TABLE 6.7 Construction of a Crime Index

Crime rates per 100,000 People, 1995		Crime Index, 1995
Homicide	8.2	
Rape	37.1	
Robbery	221	
Assault	418	5,278
Burglary	988	
Larceny-theft	3,045	
Motor vehicle theft	561	

Source Federal Bureau of Investigation, *Uniform Crime Reports.*

regarding what constitutes a serious crime; for the criminologists who developed the index, serious crimes are those that pose the greatest and most direct threat to personal safety, property, and public order. Such crimes as gambling, prostitution, and commercialized vice are not included in the index because they were judged to pose a considerably weaker threat to the public order.

A second thing to note is that the seven items—the seven index crimes—constitute the universe of elements that make up the more abstract concept of "serious crimes." In fact, the concept of "serious crimes" is defined by its operational definition: A serious crime is one of the Part I offenses. This is often the case with indexes: The items used in their construction are all the elements or constructs that make up the broader phenomenon.

A third thing to note about the Crime Index is the relationship between the index and the items that measure it: the values of each item determine the level of the variable measured by the index, rather than the other way around. When the incidence of rape or robbery rises, this causes the level of the index to rise. However, causality does not flow in the other direction: Changes in the index cannot occur first and account for changes in the individual items. For example, a rise in serious crimes does not cause changes in the amounts of robbery that occur; in fact, the index (serious crimes) could rise while the robbery rate falls, as long as some of the other crimes in the index rise. With an index, changes in the items produce changes in the index, rather than the other way around. This is a characteristic of many, although not all, indexes. When we discuss scales, we will see that they involve something different (DeVellis, 1991). The Applied Scenario box discusses the development and use of an index for an applied research problem.

In developing indexes, it is often the case that the separate items are not given equal weighting as was the case with the crime index. Theoretical considerations may suggest that some indicators of a variable are more important in determining the state of the variable. This is the case with some indexes that have been developed to measure socioeconomic status (SES). Most measures of SES use a combination of two or more indicators, such as occupation, income, and education. A number of years ago, sociologists Seymour Parker and Robert Kleiner (1966), building on earlier work, developed an SES measure, described in Table 6.8, that combined all three. Since education, occupation, and income are measured in different units, they need to be transformed into common measurement units. As you can see, this was done by dividing the variables in a way such that each had seven values. Parker and Kleiner did this based on theoretical considerations of the variables as well as on the empirical distribution of persons into the various categories. For example, the occupational categories were created in part based on the relative prestige of various occupations, trying to place occupations with similar prestige ratings into the same or close categories. Then Parker and Kleiner were confronted with the decision of how much to weight each indicator of SES. Some SES measures give them equal weight. Based on theoretical considerations regarding socioeconomic status, Parker and Kleiner decided that education

TABLE 6.8 An Index of Socioeconomic Status (SES)

Education (Years Completed)	Item Value	Annual Income	Item Value	Occupation	Item Value
0–4 years	1	$0–1,000	1	Unskilled workers	1
5–8 years	2	$1,001–2,000	2	Sales personnel, semi-skilled workers	2
9–11 years	3	$2,001–3,000	3	Skilled craftsmen and clerical, minor government workers	3
12 years (high school graduate)	4	$3,001–4,000	4	Minor administrative, supervisors, office managers	4
13–15 years	5	$4,001–5,000	5	Minor professionals, medical technicians, teachers	5
16 years (college graduate)	6	$5,001–6,000	6	Major administrative, managers	6
17 years or more	7	$6,001 and over	7	Major professionals (doctors, university professors)	7

Source Reprinted and adapted with the permission of The Free Press, a Division of Simon & Schuster, Inc. from *Mental Illness in the Urban Negro Community* by Seymour Parker and Robert J. Kleiner. Copyright © 1966 by The Free Press.

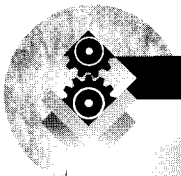
was by far the most important determinant of a person's SES and occupation least important. After lengthy analysis, they settled on the following weights: 4.4 for education, 2.5 for income, and 1.0 for occupation. Then, each individual's SES level is determined by the following formula:

$$\frac{(\textit{education value} \times 4.4) + (\textit{income value} \times 2.5) + (\textit{occupation value} \times 1.0)}{3}$$

So, a person who had received item values of 6 for education, 4 for income, and 5 for occupation, would have an SES index score of 13.8. In this case, the index of socioeconomic status is a weighted average of the values of the three separate indicators. Weighting of indexes can be done in ways that differ from this; for example, it might be a weighted total, rather than an average. But this illustration gives you an example of index construction using weights.

Evaluating Indexes

Indexes are subjected to the various kinds of measurement assessments (discussed in Chapter 5). In particular, indexes would need to pass the various assessments of validity and reliability that all measuring devices are subjected to. In addition, however, there are comparisons that can be made among the items themselves in order to assess particular items and an overall index. Especially important is something called *item analysis*. With some indexes, an assessment can be made by looking at the correlations among the various items that make up the index. The basic principle is that there should be a fairly steady and strong, although not perfect, relationship between the items of the index if all the items are good and the index is valid. If the items are at the interval-ratio level of measurement, then a correlation between each pair of items can be calculated; if the items are at the ordinal level, then a contingency table could be utilized to see if there is a relationship (see Chapter 13 and



APPLIED SCENARIO **Measuring the Success of Teenage Pregnancy Prevention Programs**

Teenage pregnancies are a serious problem in the United States because most teenagers are ill equipped to be effective parents and the demands of parenting often limit the educational and occupational opportunities available to teenage parents. Programs to prevent teenage pregnancies and to help teenage parents are common across the country, and applied social scientists are often enlisted to provide systematic assessments of how well these programs work. Social scientists Marilyn Fernandez and Holly Ruch-Ross (1998) evaluated a number of such programs in Illinois, and their research illustrates ways in which simple multiple-item indexes can be used to measure variables in applied research.

Teenage pregnancy programs typically pursue a number of goals. Virtually all such programs, for example, strive to prevent, or at least postpone, future pregnancies, and they also typically try to enhance the self-sufficiency of teenage parents by supporting them in school or assisting them in getting and keeping a job. These goals are the dependent variables that applied researchers attempt to measure. Fernandez and Ruch-Ross used these goals to develop a composite index of success, which they called a Result Score. The unit of analysis (see Chapter 4) in this research was the organization, in particular the social service or other agencies that were administering the pregnancy prevention program. The researchers wanted to distinguish the successful from the unsuccessful agencies. The Result Score was the agency's measure of success and was a simple index composed of only two items. One item was whether the rate of repeat pregnancy among clients of an agency was below the average for all the agencies. The second item was whether the clients of an agency had a school attendance rate or employment rate that was higher than the average for all agencies. Success was indicated by an agency's having a below average repeat pregnancy rate or an above average school attendance or employment rate. An agency received a score of 1 for each success; so, the index could take on three values: 2 (if an agency was successful by both measures), 1 (if successful by one measure but not the

Chapter 14 on contingency tables). We could also look at the relationship between more than two items at a time by conducting an appropriate multivariate statistical analysis.

When an index is used to measure some subjective state or attitude, then another type of assessment can be made: comparing an individual's response to each item to the results of the overall index. When people's responses to individual index items is presumed to be caused by the underlying variable, then each item should correlate with the results of the overall scale. For example, persons who score high on a self-esteem index should tend to choose the high-esteem alternative of an item that makes up that index. Any item that shows no relationship or a negative relationship would have to be assessed very carefully in terms of whether it is a good item for that index.

However, this kind of item analysis is appropriate only for some indexes; namely, those in which the values of the items that make up the index are caused by the value of the underlying phenomenon being measured. This would

other), and 0 (if successful by neither measure). Like other indexes, the Result Score measured an abstract concept, "success," by combining measures of two more concrete phenomena: repeat pregnancy rates and education/employment rates.

Fernandez and Ruch-Ross (1998) found some expected and some unexpected outcomes in their study. Not surprisingly, they found that better funded agencies had better success rates. However, unexpectedly, they found that agencies that devoted more hours of service to their clients had lower success rates! This anomalous finding will certainly motivate researchers and agencies to try to figure out what is going on, and this is exactly the point of doing such evaluations—to improve how agencies provide services.

Most of the indexes and scales discussed in Chapter 6 are based on survey questions asked of individuals. The Fernandez and Ruch-Ross index is different in that it uses the organization as the unit of analysis and the agency records as the source of data in constructing the index, rather than survey questions. (Agency records are a form of what is called "available data" and is discussed in greater detail in Chapter 10.) It is important to recognize that indexes and scales can be developed from virtually any sort of data, not just data based on survey questions or statements. The key to indexes and scales as measuring devices is not the source of the data but rather that one is constructing a composite score by using multiple indicators of a phenomenon.

Identify some organization or agency with which you have contact (your university, your place of employment, some social service agency) and consider ways that you could develop indexes and scales from the data they collect.

1. Identify what data these organizations collect and which variables they measure.
2. Develop some ways that these data could serve as the items for a multiple-item index or scale. Which abstract concepts do these indexes and scales measure?

be true of a self-esteem index where a person's underlying self-esteem level is what causes his or her responses to the items of the index. In other indexes, such as the index of SES, this is not the case, and we would not necessarily expect to see all items being correlated with the overall index.

More sophisticated statistical techniques exist for evaluating the items that make up indexes and scales. Going by such names as *factor analysis* and *Q-sort methodology*, these techniques use complex statistical procedures for deciding which items on an index or scale seem to be measuring a single dimension of the variable being studied. Basically, it involves correlating each item with the overall index and with each dimension or factor that emerges in the statistical analysis. The researcher can determine which items seem to correlate highly or cluster together and thus represent a single factor. These procedures can be used to decide whether a variable is unidimensional or whether it contains a number of distinct dimensions. It can also be used to assign weights to

the items of an index or scale and to eliminate items that don't contribute much to measuring a variable. These procedures are fairly complex and require a good grounding in basic statistics to fully understand them.

Scale Construction

A scale is made up of separate items or indicators, as is an index, but in a scale there is an intensity structure to the items. In addition, in scales, people tend to respond to the items in more of a pattern; people with a similar scale score show a more similar response pattern. Scaling can utilize a number of formats, and each format calls for some unique elements in its design.

Likert Scales

One of the most popular approaches to multiple-item measures is that developed by Rensis Likert (1932). A **Likert scale** consists of a series of statements, each followed by a series of response alternatives for the respondent to express himself or herself about the statement. An illustration of a Likert scale is presented in Table 6.6, with response alternatives ranging from "Strongly agree" to "Strongly disagree." Some Likert scales have the intensity structure built into the items in the scale. With the Conflict Tactics Scale, for example, some items are clearly a stronger or more intense expression of the variable being measured (see Table 6.3). So, "used a knife or gun on my partner" is a stronger or more intense form of conflict resolution than is "insulted or swore at my partner." Other Likert scales, such as the self-esteem scale in Table 6.6, have the intensity built into the response format, with the "Strongly agree" to "Strongly disagree" providing the intensity structure to each item (Anderson, Basilevsky, & Hum, 1983; Nunnally, 1978).

Likert originally developed his scale with the agree-disagree format for his alternatives, and some social scientists still maintain that a true Likert scale should contain those response alternatives. However, other response alternatives are often used today, such as strongly approve–strongly disapprove or very satisfied–very dissatisfied (see Table 6.4), and most researchers still consider them Likert scales (Alwin, 1997). In a Likert scale the most common number of alternatives is five because it offers respondents a sufficient range of choices without requiring unnecessarily minute distinctions in attitudes. More or fewer than five alternatives are sometimes used, but recall the research mentioned earlier that concludes that more rather than fewer alternatives make for more valid and reliable measures. The exact wording of the response alternatives in a Likert scale must be grammatically consistent with the wording in the statements (see Table 6.4).

Note in Table 6.6 the numbers ranging from 1 to 4 in brackets next to each response alternative. These numbers are included on the scale here for purposes of illustration only; they would not be printed on a scale for actual use because their presence might influence respondents' answers. The numbers are used when scoring the scale. The numbers associated with each response are totaled to provide the overall score for each respondent. In this case—a 10-item scale—individual scores can range from a low of 10 (if alternative 1 were chosen every time) to a high of 40 (if alternative 4 were chosen every time). Remember, as discussed in Chapter 5, each item in a Likert scale is an ordinal measure, ranging from a low of "Strongly disagree" to a high of "Strongly agree." Because the total score of a Likert scale is the sum of individual ordinal items, some researchers contend that a Likert scale is therefore ordinal in nature. However, other researchers maintain that the composite score produced by a Likert scale is actually at the interval level, and these researchers use interval-ratio statistics to analyze data produced by Likert scales.

The Likert scale is one example of scales known as **summated rating scales**, in which a person's score is determined by summing the

TABLE 6.9 Calculation of Discriminatory Power (DP) Score for One Item on a Scale

		Response Value							
Quartile	***N***	**1**	**2**	**3**	**4**	**5**	**Weighted Total**	**Weighted Mean**	**DP Score**
Upper	10	0	1	2	4	3	39	3.90	
Lower	10	2	8	0	0	0	18	1.80	2.10
								2.10	

number of questions answered in a particular way. We could, for example, ask respondents to agree or disagree with statements and then assign a 1 for each statement they agree with and a 0 for each disagreement. Their scale score is then the sum of their responses. Summated rating scales can take a number of different forms, although the Likert format is the most common.

Constructing a Likert scale, as with all scales, requires considerable time and effort. One begins by developing a series of statements relating to the variable being measured using the general criteria for statements outlined previously in the chapter. A common rule of thumb is to begin with three times the number of statements desired for the final scale since many of the statements will prove unacceptable for one reason or another and be deleted.

In deciding which items will ultimately be used in a Likert scale, an important criterion is whether the scale items *discriminate* among people. That is, we want people's responses to an item to range over the four or five alternatives rather than cluster on one or two choices. Imagine a scale with an item that reads: "Persons convicted of shoplifting should have their hands amputated." If such an item were submitted to a group of college students, it is likely that most would respond with "Strongly disagree" and maybe a few "Disagrees." It is highly unlikely that any would agree. Of what use is this item to us? We cannot *compare* people—assess who is more likely to agree or disagree—because they all disagree. We cannot correlate responses to this item with the social or psychological characteristics of the students because there is little or no variation in responses to the item.

For our scale, then, we want to eliminate nondiscriminating items from consideration. *Nondiscriminating items* are those that are responded to in a similar fashion both by people who score high and by people who score low on the overall scale. Nondiscriminating items on a scale can be detected on the basis of results from a pretest in which people respond to all the preliminary items of the scale. One way of identifying nondiscriminating items is by computing a **discriminatory power score** (DP score) for each item. The DP score essentially tells us the degree to which each item differentiates between respondents with high scores and respondents with low scores on the overall scale. The first step in obtaining DP scores is to calculate the total scores of each respondent and rank the scores from highest to lowest. We then identify the upper and lower quartiles of the distribution of total scores. The *upper quartile* (Q_3) is the cutoff point in a distribution above which the highest 25% of the scores are located; the *lower quartile* (Q_1) is the cutoff point below which the lowest 25% of the scores are located. With the quartiles based on *total* scores identified, we compare the pattern of responses to *each* scale item for respondents whose scores fall above the upper quartile with the pattern for respondents whose scores fall below the lower quartile. Table 6.9 illustrates the computation of DP scores for one item on a scale to which 40 persons responded. Ten respondents are above the

upper quartile, and 10 are below the lower quartile. It can be seen that the high scorers tended to agree with this item because most had scores of 4 or 5. Low scorers tended to disagree because they are totally concentrated in the 1 and 2 score range. The next step is to compute a weighted total on this item for the two groups. This is done by multiplying each score by the number of respondents with that score. For example, for those above the upper quartile, the weighted total is:

$$(1 \times 0) + (2 \times 1) + (3 \times 2) + (4 \times 4) + (5 \times 3) = 0 + 2 + 6 + 16 + 15 = 39.$$

Next, the weighted mean (average) is computed by dividing the weighted total by the number of cases in the quartile. For the upper quartile, we have

$$39/10 = 3.9.$$

The DP score for this item is then obtained by subtracting the mean of those below the lower quartile from the mean of those above the upper quartile. In this example, we have: $3.9 - 1.8 = 2.1$. This process is repeated for every item in the preliminary scale so that each item has a calculated DP score.

Once we have DP scores for all the preliminary items, final selection can begin. The best items are those with the *highest* DP scores because this shows that people in the upper and lower quartiles responded to the items very differently. As a rule of thumb, as many items as possible should have DP scores of 1.00 or greater, and few if any should drop below 0.50. Applying this rule to the item in Table 6.9, we could conclude that it is a very good item and would include it in the final scale. Under no circumstances should an item with a negative DP score be included because this means that high scorers on the overall scale scored lower on this item than did the low scorers. If the size of the negative DP score is small, it is probably an ambiguous statement that is being variously interpreted by respondents. If the negative DP score is large, however, it is possible that the item was accidentally misscored; that is, a negative item was scored as if it were positive or vice versa.

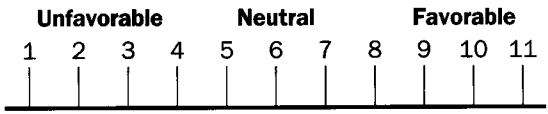
The Likert format is one of the most popular multiple-item formats because of the many advantages it possesses. First, it offers respondents a range of choices rather than the limited Yes-No alternatives used in some other scales. This makes Likert measures valuable if our theoretical assessment of the manifestations of a variable is that they range along a continuum rather than being either present or absent. Second, data produced by Likert-type scales are at least ordinal level and many consider them interval level, which enables us to use more powerful statistical procedures than with nominal-level data. Third, Likert measures are fairly straightforward to construct.

Likert scales have the same disadvantages as many other scales. In particular, one must be careful in interpreting a single score based on a Likert scale because it is a summary of so much information (separate responses to a number of items). Whenever data are summarized, some information is lost. (Your grade in this course is a summary measure of your performance, and in calculating it your instructor loses information regarding those high—or low—scores you received on individual exams.) The summary score might hide information about patterns of variation in responses or about possible multidimensionality of the scale.

Thurstone Scales

Another approach to scaling was developed by L. L. Thurstone and E. J. Chave (1929). **Thurstone scales** are constructed so that they use *equal-appearing intervals*—that is, it is assumed that the distance between any two adjacent points on the scale is the same. This feature, it is argued, provides some justification for treating the data as interval-level and using

FIGURE 6.1 Equal-Appearing Intervals as Used in Thurstone Scale Construction



all the powerful statistical procedures that require interval-level data.

Construction of a Thurstone scale begins much the same way as for Likert scales: with the selection of many statements that relate to the variable being measured. Once a sufficient number of statements is at hand, the next step is to provide a value between 1 and 11 for each statement. As illustrated in Figure 6.1, Thurstone scales utilize an 11-point scale ranging from 1 (the least favorable statement regarding an object, event, or issue) to 11 (the most favorable). Point 6 on the scale is labeled "neutral" and is used for statements that are neither favorable nor unfavorable. For example, the statement "Teenage girls who get pregnant are immoral" would be considered highly unfavorable toward teenage pregnancies.

The task of rating each statement as to how favorable or unfavorable it is with regard to the measured variable is assigned to a group of people known as "judges." With each of the preliminary statements printed on a separate card, the judges rate the items by placing them in piles corresponding to points on the 11-point scale. The judges place in each pile statements that they assess to be roughly equivalent in terms of their favorability. This use of judges affords some confidence that a Thurstone scale has the intensity structure among the items necessary to be considered a scale rather than an index.

Once the scale values are computed for all the preliminary items, the next step is to determine which items are the least ambiguous and therefore best for inclusion in the final scale. If the judges differed widely in their ratings of an item, it is likely something is unclear about the statement itself that leads to varying interpretations. Therefore the degree of agreement among judges about the rating of an item is used as one indicator of ambiguity.

Scales should include the items with the most agreement among judges, and there should be a roughly equal number of items for each of the 11 scale values ranging from unfavorable to favorable, moving upward in half-point increments. This would mean that a minimum of 21 items is required, although some argue that if reliability of .90 or better is desired, as many as 50 statements may be needed (Seiler & Hough, 1970). Regardless of the number actually used, the last step in Thurstone scale construction is to order the items randomly for presentation to respondents.

Table 6.10 presents the first 13 statements contained in the original 45-item Thurstone scale developed by Thurstone and Chave, with the scale value of each item indicated in parentheses. This particular scale is designed so that items with high scale values are "Unfavorable" toward the church, and items with low scale values are "Favorable." The scale values would not, of course, be included on a working version of the scale and are presented here for illustration only. Note that respondents are required only to check the statements with which they agree, making the Thurstone format particularly easy for respondents.

Scoring a Thurstone scale differs from the simple summation procedure used with Likert scales. Because the respondents will agree to differing numbers of statements with different values, the simple sum of the item values is worthless; two people could both agree with four statements, but these may be different statements at different levels of intensity, which would indicate quite different attitudes. Rather, a respondent's score is either the mean or median of the scale values of the items that the person agrees with. For example, if a person agreed with statements 2, 4, 8, and 12 (a

TABLE 6.10 Attitude Toward Church Scale

Check (✓) every statement below that expresses your sentiment toward the church. Interpret the statements in accordance with your own experience with churches.

(8.3)*	1. I think the teaching of the church is altogether too superficial to have much social significance.
(1.7)	2. I feel the church services give me inspiration and help me to live up to my best during the following week.
(2.6)	3. I think the church keeps business and politics up to a higher standard than they would otherwise tend to maintain.
(2.3)	4. I find the services of the church both restful and inspiring.
(4.0)	5. When I go to church, I enjoy a fine ritual service with good music.
(4.5)	6. I believe in what the church teaches but with mental reservations.
(5.7)	7. I do not receive any benefit from attending church services, but I think it helps some people.
(5.4)	8. I believe in religion, but I seldom go to church.
(4.7)	9. I am careless about religion and church relationships, but I would not like to see my attitude become general.
(10.5)	10. I regard the church as a static, crystallized institution, and as such it is unwholesome and detrimental to society and the individual.
(1.5)	11. I believe church membership is almost essential to living at its best.
(3.1)	12. I do not understand the dogmas or creeds of the church, but I find that the church helps me to be more honest and creditable.
(8.2)	13. The paternal and benevolent attitude of the church is quite distasteful to me.

* Scale value.

Source L. L. Thurstone and E. J. Chave, *The Measurement of Attitude.* Chicago: University of Chicago Press (1929). Used with permission of the University of Chicago Press.

total of four statements) in Table 6.10, that person's Thurstone scale score would be 3.13. Another person, agreeing with 1, 7, 10, and 13 (still four statements), would have a score of 8.18. This scoring procedure distributes respondents along the original 11-point scale.

Thurstone and Likert scaling techniques are essentially interchangeable methods of measuring attitudes. A major advantage of the Thurstone technique is that it provides interval-level data. However, if you claim that Likert scales are also interval level or if the interval-data properties are not needed, the Likert technique is probably preferable owing to its higher reliability with fewer items and its greater ease of construction. A second advantage of Thurstone scales is that people can respond to the items more quickly than with a Likert scale because they need only indicate whether they agree with an item and need not ponder to what degree they agree or disagree. However, because reliability calls for Thurstone scales to be longer, this advantage may be minimal. In fact, this can even become a disadvantage if the longer scale leads people to be overly quick or careless in responding to statements. Another major disadvantage of Thurstone scales is that they are costly and difficult to construct.

Semantic Differential Scales

Another scaling format, which has proved quite popular, is the semantic differential (SD) scale developed by Osgood, Suci, and Tannenbaum (1957). The **semantic differential** format presents the respondent with a stimulus, such as a person or event, that is to be rated on a

TABLE 6.11 Semantic Differential Scale Assessing Attitudes Toward the Elderly

	Scale							
Active	7	6	5	4	3	2	1	Passive
Competent	—	—	—	—	—	—	—	Incompetent
High IQ	—	—	—	—	—	—	—	Low IQ
Powerful	—	—	—	—	—	—	—	Weak
Healthy	—	—	—	—	—	—	—	Sickly
Secure	—	—	—	—	—	—	—	Insecure
Creative	—	—	—	—	—	—	—	Uncreative
Fast	—	—	—	—	—	—	—	Slow
Attractive	—	—	—	—	—	—	—	Ugly
Pleasant	—	—	—	—	—	—	—	Unpleasant
Reliable	—	—	—	—	—	—	—	Unreliable
Energetic	—	—	—	—	—	—	—	Lazy
Calm	—	—	—	—	—	—	—	Irritable
Flexible	—	—	—	—	—	—	—	Rigid
Educated	—	—	—	—	—	—	—	Uneducated
Generous	—	—	—	—	—	—	—	Selfish
Wealthy	—	—	—	—	—	—	—	Poor
Good memory	—	—	—	—	—	—	—	Poor memory
Involved	—	—	—	—	—	—	—	Socially isolated

Source William C. Levin, "Age Stereotyping: College Student Evaluations," *Research on Aging*, Vol 10 (March 1988), pp. 134–148.

scale between a series of polar opposite adjectives. Normally, the scale has 7 points, but scales can have fewer or more points if theoretical or methodological considerations call for it. Table 6.11 illustrates an SD designed to measure people's attitudes toward the elderly. In this study, college students were shown pictures of people of varying ages and then asked to describe the characteristics of each person by placing an X on the line between each adjective pair that best represented their assessment of the person. So, on the first line, placing an X over the 6 means that you view the person as quite active, whereas placing an X over the 1 is an assessment of very passive. In this example, all the positive adjectives are on the left and the negative adjectives are on the right. Sometimes the positive responses to some adjectives are put on the right in order to discourage disinterested respondents from placing all their responses in the same column.

Semantic differential responses are analyzed somewhat differently from Likert or Thurstone scales. First, the responses to the adjectives are investigated to determine if they reflect some underlying, more abstract, dimension or factor. The adjectives that make up each factor are presumed to be indicators of one underlying attitudinal dimension. Identification of these attitudinal dimensions, or factors, can be accomplished with a rather complex statistical procedure called *factor analysis*,

which basically correlates responses to each adjective pair with responses to every other adjective pair. For example, recent analyses of stereotypes toward elderly people using SDs suggest that the adjective pairs yield four factors (Intrieri, von Eye, & Kelly, 1995): acceptability (socially at ease and pleasing to others), instrumentality (vitality and active pursuit of goals), autonomy (self-sufficiency and active participation in life), and integrity (personal satisfaction and peacefulness with oneself). In Table 6.11, for example, the pairs Active–Passive, Powerful–Weak, and Energetic–Lazy would be some of the indicators of the instrumentality factor. One of the challenges of analyzing SDs is figuring out the nature of the abstract dimension that is reflected in a particular grouping of adjective pairs.

Once the factors being tapped by an SD have been determined, then the SD can be scored. One way to do this is to treat the response to each adjective pair separately. This approach is appropriate when the attitude dimension being tapped is validly measured by one item. Thus, if we were specifically interested in whether or not college students viewed the elderly as socially involved, the last item in Table 6.11 could be used as a measure of the variable. We could compare whether the average social involvement score given to the elderly differs from that given to other adults. Often, however, we are interested in one or more of the abstract dimensions which are more validly measured by a number of adjective pairs. For this to be accomplished, responses on the adjective pairs that constitute each dimension can be summed to provide an overall score on each of the dimensions measured—another variant of the summated ratings scale. For example, the college students who were given the SD in Table 6.11 consistently stereotyped the elderly as less instrumental than young adults.

Semantic differentials have several advantages when compared both to Likert and to Thurstone formats (Nunnally, 1978). Unlike the other scaling techniques that require 20 or more items for adequate reliability, SDs require only four to eight adjective pairs for each dimension to reach reliabilities of .80 or better. This brevity means that an SD can be filled out quickly (Heise, 1970; Miller, 1991). Another advantage is that SDs are much easier and less time-consuming to construct than either Likert or Thurstone scales. Adjective pairs are easier to develop than are unambiguous and unbiased statements about an issue. In addition, adjective pairs from prior studies are more readily adaptable to new studies because of their general and nonspecific nature. This is particularly important if a measuring scale is needed quickly. If, for example, we wanted people's reactions to some unanticipated event while it is still fresh in their minds, time would be of the essence. Only an SD-type scale could be readied in time.

About the only disadvantage of an SD is that identifying the abstract dimensions tapped by the adjective pairs is somewhat subjective and judgmental. The validity of the conclusions drawn is only as good as the judgment of those who identify the dimensions.

Guttman Scales

At the outset of this chapter, I noted that efforts are made to create scales that are unidimensional; that is, they measure a single variable or a single aspect of a variable. With a **Guttman scale**, the procedures used in construction give us the greatest confidence that the resulting scale is unidimensional (Guttman, 1944).

Researchers using Guttman scales achieve unidimensionality by developing the items in such a way that, in a perfect Guttman scale, there is only one pattern of responses that will yield any given score on the scale. (In fact, some argue that this is a characteristic of a true scale.) For example, if an individual's score is 5, we would expect that he or she had agreed with the first five items on the scale. This can be contrasted to other scaling techniques that allow obtaining the same score by agreeing or disagreeing with any number of items and having completely different response patterns.

TABLE 6.12 Response Patterns in a Guttman Scale

		Guttman Scale Patterns							
Response Alternatives		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Error Pattern**
Harder Items	Have person as close kin by marriage	No	No	No	No	No	No	Yes	No
	Have person in my club as personal friend	No	No	No	No	No	Yes	Yes	Yes
	Have person on my street as neighbors	No	No	No	No	Yes	Yes	Yes	No
	Have person working alongside me on my job	No	No	No	Yes	Yes	Yes	Yes	Yes
	Have person as citizen in my country	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Easier Items	Have person as visitor to my country	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Guttman scaling is able to do this because the items in the scale have an inherently progressive nature relating to the intensity of the variable being measured. In the parlance of Guttman scaling, the least intense items are referred to as "easy" because more people are likely to agree with them; the most intense items are considered "hard" because fewer are expected to agree with them. If a person agrees with a certain item, we would expect him or her also to agree with all the less intense items; conversely, if a person disagrees with a particular item, we would also expect that person to disagree with all the more intense items.

The Bogardus Social Distance Scale in Table 6.5 can be considered a Guttman scale. The items are arranged with the "easiest" at the bottom to the "hardest" at the top. Often, only two response categories are provided with Guttman scales, either Agree/Disagree or Yes/No. Some Guttman scales make use of the Likert-type response categories, but the categories are collapsed to a dichotomy in the data analysis.

The fact that the items in a Guttman scale are progressive and cumulative leads to the basic means of assessing whether a set of items constitutes a Guttman scale. This criterion is called *reproducibility*, which is the ability of each individual's composite score to predict exactly which items he or she had agreed and disagreed with. For example, in a true Guttman scale, all persons with scores of 2 will agree with the two easiest items and disagree with the rest; persons with scores of 3 will agree with the three easiest items and disagree with the rest; and so on. In a perfect Guttman scale, each respondent's score will reproduce one of these patterns, as is illustrated in Table 6.12. There is always one more perfect response pattern in a Guttman scale than there are items in the scale because one pattern will involve disagreeing with all the items; therefore, the six-item scale in Table 6.12 would have seven possible response patterns. You can see in each of the response patterns that once the "No" response changes to a "Yes," the person then answers "Yes" to all the easier questions. In actual practice, perfect Guttman scales are virtually nonexistent. Usually, some respondents will deviate from the expected pattern. Nevertheless, Guttman scales with very high levels of reproducibility have been developed.

Constructing a Guttman scale is difficult and to an extent risky because we will not

know whether the scale we have devised will have sufficient reproducibility to qualify as a Guttman scale until after we have applied it to a sample of respondents. As with the other scaling techniques, a basic first step is creating and selecting items for inclusion in the scale. In Guttman scaling, this task is further complicated by the need for the items eventually selected to have the characteristic of progression. The procedure for selecting items for a Guttman scale is known as the *scale discrimination technique* (Edwards & Kilpatrick, 1948). As was done with both Likert and Thurstone scaling techniques, we begin by writing a large number of statements relating to the variable to be measured. These statements are rated by a group of judges along the 11-point Thurstone equal-appearing interval scale. The items on which judges are in the greatest agreement are given a Likert-type response format and presented to a pretest group. The pretest results are used to calculate discriminatory power (DP) scores as described under Likert scaling. Items for inclusion in the final Guttman scale are selected so that they cover the full Thurstone scale range and have the highest DP scores. Despite the effort involved in this approach, all it accomplishes is to increase the likelihood that the selected items will have sufficient reproducibility to constitute a Guttman scale; it does not guarantee reproducibility.

The only way to determine if we have succeeded in developing a true Guttman scale is to administer it to another pretest group and see if it has adequate reproducibility. This is done by determining how many errors occur in the response patterns. In Guttman scaling, an *error* refers to any response pattern by an individual that does not follow one of the expected patterns presented in Table 6.12. Table 6.12 presents one possible error pattern, where an individual responded "Yes" to an item that was harder than another item to which that person had responded "No."

The total number of these errors is calculated for all respondents and is used in the following formula to calculate the *coefficient of reproducibility* (R_c):

$$R_c = 1 - \frac{number\ of\ errors}{(no.\ of\ items) \times (no.\ of\ subjects)}$$

Guttman (1950) suggested that a coefficient of reproducibility of .90 is the minimum acceptable for a scale to qualify as a Guttman scale. In general, the more items in a Guttman scale, the more difficult it is to achieve a high level of reproducibility. For a very short scale, .90 would certainly be the minimum acceptable; with a longer scale, a slightly lower coefficient of reproducibility would be acceptable.

Suppose that we developed a scale and found its reproducibility too low. It is perfectly legitimate to then rearrange the order of the items or delete items in an effort to achieve the necessary reproducibility. We might, for example, delete one or two of the items containing the most error responses to see if the remaining items would produce adequate reproducibility to qualify as a Guttman scale.

The data generated by Guttman scaling is ordinal level. Given the relatively few items characteristic of these scales and the common Agree-Disagree format used, there are few possible scores for respondents to achieve. This means that large numbers of respondents will have tied scores on the scale, so many statisticians believe it is better to consider these numbers as ranks (ordinal) rather than interval- or ratio-level data (see Chapter 5). Guttman scales are unique, however, for the characteristics of unidimensionality and reproducibility. If these attributes are desired, they are apt to more than outweigh the presence of all the tied scores.

Table 6.13 provides a summary of the key features, advantages, and disadvantages of the indexes and scales I have discussed, as well as the tools that are used in their development and evaluation. Scales are most commonly used in research problems in which the unit of analysis is the individual (see Chapter 4), and

TABLE 6.13 A Comparison of Various Indexes and Scales

Measuring Device	Key Feature	Assessment Tools	Advantages	Disadvantage
Indexes	Separate indicators combine to create a single measure	Validity, reliability, item analysis	More valid and reliable than single-item measure	Not unidimensional
Likert scale	Evaluate statements with 4–7 response alternatives	Validity, reliability, item analysis, discriminatory power scores, factor analysis	Range of response alternatives, easy to respond to, easy to construct	Hard to interpret a single summary score
Thurstone scale	Equal-appearing intervals	Validity, reliability item analysis	Easy and quick to respond to, interval-level data	Difficult and costly to construct
Semantic Differential (SD) Scales	Choose points between polar-opposite adjectives	Validity, reliability, factor analysis	Easy to construct, can achieve reliability with few items, easy and quick to respond to	May not be unidimensional (depending on choice of adjectives)
Guttman Scale	Reproducibility of items	Validity, reliability, item analysis, coefficient of reproducibility	Unidimensionality, true scale with intensity structure to items	Difficult to construct

this emphasis is reflected in the preceding discussion. However, scales can also be developed to measure other units of analysis, such as the characteristics of organizations or political entities.

Avoiding Response Bias

As we saw in Chapter 5, a key issue in measurement is whether people's answers to questions are accurate reflections of their actual feelings, beliefs, or behaviors. In other words, our measure of some phenomenon should be determined by the nature of the phenomenon itself and not by systematic or random errors (review the measurement formula on p. 139). One source of such error in people's responses to questions or statements is called **response bias**: the tendency for an individual's answers to questions to be influenced by things other than their true feelings, beliefs, and behaviors. It can result in a patterned overestimation or underestimation of variables (Bradburn, 1983).

Sources of Response Bias

One source of response bias is called **response set**: Some people tend to be either yea-sayers or nay-sayers, tending either to agree or disagree with statements regardless of their content. This is sometimes called the *acquiescence response set* because it more often takes the form of people being predisposed to agree with statements. To illustrate this, look again at the self-esteem scale in Table 6.6. If the scale was constructed so that "Strongly agree" always indicated high self-esteem, then people who tend to agree with statements would score higher on self-esteem than they actually should because they tend to agree with statements irrespective of content. This would throw into question the validity of the scale because it would produce the systematic error discussed in Chapter 5.

Another source of response bias is **response pattern anxiety**: Some people become anxious if they have to repeat the same response all the time and change their responses to avoid doing so. If this occurs, then

their reactions to statements do not reflect their actual attitudes but rather their reaction to a certain response pattern, and the validity of the scale is reduced. (As students, you have probably had this experience when taking a multiple-choice exam. If several consecutive questions all have the same answer, you become concerned and may doubt answers that you are fairly sure of just because the pattern of responses differs from the more random pattern you expect.)

Another source of response bias is the **social desirability effect**: people's tendency to give socially acceptable, popular answers in order to present themselves in a good light. It is very socially unacceptable, for example, to admit using a knife or gun on your spouse, and this may affect how people respond to the Conflict Tactics Scale (Table 6.3). People may deny using a knife or gun, even if they have done so, in order to avoid appearing socially unacceptable to an interviewer. The Eye on Diversity box discusses some ways in which cultural diversity has to be taken into account in assessing response bias.

Reducing Response Bias

Researchers use a number of strategies in an attempt to reduce response bias. Response set and response pattern anxiety can be avoided by designing statements so that positive statements are not always an expression of the same attitude. Likert scales are routinely designed like this. You will note in the items in Table 6.6 that choosing "Strongly agree" on items 1, 3, 4, 7, and 10 would be an expression of high self-esteem; choosing "Strongly agree" on items 2, 5, 6, 8, and 9, on the other hand, would be an expression of low self-esteem. If "Strongly agree" were an expression of high self-esteem for all items, then some respondents would have to choose the same alternative on every item in order to express their opinion. Mixing the response pattern of items is taken into account in scoring Likert scales. The alternatives that indicate an expression of the same opinion or feeling are given the same numerical score. In our example, for instance, all high-esteem alternatives (whether they be "Strongly agree" or "Strongly disagree") are given a score of 4. Then each person's responses to all items can be summed for a total scale score.

Another technique for avoiding response bias is to present sensitive issues in a neutral and nonjudgmental context. In developing the CTS (see Table 6.3), Straus and his colleagues presented questions about violent acts in the context of disagreements and conflicts, which would presumably appear more socially acceptable to people than abuse and violence.

A third way to reduce response bias has to do with the ordering of questions: Questions can be asked in a hierarchical order, beginning with the less sensitive and gradually moving on to the more sensitive issues. The CTS does this by beginning with a few items that are positive ways of resolving conflict ("I explained my side of a disagreement to my partner") before moving onto questions about psychological and physical abuse. Questions about the use of violence don't appear until well into the instrument. The rationale for this design is that people feel less reticent about divulging acts of violence if they have been given the chance to show that such acts were "the last straw" after attempting other means of conflict resolution.

A fourth strategy for reducing response bias is to use an interspersed pattern for the items, where socially acceptable items are interspersed with the less acceptable items. In the Conflict Tactics scale, positive items, such as "I said I was sure we could work out a problem," are followed by such items as "My partner needed to see a doctor because of a fight with me." The reason for this pattern is that a straight hierarchical ordering may open the door to a form of response set: A respondent may blindly answer "Never" to every item once items begin referring to violent acts. Interspersing sensitive with

positive items encourages participants to think more carefully about each item before responding. So the CTS actually uses a combination of hierarchical and interspersed ordering of sensitive items. For any given scale, whether a hierarchical pattern or an interspersed ordering produces the least bias is an empirical question to be settled through research on the scale itself.

A fifth technique that helps to reduce response bias is called "funneling." A researcher might ask respondents first about conflict in their city, then about conflict in their local community and among neighbors, and finally about conflict in their own families. As another example of this, Moser and Kalton (1972) suggest phrasing questions so that respondents can answer in the third person. For example, "Many men have hit their wives at one time or another. I wonder if you know under what circumstances it happens?" This can be followed with a direct question asking if the respondent has done it.