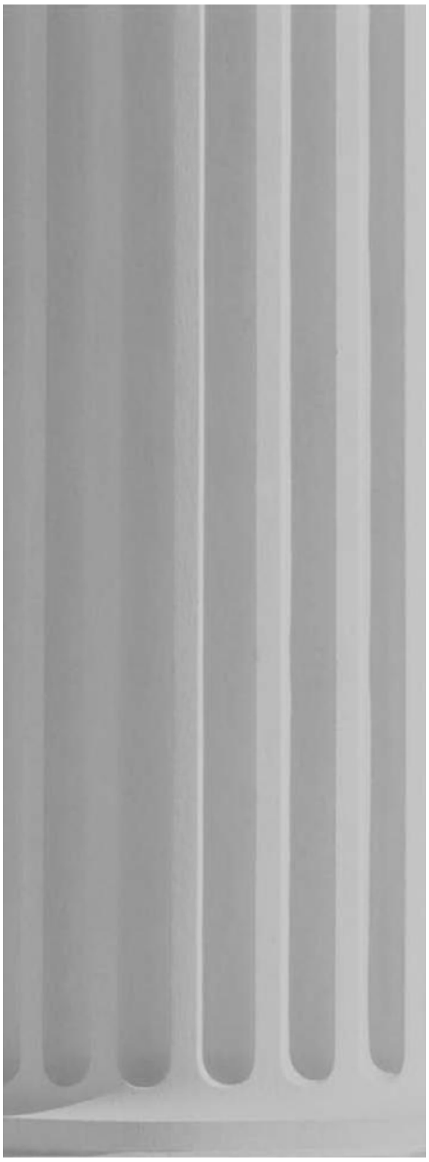
CHAPTER 5

Scales and Indexes

KEY TERMS

dichotomous response scale
Guttman scaling
index
interval response scale
Likert scaling
response scale
scaling
semantic differential
Thurstone scaling
weighted index



OUTLINE

In this chapter, I discuss the two most common approaches for creating quantitative measures of a construct: scaling and indexes. The terms *scale* and *index* are difficult to distinguish, and there are conflicting views in social research about how they should be defined and distinguished. Indexes typically combine different variables into a single score. Often, the variables combined are very different types of constructs and may even be measured in very different ways. So, an index tends to be a composite of differing elements. I consider the basic process in constructing an index and assessing its quality. Scaling evolved from the need to measure abstract or subjective constructs that may seem to be unmeasurable, such as attitudes and beliefs. I discuss general issues in scaling, including the distinction between a scale and a response format. I also explain the difference between multidimensional and unidimensional scaling. Finally, I look in depth at three types of unidimensional scales: Thurstone, Likert, and Guttman. From these discussions, you should learn not only how to use indexes and scales but also when each type is most appropriate.

5-1 Indexes

index
A quantitative score that measures a construct of interest by applying a formula or a set of rules that combines relevant data.

An **index** is a quantitative score constructed by applying a set of rules to combine two or more variables to reflect a more general construct. So, what does this mean? First, an index is a score, a numerical value, that purportedly measures something. Second, an index is a composite. It puts different variables together. Often, these variables are very different kinds of things and may even be measured in different ways and on different scales. Third, the variables are put together using a rule or set of rules. Sometimes, the rule is as simple as just adding up or averaging the scores of each variable to get a total index score. Sometimes, the rule is actually a formula or set of procedures for describing how the variables are combined. Finally, we usually construct an index because we want to measure something that none of the individual components alone does a good job of measuring. An index score is typically trying to get at something that cuts across the variables that are combined, that is more general than its composite parts.

5-1a Some Common Indexes

You are probably already familiar with several famous indexes. One of the best known is the consumer price index (CPI), which is collected every month by the Bureau of Labor Statistics of the U.S. Department of Labor (U.S. Department of Labor, 2004). Each month, the CPI index is reported and is considered to be a reflection of generally how much consumers have to pay for things. To construct this single score each month, government analysts identified eight major categories of spending for the typical consumer: food and beverages, housing, apparel, transportation, medical care, recreation, education and communication, and other goods and services. They then break down these eight areas into more than 200 specific categories. For each of these, they sample from the many items that reflect each category. For example, to represent the "apple" category that is in the "food and beverages" area, they might sample a "particular plastic bag of golden delicious apples, U.S. extra fancy grade, weighing 4.4 pounds" (U.S. Department of Labor, 2004). Each month, people call all over the country to get the current price for more than 80,000 items. Through a rather complicated weighting scheme that takes into account things like the location and the probability that the item will be purchased, these prices are combined. That is, there is a series of formulas and rules that are used each month to combine the prices into an index score. Actually, they compute thousands of different CPI scores each month to reflect different groups of consumers and different locations, although one of these is typically reported in the news as the CPI. The CPI is considered an index of consumer costs and, therefore, is a general economic indicator. It illustrates one of the most important reasons for creating an index—to track a phenomenon and its ups and downs over time.

A second well-known type of index is the socioeconomic status index (SES). Unlike the CPI, SES almost always involves the combination of several very different types of variables. Traditionally, SES is a combination of three constructs: income, education, and occupation. Income would typically be measured in dollars. Education might be measured in years or degree achieved. And occupation typically would be classified into categories or levels by status. Then, these very different elements would need to be combined to get the SES score. In one of the early classic studies in this area (Duncan, 1981), the researchers used the degree to which education and income predicted occupation as the basis for constructing the index score. This SES measure is now typically referred to as the *Duncan socioeconomic index* (SEI). For this index, an SEI score has been created for each of hundreds of occupations. The score is a weighted combination of "occupational education" (the percentage of people in that occupation who had 1 year or more of college education) and "occupational income" (the percentage of people in the occupation who earned more than a specific annual income). With the SEI, all you need to know is the occupation of a person, and you can look up the SEI score that presumably reflects the status of the occupation as related to both education and income. Almost from its inception, the measurement of socioeconomic status has been controversial, and different researchers attempt to accomplish it in a variety of ways (Hauser & Warren, 1996; Stevens & Cho, 1985).

5-1b Constructing an Index

Several steps are typically followed in constructing an index. I'll go over them briefly here, but you should know that, in practice, each one of these steps is considerably more complex than I'm able to convey in this brief description. Each step can involve sophisticated methods and considerable effort, when accomplished well. Here are the basic steps:

1. *Conceptualize the index.* It probably won't surprise you that the first thing you need to decide is what you would like the index to measure. This may seem like a simple issue at first. However, for almost anything you would like to measure with an index, different people might reasonably disagree about what it means. What is socioeconomic status? Does it include income, education, and occupation? If you measure education and occupation, won't that be highly related to income? If so, do you need a separate component that reflects income or will just the two components be sufficient? If you were trying to measure a construct like "quality of life," what components would you need to include to capture the construct? Even with a well-established measure such as the CPI, researchers worry about defining basic terms like "consumption by whom" and "prices of what"? To begin composing an index, you need first to identify the construct you are trying to reflect in the index and describe the variables that are components of the construct. There are a wide variety of ways to accomplish this step. You can make it up using your own hunches and intuitions (a surprisingly large amount of social

research happens this way). You can review the literature and use current theory as a guide. You can engage experts or key stakeholders in formal processes for conceptualizing, using approaches like brainstorming, concept mapping, or interviewing to identify what the key concept you are trying to measure means to different people. Think about several conceptual issues at this stage. What is the purpose of the index? How will it be used and by whom? Is this a one-time or short-term measure, or one that you would like to use over a long period?

2. *Operationalize and measure the components.* It is one thing to say that you would like to measure education and occupation as major components of socioeconomic status. It is quite another to figure out how to measure each one. If you are trying to measure education as it relates to status, do you simply count the number of years in school? If two people spend the same number of years in college majoring in very different subjects, should they get the same numerical value on educational status? Or should we give more "points" for someone majoring in one field than another. Should all bachelor's degrees be counted the same? If you are trying to look at occupation as it relates to status, how do you even classify occupations? How do you decide the numerical value for each occupation as it relates to status? Over time, do occupations change in status? Are new occupations created? (There weren't any Web programmers before the Internet!) If so, how do you accommodate this in an index that tries to measure changes in status over years or decades? What is the unit for which you are measuring? Are you measuring educational levels of individuals? Or are you looking at some other unit like the community or an organization? For example, if you want to measure the educational level of a community and you know the number of years a representative sample of community members went to school, it may be reasonable simply to average the number of years for the community estimate. But if you have a coding only by level of education (for example, 1 = grade school, 2 = some high school, 3 = some college, 4 = associates degree, 5 = bachelor's degree, etc.), you cannot average these values. In this case, you may need to calculate the proportion of the community that achieved a particular level (for example, the proportion of high school graduates) as an estimate of the community educational level. In any event, you need to figure out how you will measure each component of an index before you can move on to calculating the composite index score.
3. *Develop the rules for calculating the index score.* Once you have the components that you think make up the construct of interest, you need to figure out how to combine these component scores to create a single index score. There are many complications here. In the simplest case, you might be able to combine the component scores just by adding or averaging them. In essence, this is what the CPI does. This can be done for the CPI because each consumer item is measured in the same way—its price. But what if each component is measured in entirely different ways? In SES measures, you can't measure income the same way you measure education. So, you're not likely to be able to add or average the scores for income and education in any straightforward way. Even if the components are measured in a similar manner, what if you think different components should be given different emphasis in measuring the construct. For example, what if you believe that income should be considered more important than education when trying to measure socioeconomic status? How much more important? You can do several things when combining the components of an index. It helps if you can develop a model of the index that shows the index score, each of the components, and how you think theoretically these are related. You then need to develop precise rules for how to combine the components in the model. Sometimes, these rules can be stated as a set of procedures that you follow to compute the index, almost like a recipe. In other cases, the rules are essentially a formula or set of formulas (a simple average of several components is essentially a formula).
4. *Are you giving each component equal weight or you are constructing a weighted index score?* A **weighted index** is one where you combine different components of the

weighted index
A quantitative score that measures a construct of interest by applying a formula or a set of rules that combines relevant data where the data components are weighted differently.

index in different amounts or with different emphasis. You're almost certainly familiar with a weighted index because most of you have probably had a teacher that at one time or another used a weighting scheme to come up with your grade for a class. For example, imagine that your teacher measures you on three characteristics: test scores, class participation, and a class project. For the sake of argument (and this is no simple matter in itself), let's assume that you are scored on each of those on a 0 to 100 scale. If you score perfectly on all your tests, you get a 100 on the test score; if your project is perfect, you get 100 on the project score. One way to get a total index score for your course performance would be to average these three components. But what if you (or, more to the point, your professor) believe that these components should not receive equal weight? For instance, maybe participation should be weighted only half as much as the test or project component. You might reason that it doesn't matter how much you participate as long as you can do well on the tests and project. To construct this index score, you would need a formula that weights the test and project components twice as high as the participation. Here's one:

$$\text{Performance} = [(2 \times \text{Test}) + (2 \times \text{Project}) + (1 \times \text{Participation})]/5$$

Why divide by 5? I want the final index score to be on a scale of 1 to 100. Notice that the idea of weighting in index construction can get rather confusing (so what's new?). For example, your professor could have measured both your test and project performance on a 1 to 40 scale (where best performance gets a 40) and your attendance on a 1 to 20 scale. Then, to construct your index score, you might simply add the three component scores! It looks like this is not a weighted index, and technically, it isn't because you're simply adding the scores. But the truth is that you built the weighting into the measurement of each component.

5. *Validate the index score.* Once you have constructed the index, you will need to validate it. This is essentially accomplished in the same way any measure is validated (see Section 3-1, Construct Validity). If the index score is going to be used over time, it is especially important to do periodic validation studies because it's quite possible that the components or how they relate to the index score have changed in important ways over time. For instance, the classification of occupations today differs in important ways from classifications used in 1950. And the selection of consumer goods continually changes over time. Consequently, indexes like the CPI and SES have to be recalibrated or adjusted periodically if they are to be valid reflections of the construct of interest.

Indexes are essential in social research. They range from formal, complex, sophisticated national indexes that track phenomena over years or decades to simple measures developed for use in a single study (or to compute your grade for a course!).

5-2 Scaling

Scaling is the branch of measurement that involves the construction of a measure based on associating qualitative judgments about a construct with quantitative metric units. Like an index, a scale is typically designed to yield a single numerical score that represents the construct of interest. In many ways, scaling remains one of the most mysterious and misunderstood aspects of social research measurement. It attempts to do one of the most difficult of research tasks—measure abstract concepts.

scaling
The branch of measurement that involves the construction of an instrument that associates qualitative constructs with quantitative metric units.

Most people don't understand what scaling is. The basic idea of scaling is described in Section 5-2a, General Issues in Scaling. The discussion includes the important distinction between a scale and a response format. Scales are generally divided into two broad categories: unidimensional and multidimensional. The unidimensional scaling methods were developed in the first half of the twentieth century and are generally named after their creators. We'll look at three types of unidimensional scaling methods here:

- Thurstone or equal-appearing interval scaling
- Likert or summative scaling
- Guttman or cumulative scaling

In the late 1950s and early 1960s, measurement theorists developed advanced techniques for creating multidimensional scales. Although these techniques are outside the scope of this text, an understanding of the most common unidimensional scaling methods will provide a good foundation for these more complex variations.

5-2a General Issues in Scaling

S. S. Stevens (1946) came up with what I think is the simplest and most straightforward definition of scaling. He said, "Scaling is the assignment of objects to numbers according to a rule."

What does that mean? In most scaling, the objects are text statements—usually, statements of attitude or belief. In Figure 5–1, three statements describe attitudes toward immigration. To scale these statements, you have to assign numbers to them. Usually, you would like the result to be on at least an interval scale (see Section 3-3, Levels of Measurement), as indicated by the ruler in the figure. What does "according to a rule" mean? If you look at the statements, you can see that as you read down, the attitude toward immigration becomes more restrictive; if people agree with a statement on the list, it's likely that they will also agree with all of the statements higher on the list. In this case, the rule is a cumulative one. So, what is scaling? It's how you get numbers that can be meaningfully assigned to objects; it's a set of procedures. The following paragraphs introduce several approaches to scaling.

First, I have to clear up one of my pet peeves. People often confuse the idea of a scale and a response scale. A **response scale** is the way you collect responses from people on an instrument. You might use a **dichotomous response scale** like Agree/Disagree, True/False, or Yes/No, or you might use an **interval response scale** like a 1 to 5 or 1 to 7 rating. However, if all you are doing is attaching a response scale to an object or statement, you can't call that *scaling*. As you will see, scaling involves procedures that you perform independently of the respondent so that you can

response scale
A sequential numerical response format, such as a 1-to-5 rating format.

dichotomous response scale
A question with two possible responses. The better term to use is dichotomous response *format*.

interval response scale
A response measured on an interval level, where the size of the interval between potential response values is meaningful. Most 1-to-5 rating responses can be considered interval level. The better term to use is interval response *format*.

FIGURE 5–1 **Scaling as the assignment of numbers according to a rule**

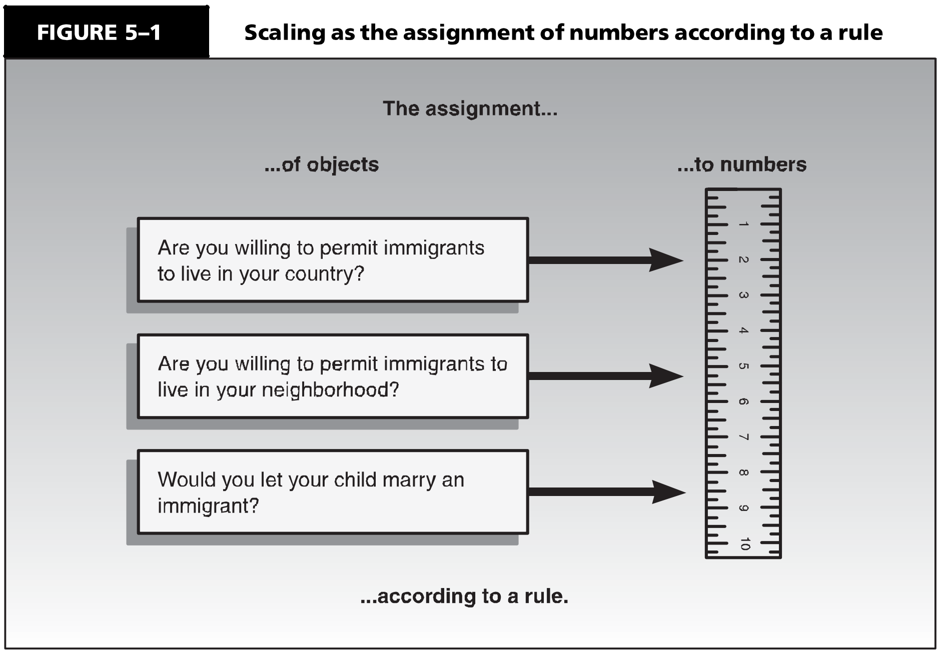

The assignment...
...of objects
...to numbers
Are you willing to permit immigrants to live in your country?
Are you willing to permit immigrants to live in your neighborhood?
Would you let your child marry an immigrant?
1
2
3
4
5
6
7
8
9
10
...according to a rule.

TABLE 5–1 **Differences between Scaling and Response Scales**

Scale	Response Scale
Results from a process	Used to collect the response for an item
Each item on a scale has a scale value	Item not associated with a scale value
Refers to a set of items	Used for a single item

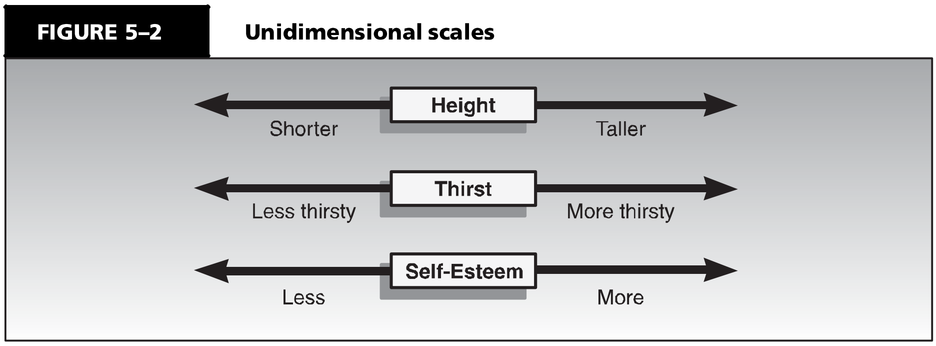
come up with a numerical value for the object. In true scaling research, you use a scaling procedure to develop your instrument (scale) and a response scale to collect the responses from participants. Simply assigning a 1 to 5 response scale for an item is *not* scaling! The differences are illustrated in Table 5–1.

Also, it's important to realize that although a scale is an instrument that can be used alone, it is often integrated into a larger and more complex instrument, such as a survey. Many surveys are designed to assess multiple topics of interest and to collect data that enables us to study the interrelationships of these topics. In many surveys, we will embed one or more scales as separate sections of the survey. When the data is collected, the analyst will compute the various scale scores by combining the responses to items from each scale according to the rules for that scale. However, just because a survey asks a set of questions on a single topic and asks you to respond on a similar response scale (such as a 1 to 5 disagree-agree response scale), you cannot automatically conclude that the set of questions constitute a scale. Got that? Please reread that sentence until you're sure you get the distinction! The set of questions on a survey cannot be considered a scale unless a scaling process was followed to identify the questions and determine how the responses would be combined. So, just because a set of questions on a survey looks like a scale, collects data using the same response scale, and is even analyzed like a scale, it isn't a real scale unless some type of scaling process was used to create it. I'll present several of the most famous scaling processes later in this chapter.

Purposes of Scaling Why do scaling? Why not just create text statements or questions and use response formats to collect the answers? First, sometimes, you do scaling to test a hypothesis. You might want to know whether the construct or concept is a single dimensional or multidimensional one (more about dimensionality later). Sometimes, you do scaling as part of exploratory research. You want to know what dimensions underlie a set of ratings. For instance, if you create a set of questions, you can use scaling to determine how well they hang together and whether they measure one concept or multiple concepts. But perhaps the most common reason we do scaling is similar to why we construct indexes: we would like to represent a construct using a single score. When a participant gives responses to a set of items, you often want to assign a single number that represents that person's overall attitude or belief. In Figure 5–1, we would like to be able to give a single number that describes a person's attitudes toward immigration, for example. Scaling is a formal procedure that helps you construct a set of items that can achieve this.

Dimensionality A scale can have any number of dimensions in it. Most scales that researchers develop have only a few dimensions. What's a dimension? Think of a dimension as a number line, as illustrated in Figure 5–2. If you want to measure a construct, you have to decide whether the construct can be measured well with one number line or whether it may need more. For instance, height is a concept that is unidimensional, or one-dimensional. You can measure the concept of height well with only a single number line (a ruler). Weight is also unidimensional; you can measure it with a scale. Thirst might also be considered a unidimensional concept; you are either more or less thirsty at any given time. It's easy to see that height and

FIGURE 5–2 Unidimensional scales



weight are unidimensional, but what about a concept like self-esteem? If you think you can measure a person's self-esteem well with a single ruler that goes from low to high, you probably have a unidimensional construct.

What would a two-dimensional concept be? Many models of intelligence or achievement postulate two major dimensions: mathematical and verbal ability. In this type of two-dimensional model, a person can be said to possess two types of achievement, as illustrated in Figure 5–3. Some people will be high in verbal skills and lower in math. For others, it will be the reverse. If a concept is truly two-dimensional, it is not possible to depict a person's level on it by using only a single number line. In other words, to describe achievement, you would need to locate a person as a point in two-dimensional (*x*, *y*) space, as shown in Figure 5–3.

semantic differential
A scaling method in which the respondent assesses an object on a set of bipolar adjective pairs.

Okay, let's push this one step further: How about a three-dimensional concept? Psychologists who study the idea of meaning theorized that the meaning of a term could be well described in three dimensions. Put in other terms, any objects can be distinguished or differentiated from each other along three dimensions. They labeled these three dimensions activity, evaluation, and potency. They called this general theory of meaning the **semantic differential**. Their theory essentially states that you can rate any object along those three dimensions. For instance, think of the idea of ballet. If you like the ballet, you would probably rate it high on activity, favorable on evaluation, and powerful on potency. On the other hand, think about the concept of a book like a novel. You might rate it low on activity (it's passive), favorable on evaluation (assuming you like it), and about average on potency. Now, think of the idea of going to the dentist. Most people would rate it low on activity (it's a passive activity), unfavorable on evaluation, and powerless on potency. (Few routine activities make you feel as powerless!) The theorists who came up with the idea of the semantic differential thought that the meaning of any concepts could be described well by rating the concept on these three dimensions. In other words, to describe the meaning of an object, you have to locate it as a dot somewhere within the cube (three-dimensional space), as shown in Figure 5–4.

FIGURE 5–3 A two-dimensional scale

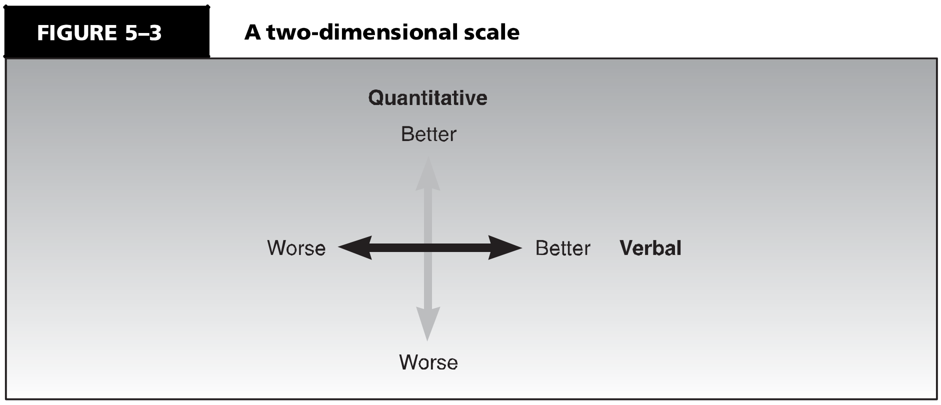
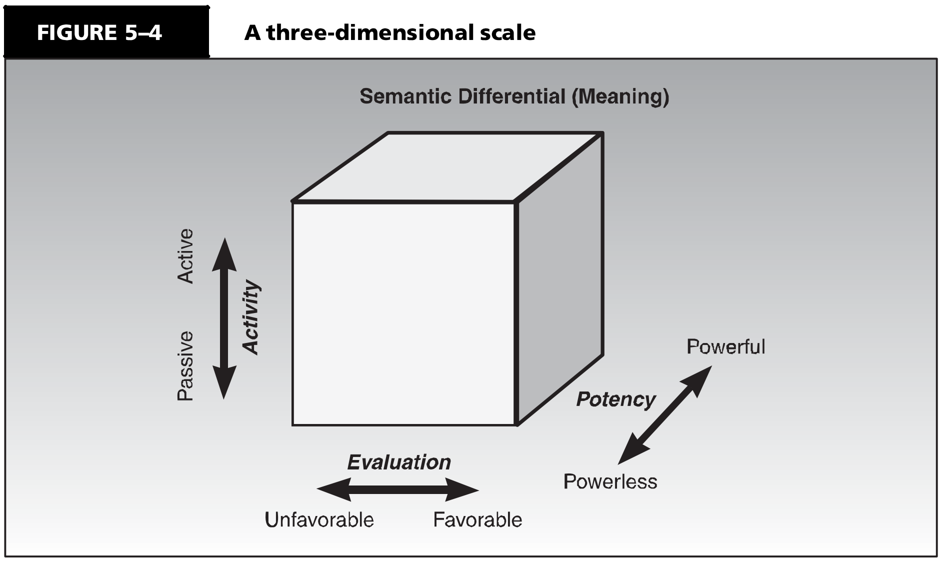


FIGURE 5–4 A three-dimensional scale



Unidimensional Versus Multidimensional What are the advantages of using a unidimensional model? Unidimensional concepts are generally easier to understand. You have either more or less of it, and that's all. You're either taller or shorter, heavier or lighter. It's also important to understand what a unidimensional scale is as a foundation for comprehending the more complex multidimensional concepts. But the best reason to use unidimensional scaling is that you believe the concept you are measuring is unidimensional in reality. As you've seen, many familiar concepts (height, weight, temperature) are actually unidimensional. However, if the concept you are studying is, in fact, multidimensional in nature, a unidimensional scale or number line won't describe it well. If you try to measure academic achievement on a single dimension, you would place every person on a single line, ranging from low to high achievers. How would you score someone who is a high math achiever and terrible verbally, or vice versa? A unidimensional scale can't capture that more general type of achievement; you would need at least two unidimensional scales.

There are three major types of unidimensional scaling methods. They are similar in that they each measure the concept of interest on a number line. However, they differ considerably in how they arrive at scale values for different items. The three methods are Thurstone, or equal-appearing interval scaling; Likert, or summative scaling; and Guttman, or cumulative scaling. Each of these approaches is described in the following sections.

5-2b Thurstone Scaling

Thurstone was one of the first and most productive scaling theorists. He actually invented three different methods for developing a unidimensional scale, which can be considered different ways to do **Thurstone scaling**: the *method of equal-appearing intervals*, the *method of successive intervals*, and the *method of paired comparisons*. The three methods differed in how the scale values for items were constructed, but in all three cases, respondents rated the resulting scale the same way. To illustrate Thurstone's (1925) approach, I'll show you the easiest method of the three to implement: the method of equal-appearing intervals.

Thurstone scaling
The process of developing a scale in which the scale items have interval-level numerical values where the final score is the average scale value of all items with which the respondent agreed.

Developing the focus. The method of equal-appearing intervals starts like almost every other scaling method—with a large set of statements to which people respond. Oops! I did it again! You can't start with the set of statements; you have to first define the focus for the scale you're trying to develop. Let this be a warning to all of you: Methodologists like me often start our descriptions with the first objective, methodological step (in this case, developing a set of statements) and forget to mention critical foundational issues like the development of the focus for a project. So, let's try this again. . . .

The method of equal-appearing intervals starts like almost every other scaling method—with the development of the focus for the scaling project. Because this is a unidimensional scaling method, you have to be able to assume that the concept you are trying to scale is reasonably thought of as one-dimensional. The description of this concept should be as clear as possible so that the person(s) who will create the statements has a clear idea of what you are trying to measure. I like to state the focus for a scaling project in the form of an open-ended statement to give to the people who will create the draft or candidate statements. You want to be sure that everyone who is generating statements has some idea of what you are after in this focus command. You especially want to be sure that technical language and acronyms are spelled out and understood.

Generating potential scale items. In this phase, you're ready to create statements. Who should create the statements for a scale? That depends. You might have experts who know something about the phenomenon you are studying. Because the people affected are likely to be expert about what they're experiencing, you might sample them to generate statements. For instance, if you are trying to create a scale for quality of life for people who have a certain type of health condition, you might want to ask them to create potential items. Finally, you can make up the items. Obviously, each of these approaches has advantages and disadvantages, so in many situations, you may want to use some or all of them.

You want a large set of candidate statements—usually, as many as 80 to 100—because you are going to select your final scale items from this pool. You also want to be sure that all of the statements are worded similarly—that they don't differ in grammar or structure. For instance, you might want them each to be worded as a statement with which respondents agree or disagree. You don't want some of them to be statements, while others are questions.

Rating the scale items. So, now you have a set of items or statements. The next step is to have a group of people called *judges* rate each statement on a 1 to 11 scale in terms of how much each statement indicates a *favorable* attitude toward the construct of interest. Pay close attention here! You *don't* want the judges to tell you what their attitudes on the statements are, or whether they would agree with the statements. You want them to rate the favorableness of each statement in terms of the construct you are trying to measure, where 1 = extremely unfavorable attitude toward the construct and 11 = extremely favorable attitude towards the construct. One easy way to actually accomplish this is to type each statement on a separate index card and have each judge rate them by sorting them into 11 piles, as shown in Figure 5–5. Who should the judges be? As with generating the items, there is no simple answer. Generally, you want to have people who are "experts"

FIGURE 5–5 **Rating statements on a 1 to 11 scale by sorting them manually**

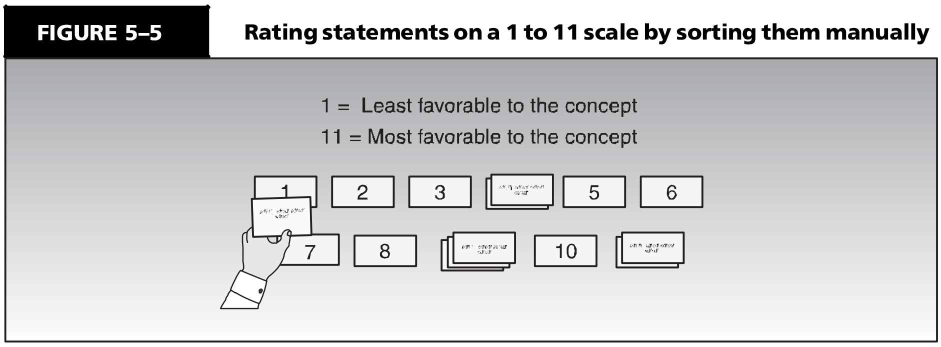
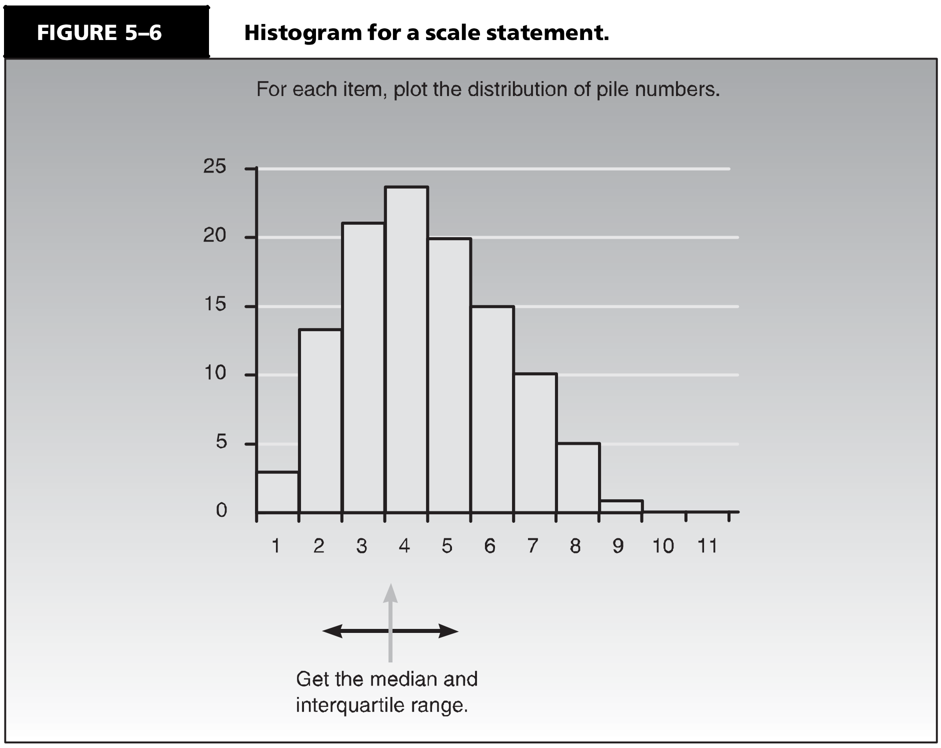

1 = Least favorable to the concept
11 = Most favorable to the concept
1
2
3
5
6
7
8
10

FIGURE 5–6 **Histogram for a scale statement.**



on the construct of interest do this. But there are many kinds of expertise, ranging from academically trained and credentialed experts to the people who are most directly experienced with the phenomenon.

Computing scale score values for each item. The next step is to analyze the rating data. For each item or statement, you need to compute the median and the interquartile range. The *median* is the value above and below which 50 percent of the ratings fall. The first quartile (Q1) is the value below which 25 percent of the cases fall and above which 75 percent of the cases fall—in other words, the 25th percentile. The median is the 50th percentile. The third quartile, Q3, is the 75th percentile. The interquartile range is the difference between third and first quartile, or Q3–Q1. Figure 5–6 shows a histogram for a single item and indicates the median and interquartile range.

You can compute these values easily with any introductory statistics program or with most spreadsheet programs. To facilitate the final selection of items for your scale, you might want to sort the table of medians and interquartile ranges in ascending order by median and, within that, in descending order by interquartile range.

Selecting the final scale items. Now you have to select the final statements for your scale. You should select statements that are at equal intervals across the range of medians. Ideally, one statement would be selected for each of the 11 median values. Within each value, you should try to select the statement that has the smallest interquartile range (the statement with the least amount of variability across judges). You don't want the statistical analysis to be the only deciding factor here. Look over *the candidate statements at each level, and select the statement that makes the most* sense. If you find that the best statistical choice is a confusing statement, select the next best choice.

Administering the scale. You now have a scale—a yardstick you can use for measuring the construct of interest. Each of your final scale items has a scale score—the median value. And the item scores should range across the spectrum of potential attitudes or beliefs on this construct (because you selected items throughout the median range). You can now give the final set of items to respondents and ask them to agree or disagree with each statement. To get an individual's final scale score, average only the scale scores of all the items that person agreed with. When you average the scale items for the statements with which the respondent agreed, you get an average score that has to range between 1 and 11. If they agreed with scale items that were low in favorableness to the construct, then the average of the items they agreed to should be low. If they agreed with items that your judges had said were highly favorable to the construct, then their final score will be on the higher end of the scale.

You should see a couple of things from this discussion. First, you use the judges to create your scale. Think of the scale as a ruler that ranges from 1 to 11 with one scale item or statement at each of the 11 points on the ruler. Second, when you give the set of scale items to a respondent and ask them to tell you which ones they agree with, you are essentially trying to measure them with that ruler. Their scale score—where you would mark the individual on your 11-point ruler—is the average item value for the items the respondent agreed with.

The other Thurstone scaling methods—the method of successive intervals and the method of paired comparisons—are similar to the method of equal-appearing intervals. All of them begin by focusing on a concept that is assumed to be unidimensional and involve generating a large set of potential scale items. All of them result in a scale consisting of relatively few items that the respondent rates on an Agree/Disagree basis. The major differences are in how the data from the judges is collected. For instance, the method of paired comparisons requires each judge to make a judgment about each pair of statements. With lots of statements, this can become time-consuming.

5-2c Likert Scaling

Likert scaling
The process of developing a scale in which the ratings of the items are summed to get the final scale score. Ratings are usually done using a 1 to 5 Disagree-to-Agree response format Likert scales are also sometimes called summated scales.

Like Thurstone or Guttman scaling, **Likert scaling** (Murphy & Likert, 1938) is a unidimensional scaling method. Here, I'll explain the basic steps in developing a Likert or summative scale. You may remember learning the term *Likert scale* in Chapter 4. A Likert scale is a type of response scale and is different from Likert scaling (see the discussion in Section 5-2a, General Issues in Scaling, and Table 5–1 for the differences between response scales and scaling).

Defining the focus. As in all scaling methods, the first step is to define what it is you are trying to measure. Because this is a unidimensional scaling method, it is assumed that the concept you want to measure is one-dimensional in nature. You might operationalize the definition as an instruction to the people who are going to create or generate the initial set of candidate items for your scale.

Generating the items. Next, you have to create the set of potential scale items. These should be items that can be rated on a 1 to 5 or 1 to 7 disagree-agree response scale. Sometimes, you can create the items by yourself based on your intimate understanding of the subject matter. More often than not, though, it's helpful to engage a number of people in the item creation step. For instance, you might use some form of brainstorming to create the items. It's desirable to have as large a set of potential items as possible at this stage; about 80 to 100 would be best.

Rating the items. The next step is to have a group of judges rate the items. Usually, you would use a 1 to 5 rating scale where:

1 = Strongly unfavorable to the concept
2 = Somewhat unfavorable to the concept
3 = Undecided
4 = Somewhat favorable to the concept
5 = Strongly favorable to the concept

Notice that, as in other scaling methods, the judges are not telling you what they believe; they are judging how favorable each item is with respect to the construct of interest.

Who should the judges be? As in any scaling method, that's not an easy question to answer. Some argue that experts familiar with the process should be used. Others suggest that you should use a random sample of the same types of people who are ultimately your respondents of interest for the scale. There are advantages and disadvantages to each.

Selecting the items. The next step is to compute the intercorrelations between all pairs of items, based on the ratings of the judges. In making judgments about which items to retain for the final scale, there are several analyses you can perform:

- Throw out any items that have a low correlation with the total (summed) score across all items. In most statistics packages, it is relatively easy to compute this type of item-total correlation. First, you create a new variable that is the sum of all of the individual items for each respondent. Then, you include this variable in the correlation-matrix computation. (If you include it as the last variable in the list, the resulting item-total correlations will all be the last line of the correlation matrix and will be easy to spot.) How low should the correlation be for you to throw out the item? There is no fixed rule here; you might eliminate all items that have a correlation with the total score less than .6, for example. (The idea of correlation is covered in Section 12-3d, Correlation.)
- For each item, get the average rating for the top quarter of judges and the bottom quarter. Then, do a *t*-test of the differences between the mean value for the item for the top and bottom quarter judges. (An in-depth discussion of *t*-tests appears in Chapter 12.) Higher *t*-values mean that there is a greater difference between the highest and lowest judges. In more practical terms, items with higher *t*-values are better discriminators, so you want to keep these items. In the end, you will have to use your judgment about which items are most sensibly retained. You want a relatively small number of items on your final scale (from 10 to 15), and you want them to have high item-total correlations and high discrimination (that is, high *t*-values).

Administering the scale. You're now ready to use your Likert scale. Each respondent is asked to rate each item on some response scale. For instance, respondents could rate each item on a 1 to 5 response scale where:

1 = Strongly disagree
2 = Disagree
3 = Undecided
4 = Agree
5 = Strongly agree

There are a variety of possible response scales (1 to 7, 1 to 9, 0 to 4). All of these odd-numbered scales have a middle value, which is often labeled *neutral or undecided.* It is also possible to use a forced-choice response scale with an even number of responses and no middle neutral or undecided choice. In this situation, respondents are forced to decide whether they lean more toward the "agree" or "disagree" end of the scale for each item.

The final score for the respondent on the scale is the sum of his or her ratings for all of the items. (This is why this is sometimes called a *summated scale.*) On some scales, you will have items that are reversed in meaning from the overall direction of the scale. These are called *reversal items.* You will need to reverse the response value for each of these items before summing for the total. That is, if the respondent gave a 1, you make it a 5; if a respondent gave a 2, you make it a 4; 3 = 3; 4 = 2; and 5 = 1. Researchers disagree about whether you should have a "neutral" or

TABLE 5–2 **The Employment Self-Esteem Likert Scale**

Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	1. I feel good about my work on the job.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	2. On the whole, I get along well with others at work.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	3. I am proud of my ability to cope with difficulties at work.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	4. When I feel uncomfortable at work, I know how to handle it.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	5. I can tell that other people at work are glad to have me there.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	6. I know I'll be able to cope with work for as long as I want.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	7. I am proud of my relationship with my supervisor at work.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	8. I am confident that I can handle my job without constant assistance.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	9. I feel like I make a useful contribution at work.
Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	10. I can tell that my coworkers respect me.

"undecided" point on the scale (an odd number of responses) or whether the response scale should be a "forced choice" one with no neutral point and an even number of responses (as in a 1 to 4 scale).

Table 5–2 shows an example of a hypothetical ten-item Likert scale that attempts to estimate the level of self-esteem (Rosenberg, 1965) a person has on the job. Notice that this instrument has no center or neutral point in the response scale; the respondent has to declare whether he or she is in agreement or disagreement with the item.

5-2d Guttman Scaling

Guttman scaling
The process of developing a scale in which the items are assigned scale values that allow them to be placed in a cumulative ordering with respect to the construct being scaled.

Guttman scaling (Guttman, 1950) is also sometimes known as *cumulative scaling* or *scalogram analysis*. In Chapter 4, I introduced the term *Guttman scale* in Section 4-1a, Types of Questions. A Guttman scale is a type of response scale and is different from Guttman scaling (see the discussion in 5-2a, General Issues in Scaling, and Table 5–1 for the differences between response scales and scaling). The purpose of Guttman scaling is to establish a one-dimesional continuum for a concept you want to measure. What does that mean? Essentially, you would like a set of items or statements so that a respondent who agrees with any specific question in the list will also agree with al previous questions. Put more formally, you would like to be able to predict item responses perfectly knowing only the total score for the respondent. For example, imagine a ten-item cumulative scale. If the respondent scores a 4, it should mean that he or she agreed with the first four statements. If the respondent scores an 8, it should mean he or she agreed with the first eight. The object is to find a set of items that perfectly matches this pattern. In practice, you would seldom expect to find this cumulative patern perfectly. So, you use scalogram analysis to examine how closely a set of items corresponds with this idea of cumulativeness. Here, I'll explain how you develop a Guttman scale.

Define the focus. As in all of the scaling methods, you begin by defining the focus for your scale. Let's imagine that you want to develop a cumulative scale that measures U.S. citizen attitudes toward immigration. You would want to be sure to specify in your definition whether you are talking about any type of immigration (legal and illegal) from anywhere (Europe, Asia, Latin and South America, Africa).

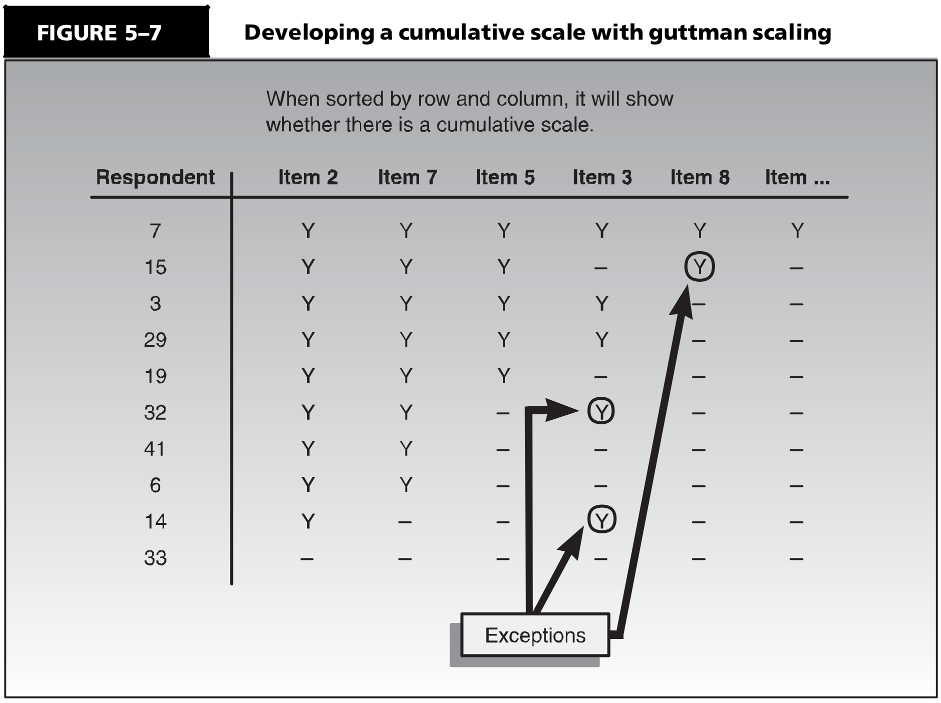
Develop the items. Next, as in all scaling methods, you would develop a large set of items that reflect the concept. You might do this yourself, or you might engage a knowledgeable group to help. Of course, as with all scaling methods, you would want to come up with many more statements (about 80 to 100 is desirable) than you will ultimately need.

Rate the items. Next, you would want to have a group of judges rate the statements or items in terms of how favorable they are to the concept of interest. They would give a *Yes* if the item is favorable toward the construct and a *No* if it is not. Notice that you are not asking the judges whether they personally agree with the statement. Instead, you're asking them to make a judgment about how the statement is related to the construct of interest.

Develop the cumulative scale. The key to Guttman scaling is in the analysis. You construct a matrix or table that shows the responses of all the judges on all of the items. You then sort this matrix so that judges who agree with more statements are listed at the top and those who agree with fewer are at the bottom. For judges with the same number of agreements, sort the statements from left to right from those that most agreed to, to those that fewest agreed to. You might get a table something like the one in Figure 5–7. Notice that the scale is nearly cumulative when you read from left to right across the columns (items). Specifically, a person who agreed with item 7 always agreed with item 2. Someone who agreed with item 5 always agreed with items 7 and 2. The matrix shows that the cumulativeness of the scale is not perfect, however. While, in general, a person agreeing with item 3 tended to also agree with 5, 7, and 2, there are several exceptions to that rule.

Although you can examine the matrix if there are only a few items in it, if there are many items, you need to use a data analysis called *scalogram analysis* to determine the subsets of items from the pool that best approximate the

FIGURE 5–7 **Developing a cumulative scale with guttman scaling**



When sorted by row and column, it will show whether there is a cumulative scale.

Respondent	Item 2	Item 7	Item 5	Item 3	Item 8	Item ...
7	Y	Y	Y	Y	Y	Y
15	Y	Y	Y	–	Ⓨ	–
3	Y	Y	Y	Y	–	–
29	Y	Y	Y	Y	–	–
19	Y	Y	Y	–	–	–
32	Y	Y	–	Ⓨ	–	–
41	Y	Y	–	–	–	–
6	Y	Y	–	–	–	–
14	Y	–	–	Ⓨ	–	–
33	–	–	–	–	–	–

Exceptions

cumulative property. Then, you review these items and select your final scale elements. There are several statistical techniques for examining the table to find a cumulative scale. Because there is seldom a perfectly cumulative scale, you usually have to test how good it is. These statistics also estimate a scale score value for each item. This scale score is used in the final calculation of a respondent's score.

Administering the scale. After you've selected the final scale items, it's relatively simple to administer the scale. You simply present the items and ask respondents to check items with which they agree.

Each scale item has a scale value associated with it (obtained from the scalogram analysis). To compute a respondent's scale score, you simply sum the scale values of every item the respondent agrees with. In this example, the final value should be an indication of the respondent's view on the construct of interest.

5-3 Indexes and Scales

At this point, you should have a much clearer sense of how indexes and scales are similar to and different from each other. One clear commonality is that both an index and a scale yield a single numerical score or value that is designed to reflect the construct of interest.

But there are lots of ways in which scales and indexes are different. Indexes very often are used to combine component scores that differ greatly from one another and are measured in different ways, like income, occupation, and education in SES. Scales typically involve rating a set of similar items on the same response scale, as in the 1 to 5 Likert response format. Indexes often combine numerical values that are counts or are more objectively observable (like prices). Scales very often are constructed to get at more subjective and judgmental constructs like attitudes or beliefs.

Needless to say, there's considerable disagreement among researchers about whether and how indexes and scales can be defined and distinguished. Some researchers argue that a scale is a particular type or subset of an index. Others argue that they are very different things altogether. Some maintain that a unique feature of scaling is the sophistication of the methodology used to select the items; others contend that good index development can get as sophisticated and advanced as any scaling procedure. And so it goes. However we define them, it should be clear to you that both indexes and scales are essential tools in social research.

Summary

A lot of territory was covered in this chapter. We began by learning about indexes. We briefly looked at two of the most famous indexes: the consumer price index (CPI) and socioeconomic status (SES). I then went through the basic steps for how to construct an index score. Next, I showed you what a scale is and described the basic univariate scale types: Thurstone, Likert, and Guttman. You saw that scales can be used as stand-alone instruments, but they can also be integrated into a larger survey. Based on this chapter, you should have a feel for what would be involved in creating and using either an index or scale. The next chapter introduces you to several very different forms of measurement—qualitative and unobtrusive—that aren't geared to generating a single score like an index or scale does, but that are at least as important for social research.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.