the college board is shown in its ability to predict the college success of students. 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 how well they drive. In these examples, college success and driving ability are the *criteria*. As a general matter, behavior may serve as a gauge of criterion validity for the many attitudinal measures we make in social research—for example, do "prejudiced" people actually discriminate against minorities?—although the relationship between attitudes and behavior is also an important subject of study in its own right.

Sometimes it is difficult to find behavioral criteria that can be taken to validate measures as directly as in the examples above. In those instances, however, we can often approximate such criteria by considering how the variable in question ought, theoretically, to relate to other variables. **Construct validity is based on the logical relationships among variables.**

Let's suppose, for example, that you are interested in studying "marital satisfaction"—its sources and consequences. 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 will have also 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 of marital satisfaction relates to marital fidelity in the expected fashion, that constitutes evidence of your measure's construct validity. If "satisfied" marriage partners were as likely to cheat on their spouses as 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. Whereas I have suggested here that tests of construct validity are less compelling than tests of criterion validity, however, you should realize there is room for disagreement as to which kind of test a particular variable constitutes in a given situation. It is less important that you distinguish these two types than that you understand the logic of validation that they have in common: if we have been successful in measuring some variable, then those measurements should relate in some logical fashion to other measures.

Finally, content validity refers to the degree to which a measure covers the range of meanings included within the concept. For example, a test of mathematical ability, Carmines and Zeller point out, cannot be limited to addition alone but would also need to cover subtraction, multiplication, division, and so forth. Or, if we say we are measuring prejudice *in general*, do our measurements reflect 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 can think of measurement as analogous to hitting the bull's-eye on a target, you'll see that reliability looks like a "tight pattern," regardless of where it hits, since 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 can be seen as a random error; the failure of validity is a systematic error. Notice that neither an unreliable nor an invalid measure is likely to be very useful.

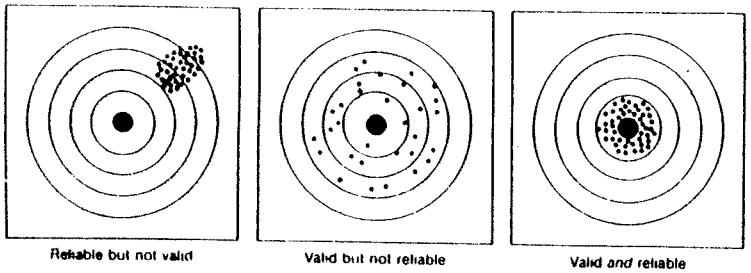
Tension between Reliability and Validity

As a footnote to these discussions, I want to point out briefly that a certain tension often exists between the criteria of reliability and validity. Often we seem to face a trade-off between the two.

If you'll recall for a moment the earlier example of measuring morale in different factories, I think you'll see that the strategy of immersing yourself in the day-to-day routine of the assembly line, observing what went on, and talking to the workers seems to provide a more valid measure of morale than counting grievances. It just seems obvious that we'd be able to get a clearer sense of whether the morale was high or low in that fashion than we would get from counting the number of grievances filed with the union.

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

Figure 5-2 An Analogy to Validity and Reliability



Suggested by an anonymous reviewer

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." Scores of books and articles have been written on the topic of anomie/anomia, and they still haven't exhausted the interesting aspects of that concept.

Yet, science needs to be specific to generate reliable measurements. Very often, then, the specification of reliable operational definitions and measurements seems to rob such concepts of their richness of meaning. I mean, morale is much more than a luck of grievances filed with the union; anomie is much more than the five items created by Leo Srole.

That is a persistent and inevitable dilemma for the social researcher, and you will be effectively forearmed against it by being forewarned. Be prepared for it, and deal with it. If there is no clear agreement on how to measure a concept, measure it several different ways. If the concept has several different dimensions, measure them all. And 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

- Concepts are mental images we use as summary devices for bringing together observations and experiences that seem to have something in common.
- Our concepts do not exist in the real world, so they can't be measured directly.
- It *is* possible to measure the things that our concepts summarize.
- Conceptualization is the process of specifying the vague mental imagery of our concepts, sorting out the kinds of observations and measurements that will be appropriate for our research.
- The interchangeability of indicators permits us to study and draw conclusions about concepts even when we can't agree on how those concepts should be defined.
- Precision refers to the exactness of the measure used in an observation or description of an attribute. For example, the description of a person as being "six feet, one and three-quarters inches tall" is more precise than the description "about six feet tall."
- Reliability refers to the likelihood that a given measurement procedure will yield the same description of a given phenomenon if that measurement is repeated. For example, estimating a person's age by asking his or her friends would be less reliable than asking the person or checking the birth certificate.
- Validity refers to the extent to which a specific measurement provides data that relate to commonly accepted meanings of a particular concept. There are numerous yardsticks for determining validity: face validity, criterion-related validity, content validity, and construct validity.
- The creation of 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 the concept.

Review Questions and Exercises

1. Pick a social science concept such as liberalism or alienation, and specify that concept so that it could be studied in a research project. Be sure to specify the dimensions you wish to include (and those you wish to exclude) in your conceptualization.
2. In a newspaper or magazine, find an instance of invalid and/or unreliable measurement. Justify your choice.

Additional Readings

Carmines, Edward G., and Zeller, Richard A., ***Reliability and Validity Assessment*** (Beverly Hills, CA: Sage, 1979). In this chapter, we've examined the basic logic of validity and reliability in social science measurement. Carmines and Zeller explore those issues in more detail and examine some of the ways for calculating reliability mathematically.

Gould, Julius, and Kolb, William, *A Dictionary of the Social Sciences* (New York: Free Press, 1964). A primary reference to the social scientific agreements on various concepts. Although the terms used by social scientists do not have ultimately "true" meanings, this reference book lays out the meanings social scientists have in mind when they use those terms.

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

Wallace, Walter, *The Logic of Science in Sociology* (Chicago: Aldine-Atherton, 1971), Chapter 3. A brief and lucid presentation of concept formation within the context of other research steps. This discussion relates conceptualization to observation on the one hand and to generalization on the other.

- Usually, short items in a questionnaire are better than long ones.
- Negative items and terms should be avoided in questionnaires because they may confuse respondents.
- Bias is the quality in questionnaire items that encourages respondents to answer in a particular way or to support a particular point of view. Avoid it.
- Operationalization begins in study design and continues throughout the research project, including the analysis of data.

Review Questions and Exercises

1. What level of measurement—nominal, ordinal, interval, or ratio—describes each of the following variables:
   a. Race (white, black, 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)

2. For each of the open-ended questions listed, construct a closed-ended question that could be used in a questionnaire.
   a. What was your family's total income last year?
   b. How do you feel about the MX missile system?
   c. How important is religion in your life?
   d. What was your main reason for attending college?
   e. What do you feel is the biggest problem facing this community?

Additional Readings

Feick, Lawrence, F., "Latent Class Analysis of Survey Questions that Include Don't Know Responses," ***Public Opinion Quarterly,*** (Winter 1989) Vol. 53: 525–547. The term "don't know" can mean a variety of things, as this analysis indicates.

Miller, Delbert, ***Handbook of Research Design and Social Measurement*** (New York: Longman, 1983). A useful reference work. This book, especially Part IV, cites and describes a wide variety of operational measures used in earlier social research. In a number of cases, the questionnaire formats used are presented. Though the quality of these illustrations is uneven, they provide excellent examples of the variations possible.

Oppenheim, A. N., ***Questionnaire Design and Attitude Measurement*** (New York: Basic Books, 1966). An excellent and comprehensive treatment of the construction of questionnaires and their relation to measurement in general. Although the illustrations of questionnaire formats are not always the best, this comes the closest of any book available to being the definitive work on questionnaires. Its coverage ranges from the theoretical to the nitty-gritty.

Smith, Eric R. A. N., and Squire, Peverill, "The Effects of Prestige Names in Question Wording," ***Public Opinion Quarterly,*** (Spring 1990) Vol. 54: 97–116). Not only do prestigious names affect the overall responses given to survey questionnaires, but they also affect such things as the correlation between education and the number of don't know answers.

Tourangeau, Roger et al., "Carryover Effects in Attitude Surveys," ***Public Opinion Quarterly,*** (Winter 1989) Vol. 53: 495–524. The authors asked six target questions in a telephone survey of 1,100 respondents, varying the questions immediately preceding the target questions. They found substantial differences.

lexes, Scales, and Typologies

What You'll Learn in This Chapter

Now we conclude the discussion of measurement, begun in Chapters 5 and 6. You'll learn the logic and skills of constructing composite measures from among several indicators of variables.

Introduction

This chapter discusses the construction of indexes and scales as composite measures of variables. A short section at the end of the chapter considers typologies. Each of these types of composite or cumulative measures combines several empirical indicators of a variable into a single measure.

Composite measures are frequently used in social science research, for several reasons. First, despite the care taken in designing studies to provide valid and reliable measurements of variables, the researcher seldom is able to develop in advance single indicators of complex concepts. That is especially true with regard to attitudes and orientations. The survey researcher, for example, is seldom able to devise single questionnaire items that adequately tap respondents' degrees of prejudice, religiosity, political orientations, alienation, and the like. More likely, you will devise several items, each of which provides *some* indication of the variables. Each of these, however, is likely to prove invalid or unreliable for many respondents.

You should realize that *some* variables are rather easily measured through single indicators. We may determine a survey respondent's sex by asking: Sex: ☐ Male ☐ Female. We may determine a newspaper's circulation by merely looking at the figure the newspaper reports. The number of times an experimental stimulus is administered to an experimental group is clearly defined in the design of the experiment. Nonetheless, social scientists, using a variety of research methods, frequently wish to study variables that have no clear and unambiguous single indicators.

Second, you may wish to employ a rather refined ordinal measure of your variable, arranging cases in several ordinal categories from—for example—very low to very high on a variable such as alienation. A single data item might not have enough categories to provide the desired range of variation, but an index or scale formed from several items would.

Finally, indexes and scales are *efficient* devices for data analysis. If considering a single data item gives us only a rough indication of a given variable, considering several data items may give us a more comprehensive and more accurate indication. For example, a single newspaper editorial may give us some indication of the political orientations of that newspaper. Examining several editorials would probably give us a better assessment, but the manipulation of several data items simultaneously could be very complicated. Indexes and scales (especially scales) are efficient *data-reduction devices*. Several indicators may be summarized in a single numerical score, while sometimes very nearly maintaining the specific details of all the individual indicators.

Indexes versus Scales

The terms index and scale are typically used imprecisely and interchangeably in social research literature. Before considering the distinctions this book will make between indexes and scales, let's first see what they have in common.

Both scales and indexes are typical *ordinal* measures of variables. Both rank-order people (or other units of analysis) in terms of specific variables such as religiosity, alienation, socioeconomic status, prejudice, or intellectual sophistication. A person's score on a scale or index of religiosity, for example, gives an indication of his or her relative religiosity vis-à-vis other people.

As the terms will be used in this book, both scales and indexes are *composite measures of variables*: measurements based on more than one data item. Thus, a survey respondent's score on an index or scale of religiosity would be determined by the specific responses given to several questionnaire items, each of which would provide some indication of his or her religiosity. Similarly, a person's IQ score is based on answers to a large number of test questions. The political orientation of a newspaper might be represented by an index or scale score reflecting the newspaper's editorial policy on a number of political issues.

In this book, we shall distinguish indexes and scales through the manner in which scores are assigned. An *index* is constructed through the simple accumulation of scores assigned to *individual* attributes. A *scale* is constructed through the assignment of scores to *patterns* of attributes. Thus, a scale takes advantage of any *intensity structure* that may exist among attributes. A simple example should clarify this distinction.

Figure 7-1 provides a graphic illustration of the difference between indexes and scales. Let's assume we want to develop a measure of political activism, distinguishing those people who are very active in political affairs, those who don't participate much at all, and those who are somewhere in between.

The first part of Figure 7-1 illustrates the logic of indexes. I've represented six different political actions. Although you and I might disagree on some specifics, I think we could agree that the six actions represent roughly the same degree of political activism. Although some people might give money more easily than write letters to the editor—or vice versa—the six actions are probably more or less equal if we consider the population as a whole.

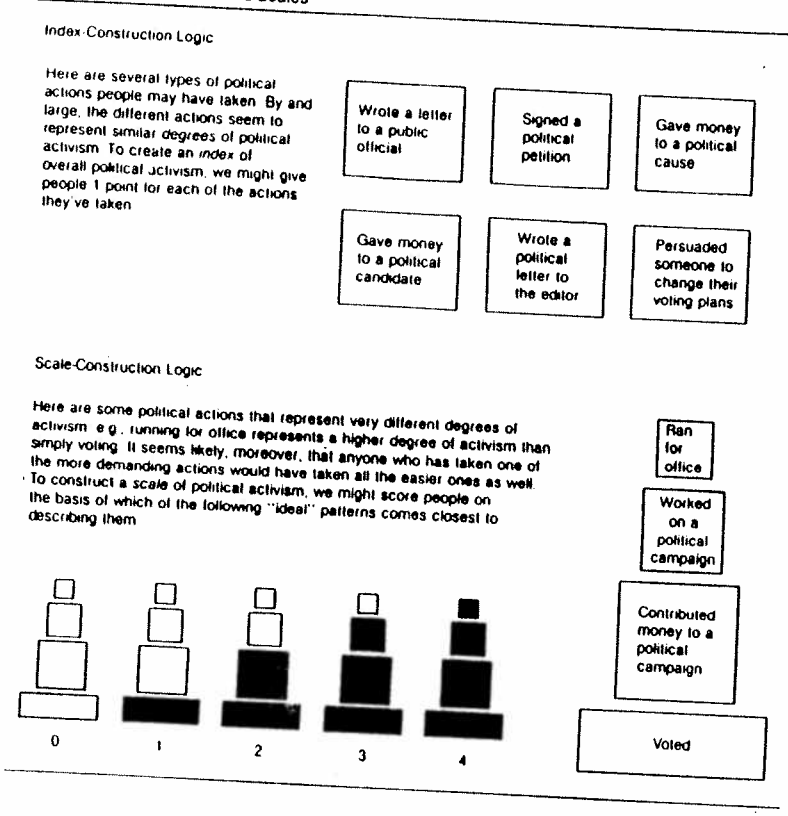
We could construct an index of political activism, using the six items, by giving each person 1 point for each of the actions he or she has taken. So if you wrote to a public official and signed a petition, you'd get a total of 2 points. If I gave money to a candidate and persuaded someone to change their vote, I'd get the same score as you. Using this approach, we'd conclude that you and I had the same degree of political activism, even though we had taken different actions.

The second part of Figure 7-1 describes the logic of scale construction. In this case, the actions clearly represent *different* degrees of political activism—ranging from simply voting to running for office. Moreover, it seems safe to assume a *pattern* of actions in this case. For example, all those who contributed money probably also voted. Those who worked on a campaign probably also gave some money and voted. This suggests that most people will only fall into one of five "ideal" action patterns, represented by the small illustrations at the bottom of the figure. The discussion of scales, later in this chapter, describes ways of identifying people with the type they most closely represent.

It should be apparent that scales are generally superior to indexes, if for no other reason than that scale scores convey more information than index scores. Still, you should be wary of the common misuse of the term *scale*; clearly, calling a given measure a scale rather than an index does not make it better. You should be cautioned against two other misconceptions about scaling. First, whether the combination of several data items results in a scale almost always depends on the particular sample of observations under study. Certain items may form a scale among one sample but not among another, and you should not assume that a given set of items *is* a scale because it has formed a scale among a given sample. Second, the use of certain *scaling techniques* to be discussed does not assure the creation of a scale any more than the use of items that have previously formed scales can offer such assurance.

An examination of the substantive literature based on social science data will show that indexes are used much more frequently

Figure 7-1 Indexes versus Scales



than scales. Ironically, however, the methodological literature contains little if any discussion of index construction, but discussions of scale construction abound. There appear to be two reasons for this disparity. First, indexes are more frequently used because scales are often difficult or impossible to construct from the data at hand. Second, methods of index construction are not discussed because they seem obvious and straightforward.

Index construction is not a simple undertaking, though the general failure to develop index construction techniques has resulted in the creation of many bad indexes in social research. With this in mind, I have devoted most of this chapter to the methods of index construction. Once you fully understand the logic of this activity, you will be better equipped to attempt the construction of scales. Indeed, the carefully constructed index may turn out to be a scale anyway.

Index Construction

Let's look now at the several steps involved in the creation of an index: selecting possible items, examining their empirical relationships, combining some items into an index, and validating it. I have presented these steps in some detail, since they are not all obvious. You should come away from this section able to create a composite measure that will fully support your subsequent analyses.

Item Selection

A composite index is created to measure some variable. The first criterion for selecting items to be included in the index is *face validity* (or logical validity). If you want to measure political conservatism, for example, each of your items should appear *on its face* to indicate conservatism (or its opposite: liberalism). Political party affiliation would be one such item. If people were asked to approve or disapprove of the views of a well-known conservative public figure, their responses might logically provide another indication of their conservatism. In constructing an index of religiosity, you might consider items such as church attendance, acceptance of certain religious beliefs, and frequency of prayer; each of these appears to offer some indication of religiosity.

The methodological literature on conceptualization and measurement stresses the need for *unidimensionality* in scale and index construction: A composite measure should represent only one dimension. Thus, items reflecting religiosity should not be included in a measure of political conservatism, even though the two variables might be empirically related to one another.

At the same time, you should be aware of subtle nuances that may exist within the general dimension you are attempting to measure. Thus in the example of religiosity, the indicators mentioned previously represent different *types* of religiosity – ritual participation, belief, and so on. If you wished to focus on ritual participation in religion, you should choose items specifically indicating this type of religiosity: church attendance, communion, confession, and the like. If you wished to measure religiosity in a more general way, you would include a balanced set of items, representing each of the different types of religiosity. Ultimately, the nature of the items included will determine how specifically or generally the variable is measured.

In selecting items for an index, you must also be concerned with the amount of *variance* provided by those items. If an item is intended to indicate political conservatism, for example, you should note what proportion of respondents were identified as conservatives by the item. If a given item identified no one as a conservative or everyone as a conservative – for example, if nobody indicated approval of a radical right political figure – that item would not be very useful in the construction of an index.

To guarantee variance, you have two options. First, you may select several items on which responses divide people about equally in terms of the variable; for example, about half conservative and half liberal. Although no single response would justify characterization of a person as very conservative, a person who responded as a conservative on all items might be so characterized.

The second option is to select items differing in variance. One item might identify about half the subjects as conservative, and

another might identify few of the respondents as conservative. (*Note*: This latter option is necessary for scaling, but it is reasonable for index construction as well.)

Bivariate Relationships among Items

The second step in index construction is to examine the bivariate relationships among the items being considered for inclusion.[6] If each of the items is indeed valid on its face, then the several items should be related to one another empirically. For example, if several items all reflect conservatism or liberalism, then respondents who appear conservative in terms of one item should appear conservative in terms of others. Recognize, however, that such items will seldom if ever be perfectly related to one another; persons who appear conservative on one item will appear liberal on another. (This disparity creates the need for constructing composite measures in the first place.) Nevertheless, persons who appear conservative on Item A should be more likely to appear conservative on Item B than persons who appear liberal on Item A.

You should examine all the possible bivariate relationships among the several items being considered for inclusion in the index to determine the relative strengths of relationships among the several pairs of items. Either percentage tables or correlation coefficients (see Chapter 17), or both, may be used for this purpose. The primary criterion for evaluating these several relationships is the strength of the relationships. The use of this criterion, however, is rather subtle. (The box entitled "'Cause' and 'Effect' Indicators" examines some of those subtleties.)

Clearly, you should be wary of items that are not related to one another empirically. It is unlikely that they measure the same variable if they are unrelated. A given item that is not related to several other items probably should be dropped from consideration.

At the same time a *very* strong relationship between two items presents a different problem. If two items are perfectly related to one another, then only one is necessary for inclusion in the index, since it completely conveys the indications provided by the other. (This problem will become even clearer in the next section.)

To illustrate the testing of bivariate relationships in index construction, an example from the substantive literature may be useful. A few years ago, I conducted a survey of medical school faculty members to find out about the consequences of a "scientific perspective" on the quality of patient care provided by physicians. The primary intent was to determine whether more scientifically inclined doctors treated patients more impersonally than other doctors.

The survey questionnaire offered several possible indicators of respondents' scientific perspectives. Of those, three items appeared to provide especially clear indications of whether the doctors were scientifically oriented:

> 1. As a medical school faculty member, in what capacity do you feel you can make your greatest *teaching* contribution: as a practicing physician or as a medical researcher?
>
> 2. As you continue to advance your own medical knowledge, would you say your ultimate medical interests lie primarily in the direction of total patient management or the understanding of basic mechanisms?
>
> 3. In the field of therapeutic research, are you generally more interested in articles reporting evaluations of the effectiveness of various treatments or articles exploring the basic rationale underlying the treatments?
>
> (Babbie, 1970:27–31)

For each of these items, we might conclude that those respondents who chose the second answer are more scientifically oriented than respondents who chose the first answer. This *comparative* conclusion is a reasonable one, but we should not be misled into thinking that respondents who chose the second answer to a given item are scientists in any absolute sense. They are simply *more scientific* than those who chose the first answer to the item. To see this point more clearly, let's examine the distribution of responses to each item. From the first item—best teaching role—only about one-third of the respondents appeared scientifically oriented. (Approximately one-third said they could make their greatest teaching contribution as medical researchers.) In response to the second item—ultimate medical interests—approximately two-thirds chose the scientific answer, saying they were more interested in learning about basic mechanisms than learning about total patient management. In response to the third item—reading

"Cause" and "Effect" Indicators

by Kenneth Bollen
Department of Sociology, Dartmouth College

While it often makes sense to expect indicators of the same variable to be positively related to one another, as discussed in the text, this is not always the case.

[redacted]. For example, to measure self-esteem, we might ask a person to indicate whether they agree or disagree with the statements (1) "I am a good person" and (2) "I am happy with myself." A person with high self-esteem should agree with both statements while one with low self-esteem would probably disagree with both. Since each indicator depends on or "reflects" self-esteem, we expect them to be positively correlated. More generally, indicators that depend on the same variable should be associated with one another if they are valid measures.

But, this is not the case when the indicators are the "cause" rather than the "effect" of a variable. In this situation the indicators may correlate positively, negatively, or not at all. For example, we could use sex and race as indicators of the variable *exposure to discrimination*. Being nonwhite or female increases the likelihood of experiencing discrimination, so both are good indicators of the variable. But we would not expect the race and sex of individuals to be strongly associated.

Or, we may measure *social interaction* with three indicators: time spent with friends, time spent with family, and time spent with coworkers. Though each indicator is valid, they need not be positively correlated. Time spent with friends, for instance, may be *inversely* related to time spent with family. Here, the three indicators "cause" the degree of social interaction.

As a final example, exposure to stress may be measured by whether a person recently experienced divorce, death of a spouse, or loss of a job. Though any of these events may indicate stress, they need not correlate with one another.

[redacted]

preferences – about 80 percent chose the scientific answer.

So these three questionnaire items cannot tell us how many "scientists" there are in the sample, for none of them is related to a set of criteria for what constitutes being a scientist in any absolute sense. Using the items for this purpose would present us with the problem of three quite different estimates of how many scientists there were in the sample.

However, these items do provide us with three independent indicators of respondents' *relative* inclinations toward science in medicine. Each item separates respondents into the *more* scientific and the *less* scientific. But each grouping of more or less scientific respondents will have a somewhat different membership from the others. Respondents who seem scientific in terms of one item will not seem scientific in terms of another. Nevertheless, to the extent that each item measures the same general dimension, we should find some correspondence among the several groupings. Respondents who appear scientific in terms of one item should be more likely to appear scientific in their response to another item than those who appeared nonscientific in their response to the first. We should find an association or correlation between the responses given to two items.

Figure 7-2 shows the associations among the responses to the three items. Three bivariate tables are presented, showing the conjoint distribution of responses for each pair of items. Although each single item produces a different grouping of "scientific" and "nonscientific" respondents, we see in Figure 7-2 that the responses given to each of the items correspond, to a degree, to the responses given to each of the other items.

An examination of the three bivariate relationships presented in Figure 7-2 supports the suggestion that the three items all measure the same variable: scientific orientations. To see why this is so, let's begin by looking at the first bivariate relationship in the table. The table shows that faculty who responded that "researcher" was their best teaching role were more likely to identify their ultimate medical interests as "basic mechanisms" than those who answered "physician." The data show that 87 percent of the "researchers" also chose the scientific response to the second item, as opposed to 51 percent of the "physicians." (*Note*: The fact that the "physicians" are about evenly split in their ultimate medical interests is irrelevant. It is only relevant that they are *less* scientific in their medical interests than the "researchers.") The strength of this relationship may be summarized as a 36 percentage point difference.

The same general conclusion applies to the other bivariate relationships. The strength of the relationship between reading preferences and ultimate medical interests may be summarized as a 38 percentage point difference; the strength of the relationship between reading preferences and the two teaching roles may be summarized as a 21 percentage point difference.

Initially, the three items were selected on the basis of face validity – each appeared to give some indication of faculty members' orientations to science. By examining the bivariate relationship between the pairs of items, we have found support for the expectation that they all measure basically the same thing. However, that support is not sufficient justification to include the items in a composite index.

Multivariate Relationships among Items

Before combining them in a single index, we need to examine the multivariate relationships among the several variables. Recall that the primary purpose of index construction is to develop a method of classifying subjects in terms of some variable such as political conservatism, religiosity, scientific orientation, or whatever. An index of political conservatism should identify those who are very conservative, moderately conservative, not very conservative, and not at all conservative (or moderately liberal and very liberal, respectively, in place of the last two categories). The several gradations of the variable are provided by the combination of responses given to the several items included in the index. Thus, the respondent who appeared conservative on all items would be considered very conservative overall.

Figure 7-2 Bivariate Relationships among Scientific Orientation Items

A.

Ultimate Medical Interest	Best Teaching Role: Physician	Best Teaching Role: Researcher
Total patient management	49%	13%
Basic mechanisms	51%	87%
	100% (268)	100% (159)

B.

Ultimate Medical Interest	Reading Preferences: Effectiveness	Reading Preferences: Rationale
Total patient management	68%	30%
Basic mechanisms	32%	70%
	100% (78)	100% (349)

C.

Best Teaching Role	Reading Preferences: Effectiveness	Reading Preferences: Rationale
Physician	85%	64%
Researcher	15%	36%
	100% (78)	100% (349)

For an index to provide meaningful gradations in this sense, each item must add something to the evaluation of each respondent. Recall that in the preceding section it was suggested that two items perfectly related to one another should not be included in the same index. If one item were included, the other would add nothing to our evaluation of respondents. The examination of multivariate relationships among the items is

another way of eliminating deadwood. It also determines the overall power of the particular collection of items in measuring the variable under consideration.

The purposes of this multivariate examination will become clearer if we return to the earlier example of measuring scientific orientations among medical school faculty members. Figure 7-3 presents the trivariate relationships among the three items.

Figure 7-3 has been presented somewhat differently from Figure 7-2. In this instance, the sample respondents have been categorized in four groups according to (1) their best teaching roles and (2) their reading preferences. The numbers in parentheses indicate the number of respondents in each group. (Thus 66 of the faculty members who said they could best teach as physicians also said they preferred articles dealing with the effectiveness of treatments.) For each of the four groups, the percentage that say they are ultimately more interested in basic mechanisms has been presented. (Of the 66 faculty mentioned, 27 percent are primarily interested in basic mechanisms.)

The arrangement of the four groups is based on a previously drawn conclusion regarding scientific orientations. The group in the upper left corner of the table is presumably the least scientifically oriented, based on best teaching role and reading preference. The group in the lower right corner is presumably the most scientifically oriented in terms of those items.

Recall that expressing a primary interest in basic mechanisms was also taken as an indication of scientific orientations. As we should expect, then, those in the lower right corner are the most likely to give this response (89 percent) and those in the upper left corner are the least likely (27 percent). The respondents who gave mixed responses in terms of teaching roles and reading preferences have an intermediate rank in their concern for basic mechanisms (58 percent in both cases).

This table tells us many things. First, we may note that the original relationships between pairs of items are not significantly affected by the presence of a third item. Recall, for example, that the relationship between teaching role and ultimate medical interest was summarized as a 36 percentage point difference. Looking at Figure 7-3, we see that among only those respondents who are most interested in articles dealing with the effectiveness of treatments, the relationship between teaching role and ultimate medical interest is 31 percentage points (58 percent minus 27 percent: first row), and the same is true among those most interested in articles dealing with the rationale for treatments (89 percent minus 58 percent: second row). The original relationship between teaching role and ultimate medical interest is essentially the same as in Figure 7-2, even among those respondents judged as scientific or nonscientific in terms of reading preferences.

The same conclusion may be drawn from the columns in Figure 7-3. Recall that the original relationship between reading preferences and ultimate medical interests was summarized as a 38 percentage point difference. Looking only at the "physicians" in Figure 7-3, we see that the relationship between the other two items is now 31 percentage points. The same relationship is found among the "researchers" in the second column.

The importance of these observations becomes clearer when we consider what might have happened. Figure 7-4 presents hypothetical data to illustrate that. These data tell a much different story than the actual data reported in Figure 7-3. In this instance, it is evident that the original relationship between teaching role and ultimate medical interest persists, even when reading preferences are introduced into the picture. In each row of the table the "researchers" are more likely to express an interest in basic mechanisms than the "physicians." Looking down the columns, however, we note that there is no relationship between reading preferences

Figure 7-3 Trivariate Relationships among Scientific Orientation Items

Percentage Interested in Basic Mechanisms

Reading Preferences	Best Teaching Role: Physician	Best Teaching Role: Researcher
Effectiveness	27% (66)	58% (12)
Rationale	58% (219)	89% (130)

and ultimate medical interest. If we know whether a respondent feels he or she can best teach as a physician or as a researcher, knowing the respondent's reading preference adds nothing to our evaluation of his or her scientific orientation. If something like Figure 7-4 resulted from the actual data, we would conclude that reading preference should not be included in the same index as teaching role, since it contributes nothing to the composite index.

This example used only three questionnaire items. If more were being considered, then more complex multivariate tables would be in order, constructed of four, five, or more variables. The purpose of this step in index construction, again, is to discover the simultaneous interaction of the items to determine whether they are all appropriate for inclusion in the same index.

Index Scoring

When you have chosen the best items for the index, you next assign scores for particular responses, thereby creating a single composite index out of the several items. There are two basic decisions to be made in this step.

First, you must decide the desirable range of the index scores. Certainly a primary advantage of an index over a single item is the range of gradations it offers in the measurement of a variable. As noted earlier, political conservatism might be measured from "very conservative" to "not at all conservative" (or "very liberal"). How far to the extremes, then, should the index extend?

In this decision, the question of variance enters once more. Almost always, as the possible extremes of an index are extended, fewer cases are to be found at each end. The researcher who wishes to measure political conservatism to its greatest extreme may find there is almost no one in that category.

The first decision, then, concerns the conflicting desire for (1) a range of measurement in the index and (2) an adequate number of cases at each point in the index. You will be forced to reach some kind of compromise between these conflicting desires.

The second decision concerns the actual assignment of scores for each particular response. Basically you must decide whether to give each item an equal weight in the index or to give them different weights. Although there are no firm rules, I suggest—and practice tends to support this method—that items be weighted equally unless there are compelling reasons for differential weighting. That is, the burden of proof should be on differential weighting; equal weighting should be the norm.

Of course, this decision must be related to the earlier issue regarding the balance of

Figure 7-4 Hypothetical Trivariate Relationship among Scientific Orientation Items

Percentage Interested in Basic Mechanisms

Reading Preferences	Best Teaching Role: Physician	Best Teaching Role: Researcher
Effectiveness	51% (66)	87% (12)
Rationale	51% (219)	87% (130)

items chosen. If the index is to represent the composite of slightly different aspects of a given variable, then you should give each aspect the same weight. In some instances, however, you may feel that, say, two items reflect essentially the same aspect, and the third reflects a different aspect. If you wished to have both aspects equally represented by the index, you might decide to give the different item a weight equal to the combination of the two similar ones. In such a situation, you might want to assign a maximum score of 2 to the different item and a maximum score of 1 to each of the similar ones.

Although the rationale for scoring responses should take such concerns as these into account, you will typically experiment with different scoring methods, examining the relative weights given to different aspects but at the same time worrying about the range and distribution of cases provided. Ultimately, the scoring method chosen will represent a compromise among these several demands. (*Note*: In this activity, as in most research activities, the decision is open to revision on the basis of later examinations. Validation of the index, to be discussed shortly, may lead you to recycle your efforts and to construct a completely different index.)

In the example taken from the medical school faculty survey, I decided to weight the items equally, since they had been chosen, in part, on the basis of their representing slightly different aspects of the overall variable—scientific orientation. On each of the items, the respondents were given a score of 1 for choosing the "scientific" response to the item and a score of 0 for choosing the "nonscientific" response. Each respondent, then, had a chance of receiving a score of 0, 1, 2, or 3, depending on the number of "scientific" responses he or she chose. This scoring method provided what was considered a useful range of variation—four index categories—and also provided enough cases in each category for analysis.

Here's a similar example of index scoring, from a recent study of work satisfaction. One of the key variables was "job-related depression," measured by an index composed of the following four items, which asked workers how they felt when thinking about themselves and their jobs:

- "I feel downhearted and blue."
- "I get tired for no reason."
- "I find myself restless and can't keep still."
- "I am more irritable than usual."

The researchers, Amy Wharton and James Baron, report: "Each of these items was coded: 4 = often, 3 = sometimes, 2 = rarely, 1 = never."

(Wharton, 1987:578)

They go on to explain how they measured other variables examined in the study:

> Job-related self-esteem was based on four items asking respondents how they saw themselves in their work: happy/sad; successful/not successful; important/not important; doing their best/not doing their best. Each item ranged from 1 to 7, where 1 indicates a self-perception of not being happy, successful, important, or doing one's best.

As you look through the social research literature, you will find numerous, similar examples of cumulative indexes being used to measure variables.

Handling Missing Data

Regardless of your data-collection method, you will frequently face the problem of missing data. In a content analysis of the political orientations of newspapers, for example, you may discover that a particular newspaper has never taken an editorial position on one of the issues being studied—it may never have taken a stand on the United Nations, for example. In an experimental design involving several retests of subjects over time, some subjects may be unable to participate in some of the sessions. In virtually every survey, some respondents fail to answer some questions (or choose a "don't know" response). Although missing data present problems at all stages of analysis, it is especially troublesome in index construction. There are, however, several methods of dealing with the problem of missing data.

First, if there are relatively few cases with missing data, you may decide to exclude them from the construction of the index and the analysis. (In the medical school faculty example, this was the decision I made regarding missing data.) The primary concerns in this instance are whether the numbers available for analysis will still be sufficient and whether the exclusion will result in a biased sample whenever the index is used in the analysis. The latter possibility can be examined through a comparison—on other relevant variables—of those who would be included and excluded from the index.

Second, you may sometimes have grounds for treating missing data as one of the available responses. For example, if a questionnaire has asked respondents to indicate their participation in a number of activities by checking "yes" or "no" for each, many respondents may have checked some of the activities "yes" and left the remainder blank. In such a case, you might decide that a failure to answer meant "no," and score missing data in this case as though the respondents had checked the "no" space.

Third, a careful analysis of missing data may yield an interpretation of their meaning. In constructing a measure of political conservatism, for example, you may discover that respondents who failed to answer a given question were generally as conservative on other items as those who gave the conservative answer. As another example, a recent study measuring religious beliefs found that people who answered "don't know" about a given belief were almost identical to the "disbelievers" in their answers about other beliefs. (*Note*: You should not take these examples as empirical guides in your own studies, but only as suggestive of ways to analyze your own data.) Whenever the analysis of missing data yields such interpretations, then, you may decide to score such cases accordingly.

There are a number of other ways for handling this problem. If an item has several possible values, you might assign the middle value to cases with missing data; for example, you could assign a 2 if the values are 0, 1, 2, 3, and 4. For a continuous variable such as age, you could similarly assign the mean to cases with missing data. Or, missing data can be supplied by assigning values at random. All of these are conservative solutions in that

they work against any relationships you may expect to find.

If you're creating an index out of several items, it sometimes works to handle missing data by using proportions based on what is observed. Suppose your index is composed of six indicators, and you only have four observations for a particular subject. If the subject has earned 4 points out of a possible 4, you might assign an index score of 6; if the subject has 2 points (half the possible score on four items), you could assign a score of 3 (half the possible score on six observations).

The choice of a particular method to be used depends so much on the research situation as to preclude the suggestion of a single "best" method or a ranking of the several I have described. Excluding all cases with missing data can bias the representativeness of the findings, but including such cases by assigning scores to missing data can influence the nature of the findings. The safest and best method would be to construct the index using alternative methods and see whether the same findings follow from each. Understanding your data is the final goal of analysis anyway.

Index Validation

Up to this point, we have discussed all the steps in the selection and scoring of items that result in a composite index purporting to measure some variable. If each of the preceding steps is carried out carefully, the likelihood of the index actually measuring the variable is enhanced. To demonstrate success, however, there must be *validation* of the index. In the basic logic of validation, we assume that the composite index provides a measure of some variable; that is, the successive scores on the index arrange cases in a rank order in terms of that variable. An index of political conservatism rank-orders people in terms of their relative conservatism. If the index does that successfully, then persons scored as relatively conservative on the index should appear relatively conservative in all *other indications* of political orientation, such as questionnaire items. There are several methods for validating a composite index.

Item Analysis The first step in index validation is an internal validation called item analysis. In item analysis, you examine the extent to which the composite index is related to (or predicts responses to) the items in the index itself. Simply create tables in which the index is the independent variable and one of the items is the dependent variable. If the index has been carefully constructed through the examination of bivariate and multivariate relationships among several items, this step should confirm the validity of that index, with each individual item correlating with index scores.

In a complex index containing many items, this step provides a convenient test of the independent contribution of each item to the index. If a given item is found to be poorly related to the index, it may be assumed that other items in the index cancel out the contribution of that item. If the item in question contributes nothing to the index's power, it should be excluded.

Although item analysis is an important first test of the index's validity, it is scarcely a sufficient test. If the index adequately measures a given variable, it should successfully predict other indications of that variable. To test that, we must turn to items not included in the index.

External Validation Persons scored as politically conservative on an index should appear conservative in their responses to other items in the questionnaire. Of course, we are talking about *relative* conservatism, as we are unable to make a final absolute definition of what constitutes conservatism. However, those respondents scored as the most conservative on the index should be the most conservative in answering other questions. Those scored as the least conservative on the index should be the least conservative on other items. Indeed, the ranking of groups of respondents on the index should predict the ranking of those groups in answering other questions dealing with political orientations.

In our example of the scientific orientation index, several questions in the questionnaire offered the possibility of further validation. Table 7-1 presents some of those items.

These items provide several lessons regarding index validation. First, we note that the index strongly predicts the responses to the validating items in the sense that the rank order of scientific responses among the four groups is the same as the rank order provided by the index itself. At the same time, each item gives a different *description* of scientific orientations overall. For example, the last validating item indicates that the great majority of *all* faculty were engaged in research during the preceding year. If this were the only indicator of scientific orientation, we would conclude that nearly all faculty were scientific. Nevertheless, those scored as more scientific on the index are more likely to have engaged in research than those who were scored as relatively less scientific. The third validating item provides a different descriptive picture: Only a minority of the faculty overall say they would prefer duties limited exclusively to research. Nevertheless, the percentages giving this answer correspond to the scores assigned on the index.

Bad Index versus Bad Validators Nearly every index constructor at some time must face the apparent failure of external items to validate the index. If the internal item analysis shows inconsistent relationships between the items included in the index and the index itself, something is wrong with the index. But if the index fails to predict strongly the external validation items, the conclusion to be drawn is more ambiguous. You must choose between two possibilities: (1) the index does not adequately measure the variable in question, or (2) the validation items do not adequately measure the variable and thereby do not provide a sufficient test of the index.

The researcher who has worked long and conscientiously on the construction of an index will find the second conclusion very compelling. Typically, you will feel you have included the best indicators of the variable in the index; the validating items are, therefore, second-rate indicators. Nevertheless, you should recognize that the index is purportedly a very powerful measure of the variable; thus, it should be somewhat related to any item that taps the variable even poorly.

When external validation fails, you should reexamine the index before deciding that the validating items are insufficient. One method of doing that is to examine the relationships between the validating items and the individ-

Table 7-1 Validation of Scientific Orientation Index

	Index of Scientific Orientation			
	Low 0	1	2	High 3
Percentage interested in attending scientific lectures at the medical school	34	42	46	65
Percentage who say faculty members should have experience as medical researchers	43	60	65	89
Percentage who would prefer faculty duties involving research activities only	0	8	32	66
Percentage who engaged in research during preceding academic year	61	76	94	99

onal items included in the index. If you discover that some of the index items relate to the validators and others do not, that will improve your understanding of the index as it was initially constituted.

There is no cookbook solution to this dilemma; it is an agony serious researchers must learn to survive. Ultimately, the wisdom of your decision to accept an index will be determined by the usefulness of that index in your later analyses. Perhaps you will initially decide that the index is a good one and that the validators are defective, and later find that the variable in question (as measured by the index) is not related to other variables in the ways you expected. Then you may have to compose a new index.

Likert Scaling

Earlier in this chapter, I defined a scale as a composite measure based on *intensity structure* among the items composing the measure. In scale construction, response patterns across several items are scored, whereas in index construction, individual responses are scored, and those independent scores are summed. By this definition, the measurement method developed by Rensis Likert, called Likert scaling, represents a systematic and refined means for constructing indexes from questionnaire data. I'll discuss this method here, therefore, rather than in the sections on scaling to follow.

The Likert format is associated with a question format that is frequently used in contemporary survey questionnaires. Basically, the respondent is presented with a statement in the questionnaire and is asked to indicate whether he or she "strongly agrees," "agrees," "disagrees," "strongly disagrees," or is "undecided." Modifications of the wording of the response categories (for example, "approve") may be used, of course.

The particular value of this format is the unambiguous ordinality of response categories. If respondents were permitted to volunteer or select such answers as "sort of agree," "pretty much agree," "really agree," and so forth, the researcher would find it impossible to judge the relative strength of agreement intended by the various respondents. The Likert format resolves this dilemma.

The Likert format also lends itself to a rather straightforward method of index construction. Because identical response categories are used for several items intended to measure a given variable, each such item can be scored in a uniform manner. With five response categories, scores of 0 to 4 or 1 to 5 might be assigned, taking the direction of the items into account (for example, assign a score of 5 to "strongly agree" for positive items and to "strongly disagree" for negative items). Each respondent would then be assigned an overall score representing the summation of the scores he or she received for responses to the individual items.

The Likert method is based on the assumption that an overall score based on responses to the many items reflecting a particular variable under consideration provides a reasonably good measure of the variable. These overall scores are not the final product of index construction; rather, they are used in an *item analysis* to select the *best* items. Essentially, each item is correlated with the large, composite measure. Items that correlate highest with the composite measure are assumed to provide the best indicators of the variable, and only those items would be included in the index ultimately used for analyses of the variable.

It should be noted that the uniform scoring of Likert-item response categories assumes that each item has about the *same intensity* as the rest. That is the key respect in which the Likert method differs from scaling as the term is used in this book.

You should also realize that Likert-type items can be used in a variety of ways: You

are by no means bound to the method described. Such items can be combined with other types of items in the construction of simple indexes; similarly, they can be used in the construction of scales. However, if all the items being considered for inclusion in a composite measure are in the Likert format, then the method I described should be considered.

Semantic Differential

As we've seen, Likert-type items ask respondents to agree or disagree with a particular position. The semantic-differential format asks them to choose between two opposite positions. Here's how it works.

Suppose you are conducting an experiment to evaluate the effectiveness of a new music appreciation lecture on subjects' appreciation of music. Let's say that you have created experimental and control groups as described in Chapter 8. Now you want to play some musical selections and have the subjects report their feelings about them. A good way to tap those feelings would be to use a semantic differential format.

To begin, you must determine the *dimensions* along which each selection should be judged by subjects. Then you need to find two *opposite* terms, representing the polar extremes along each dimension. Let's suppose one dimension that interests you is simply whether subjects enjoyed the piece or not. Two opposite terms in this case could be "enjoyable" and "unenjoyable." Similarly, you might want to know whether they regarded the individual selections as "complex" or "simple," "harmonic" or "discordant," and so forth.

Once you have determined the relevant dimensions and have found terms to represent the extremes of each, you might prepare a rating sheet to be completed by each subject for each piece of music. Figure 7-5 shows an example of what it might look like.

On each line of the rating sheet, the subject would indicate how he or she felt about the piece of music: whether it was enjoyable or unenjoyable, for example, and whether it was "somewhat" that way or "very much" so. To avoid creating a biased pattern of responses to such items, it's a good idea to vary the placement of terms that are likely to be related to each other. Notice, for example, that "discordant" and "traditional" are on the left side of the sheet, with "harmonic" and "modern" on the right side. Very likely, those selections scored as "discordant" would also be scored as "modern" as opposed to "traditional."

Both the Likert and semantic differential formats have a greater rigor and structure than other question formats. Despite common references to Likert scales and semantic differential scales, these formats produce

Figure 7-5 Semantic Differential: Feelings about Musical Selections

	Very Much	Somewhat	Neither	Somewhat	Very Much	
Enjoyable	☐	☐	☐	☐	☐	Unenjoyable
Simple	☐	☐	☐	☐	☐	Complex
Discordant	☐	☐	☐	☐	☐	Harmonic
Traditional	☐	☐	☐	☐	☐	Modern
			etc.			

data suitable to both indexing and scaling, as the latter terms are distinguished from one another in this chapter.

Now we'll turn our attention from the creation of cumulative indexes to an examination of scaling techniques. Although many methods of scaling are available, I'm going to limit our discussion to three: the Bogardus, Thurstone, and Guttman scales.

Scale Construction

Good indexes provide an ordinal ranking of cases on a given variable. All indexes are based on this kind of assumption: A senator who voted for seven conservative bills is considered to be more conservative than one who only voted for four of them. What an index may fail to take into account, however, is that not all indicators of a variable are equally important or equally strong. The first senator might have voted in favor of seven mildly conservative bills, whereas the second senator might have voted in favor of four extremely conservative bills. (The second senator might have considered the other seven bills too liberal and voted against them.)

Scales offer more assurance of ordinality by tapping *structures* among the indicators. The several items going into a composite measure may have different *intensities* in terms of the variable. The three scaling procedures described will illustrate the variety of techniques available.

Bogardus Social Distance Scale

Let's suppose you are interested in the extent to which Americans are willing to associate with, say, Albanians. You might ask the following questions:

1. Are you willing to permit Albanians to live in your country?
2. Are you willing to permit Albanians to live in your community?
3. Are you willing to permit Albanians to live in your neighborhood?
4. Would you be willing to let an Albanian live next door to you?
5. Would you let your child marry an Albanian?

Note that the several questions increase in the closeness of contact the respondents may or may not want with Albanians. Beginning with the original concern to measure willingness to associate with Albanians, we have developed several questions indicating differing degrees of intensity on this variable. The kinds of items presented constitute a Bogardus social distance scale.

The clear differences of intensity suggest a structure among the items. Presumably if a person is willing to accept a given kind of association, he or she would be willing to accept all those preceding it in the list—those with lesser intensities. For example, the person who is willing to permit Albanians to live in the neighborhood will surely accept them in the community and the nation but may or may not be willing to accept them as next-door neighbors or relatives. This, then, is the logical structure of intensity inherent among the items.

Empirically, one would expect to find the largest number of people accepting co-citizenship and the fewest accepting intermarriage. In this sense, we speak of "easy items" (for example, residence in the United States) and "hard items" (for example, intermarriage). More people agree to the easy items than to the hard ones. With some inevitable exceptions, logic demands that once a person has refused a relationship presented in the scale, he or she will also refuse all those harder ones that follow it.

The Bogardus social distance scale illustrates the important economy of scaling as a data-reduction device. By knowing *how many* relationships with Albanians a given respondent will accept, we know *which* relationships were accepted. Thus, a single number can accurately summarize five or six data items without a loss of information.

Thurstone Scales

Often the inherent structure of the Bogardus social distance scale is not appropriate to the variable being measured. Indeed, such a logical structure among several indicators is seldom apparent. Thurstone scaling is an attempt to develop a format for generating groups of indicators of a variable that have at least an *empirical* structure among them. One of the basic formats is that of "equal-appearing intervals."

A group of judges is given perhaps a hundred items felt to be indicators of a given variable. Each judge is then asked to estimate how strong an indicator of a variable each item is—by assigning scores of perhaps 1 to 13. If the variable were prejudice, for example, the judges would be asked to assign the score of 1 to the very weakest indicators of prejudice, the score of 13 to the strongest indicators, and intermediate scores to those felt to be somewhere in between.

Once the judges have completed this task, the researcher examines the scores assigned to each item by all the judges to determine which items produced the greatest agreement among the judges. Those items on which the judges disagreed broadly would be rejected as ambiguous. Among those items producing general agreement in scoring, one or more would be selected to represent each scale score from 1 to 13.

The items selected in this manner might then be included in a survey questionnaire. Respondents who appeared prejudiced on those items representing a strength of 5 would then be expected to appear prejudiced on those having lesser strengths, and if some of those respondents did not appear prejudiced on the items with a strength of 6, it would be expected that they would also not appear prejudiced on those with greater strengths.

If the Thurstone scale items were adequately developed and scored, the economy and effectiveness of data reduction inherent in the Bogardus social distance scale will appear. A single score might be assigned to each respondent (the strength of the hardest item accepted), and that score would adequately represent the responses to several questionnaire items. And as is true of the Bogardus scale, a respondent scored 6 might be regarded as more prejudiced than one scored 5 or less.

Thurstone scaling is not often used in research today, primarily because of the tremendous expenditure of energy required for the judging of items. Ten to 15 judges would have to spend a considerable amount of time to score the many initial items. Because the quality of their judgments would depend on their experience with and knowledge of the variable under consideration, the task might require professional researchers. Moreover, the meanings conveyed by the several items indicating a given variable tend to change over time. Thus an item having a given weight at one time might have quite a different weight later on. For a Thurstone scale to be effective, it would have to be periodically updated.

Guttman Scaling

A very popular scaling technique used by researchers today is the one developed by Louis Guttman. Like both Bogardus and Thurstone scaling, Guttman scaling is based on the fact that some items under consideration may prove to be harder indicators of the variable than others. One example should suffice to illustrate this pattern.

In the earlier example of measuring scientific orientations among medical school faculty members, you'll recall that a simple

The material that should have appeared on this page was missing in the original--if your professor is able to find the material for this page elsewhere you can add it to your coursepack by trimming off the edges of the replacement page then taping that page over top of this page.

the minimum we can hope for in a mixed type pattern. In the first mixed type, for example, we would erroneously predict a scientific response to the easiest item for each of the 18 respondents in this group, making a total of 18 errors.

The extent to which a set of empirical responses form a Guttman scale is determined by the accuracy with which the original responses can be reconstructed from the scale scores. For each of the 427 respondents in this example, we will predict three questionnaire responses, for a total of 1,281 predictions. Table 7-3 indicates that we will make 44 errors using the scale scores assigned. The percentage of correct predictions is called the coefficient of reproducibility: the percentage of original responses that could be reproduced by knowing the scale scores used to summarize them. In the present example, the coefficient of reproducibility is 1,237/1,281 or 96.6 percent.

Except for the case of perfect (100 percent) reproducibility, there is no way of saying that a set of items does or does not form a Guttman scale in any absolute sense. Virtually all sets of such items *approximate* a scale. As a rule of thumb, however, coefficients of 90 or 95 percent are the commonly used standards in this regard. If the observed reproducibility exceeds the level you've set, you will probably decide to score and use the items as a scale.

The decision concerning criteria in this regard is, of course, arbitrary. Moreover, a high degree of reproducibility does not ensure that the scale constructed in fact measures the concept under consideration, although it increases confidence that all the component items measure the same thing. Also, you should realize that a high coefficient of reproducibility is more likely when few items are involved.

One concluding remark should be made with regard to Guttman scaling: It is based on the structure observed among the *actual data under examination*. This is an important point that is often misunderstood. It does not make sense to say that a set of questionnaire items (perhaps developed and used by a previous researcher) constitutes a Guttman scale. Rather, we can say only that they form a scale within a given body of data being analyzed. Scalability, then, is a sample-dependent, empirical question. Although a set of items may form a Guttman scale among one sample of survey respondents, for example, there is no guarantee that they will form such a scale among another sample. In this sense, then, a set of questionnaire items in and of themselves never form a scale, but a set of empirical observations may.

Typologies

We shall conclude this chapter with a short discussion of typology construction and analysis. Recall that indexes and scales are constructed to provide ordinal measures of given variables. We attempt to assign index or scale scores to cases in such a way as to indicate a rising degree of prejudice, religiosity, conservatism, and so forth. In such cases, we are dealing with single dimensions.

Often, however, the researcher wishes to summarize the intersection of two or more variables. You may, for example, wish to examine the political orientations of newspapers separately in terms of domestic issues and foreign policy. The fourfold presentation in Table 7-4 describes such a typology.

Table 7-4 A Political Typology of Newspapers

	Foreign Policy	
Domestic Policy	Conservative	Liberal
Conservative	A	B
Liberal	C	D

Newspapers in cell A of the table are conservative on both foreign policy and domestic policy; those in cell D are liberal on both. Those in cells B and C are conservative on one and liberal on the other. (For purposes of analysis, each cell type might be presented by a data card punch [A = 1, B = 2, C = 3, D = 4] and could be easily manipulated in examining the typology's relationship to other variables.)

Frequently, you arrive at a typology in the course of an attempt to construct an index or scale. The items that you felt represented a single variable appear to represent two. We might have been attempting to construct a single index of political orientations for newspapers but discovered—empirically—that foreign and domestic politics had to be kept separate.

In any event, you should be warned against a difficulty inherent in typological analysis. Whenever the typology is used as the *independent variable*, there will probably be no problem. In the preceding example, you might compute the percentages of newspapers in each cell that normally endorse Democratic candidates; you could then easily examine the effects of both foreign and domestic policies on political endorsements.

It is extremely difficult, however, to analyze a typology as a *dependent variable*. If you want to discover why newspapers fall into the different cells of typology, you're in trouble. That becomes apparent when we consider the ways in which you might construct and read your tables. Assume, for example, that you want to examine the effects of community size on political policies. With a single dimension, you could easily determine the percentages of rural and urban newspapers that were scored conservative and liberal on your index or scale. With a typology, however, you would have to present the distribution of the urban newspapers in your sample among types A, B, C, and D. Then you would repeat the procedure for the rural ones in the sample and compare the two distributions. Let us suppose that 80 percent of the rural newspapers are scored as type A (conservative on both dimensions) as compared with 30 percent of the urban ones. Moreover, suppose that only 5 percent of the rural newspapers are scored as type B (conservative only on domestic issues) as compared with 40 percent of the urban ones. It would be incorrect to conclude from an examination of type B that urban newspapers are more conservative on domestic issues than rural ones, since 85 percent of the rural newspapers, compared with 70 percent of the urban ones, have this characteristic. The relative sparsity of rural newspapers in type B is due to their concentration in type A. It should be apparent that an interpretation of such data would be very difficult in anything other than description.

In reality, you would probably examine two such dimensions separately, especially if the dependent variable has more categories of responses than the example given.

Don't think that typologies should always be avoided in social research; often they provide the most appropriate device for understanding the data. You should be warned, however, against the special difficulties involved in using typologies as dependent variables.

Main Points

- Single indicators of variables seldom have sufficiently clear validity to warrant their use.
- Composite measures, such as scales and indexes, solve this problem by including several indicators of a variable in one summary measure.
- Both scales and indexes are intended as

ordinal measures of variables, though scales typically satisfy this goal better than indexes.

- Indexes are based on the simple cumulation of indicators of a variable.
- Scales take advantage of any logical or empirical intensity structures that exist among the different indicators of a variable.
- Face validity is the first criterion for the selection of indicators to be included in a composite measure; the term means that an indicator seems, on face value, to provide some measure of the variable.
- If different items are indeed indicators of the same variable, then they should be related empirically to one another. If, for example, frequency of church attendance and frequency of prayer are both indicators of religiosity, then those people who attend church frequently should be found to pray more than those who attend church less frequently.
- Once an index or a scale has been constructed, it is essential that it be validated. Internal validation refers to the relationship between individual items included in the composite measure and the measure itself. External validation refers to the relationships between the composite measure and other indicators of the variable—indicators *not* included in the measure.
- Likert scaling is a measurement technique based on the use of standardized response categories (for example, strongly agree, agree, disagree, strongly disagree) for several questionnaire items. Likert-format items may be used appropriately in the construction of either indexes or scales.
- The Bogardus social distance scale is a device for measuring the varying degrees to which a person would be willing to associate with a given class of people, such as an ethnic minority. Subjects are asked to indicate whether or not they would be willing to accept different kinds of association. The several responses produced by these questions can be adequately summarized by a single score, representing the closest association that is acceptable, since those willing to accept a given association also would be willing to accept more distant ones.
- Thurstone scaling is a technique for creating indicators of variables that have a clear intensity structure among them. Judges determine the intensities of different indicators.
- Guttman scaling is probably the most popular scaling technique in social research today. It is a method of discovering and using the empirical intensity structure among several indicators of a given variable.
- A coefficient of reproducibility is a measure of the extent to which all the particular responses given to the individual items included in a scale can be reproduced from the scale score alone.
- A typology is a nominal composite measure often used in social research. Typologies may be used effectively as independent variables, but interpretation is difficult when they are used as dependent variables.

Review Questions and Exercises

1. In your own words, describe the difference between an index and a scale.

2. Make up three questionnaire items—measuring attitudes toward nuclear power—that would probably form a Guttman scale.

Additional Readings

Glock, Charles; Ringer, Benjamin; and Babbie, Earl, ***To Comfort and to Challenge: A Dilemma of the Contemporary Church*** (Berkeley: University of California Press, 1967). An empirical study illustrating composite measures. Since the construction of scales and indexes can be most fully grasped through concrete examples, this might be a useful study to examine. The authors use a variety of composite measures, and they are relatively clear about the methods used in constructing them.

Lazarsfeld, Paul; Pasanella, Ann; and Rosenberg, Morris (eds.), ***Continuities in the Language of Social Research*** (New York: Free Press, 1972), especially Section I. An excellent collection of conceptual discussions and concrete illustrations. The construction of composite measures is presented within the more general area of conceptualization and measurement.

Miller, Delbert, *Handbook of Research Design and Social Measurement* (New York: Longman, 1983). An excellent compilation of frequently used and semistandardized scales. The many illustrations reported in Part 4 of the Miller book may be directly adaptable to studies or at least suggestive of modified measures. Studying the several different illustrations, moreover, may also give a better understanding of the logic of composite measures in general.

Oppenheim, A. N., *Questionnaire Design and Attitude Measurement* (New York: Basic Books, 1966). An excellent presentation on composite measures, with special reference to questionnaires. Although Oppenheim says little about index construction, he gives an excellent presentation of the logic and the skills of scale construction—the kinds of scales discussed in Chapter 7 of the present book and many not discussed here.