CHAPTER

2

The Measurement of Attitudes

SAMUEL HIMMELFARB

Attitudes are not directly observable; their existence can only be inferred from overt responses or *indicators*. Attitudes as evaluative tendencies manifest themselves in three general classes of indicators: *cognitive*, *affective*, and *behavioral*. This chapter considers how responses belonging to the three classes have been or could be used to measure attitudes.

The chapter is organized into several sections. We begin with a general discussion of basic concepts and ideas about measurement. The next section presents some of the more common ways attitude scales are constructed. Most of these scaling techniques are quite general and may be used to construct attitude measures within any of the three indicator classes. This section is followed by a discussion of attitude measures that have been linked to a particular class of indicators and that generally have not been based on formal scaling models. Finally, we discuss various ways of assessing the reliability and validity of attitude measures and some of the factors that influence reliability and validity.

Basic Concepts and Ideas

Measurement

S. S. Stevens (1946, 1951), one of the founders of modern measurement theory, defined measurement as the assignment of numbers to objects or events according to rules. Measurement, however, requires more than number assignment by some rule. Our real number system has certain properties such as order (e.g., $4 > 2$), difference (e.g., $4 - 3 = 1$), and ratio (e.g., $6/2 = 3$). The aim of measurement is to assign numbers to objects so that the properties of the numbers that are assigned reflect the relations of the objects to each other on the attribute being measured (e.g., attitude). For example, if Person A has twice as much of the relevant attribute as Person B, we would like to assign numbers to A and B that reflect that 2-to-1 relationship.

Levels of Measurement. The relations between the real numbers assigned to objects in the measurement process may or may not reflect the actual relations that exist between the objects on the attribute being measured. This fact led Stevens to the concept of *levels of measurement*, or types of scales. In *nominal scales*, the lowest level, the numbers assigned to objects reflect only *equivalence* versus *difference*. Objects that are the same on the attribute are given the same number, and objects that are

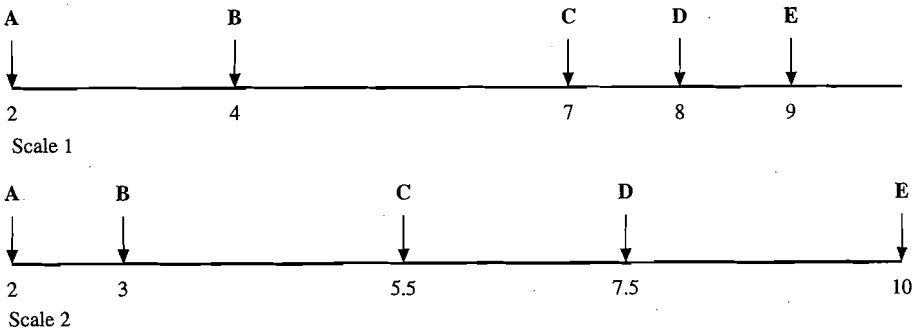

FIGURE 2.1. Two scalings that have the same ordinal properties.

different are given different numbers. For example, different numbers are assigned to ball players to reflect the fact that they are different players. The numbers assigned stand for the players' names (hence, "nominal") and imply nothing about their relative abilities. Similarly, the coding of all males as "1" and all females as "2" would yield a nominal scale of gender that reflects only one property of the number system, equivalence or difference.

The assignment of numbers on the basis of difference versus equivalence is rarely the goal of scale construction. Yet, the categorization of stimuli into same or different classes is fundamental to scaling because we must be able to distinguish between the objects being scaled. When we can also determine whether one object has more or less of the attribute than another, an *ordinal scale* can be constructed. For example, if we can discern that Person A has a less positive attitude than Person B and that Person B has a less positive attitude than Person C, numbers can be assigned to A, B, and C that reflect this ordering. Figure 2.1 illustrates an attitudinal dimension on which five persons have been located and assigned numbers that reflect the ordered relations among their attitudes. The values assigned in Scale 1 of Figure 2.1 are arbitrary and reflect only the ordinal properties of our scale. Persons A through E could have been assigned other values as long as the numbers assigned preserved the ordinal relationship between their attitudes. Such an alternative scaling is also shown in Figure 2.1 (Scale 2). A change from one set of values to another, even the arbitrary changes implemented in Figure 2.1, is called a *monotonic transformation*, if it preserves the ordering among the objects that are assessed. The two scales shown in Figure 2.1 thus preserve the ordinal relations among the persons, but not the distances between them on the attitudinal dimension.

When we can ascertain not only the order but also the exact size of the differences between objects, an *interval scale* of measurement can be constructed. To determine

CHAPTER

2

The Measurement of Attitudes

SAMUEL HIMMELFARB

Attitudes are not directly observable; their existence can only be inferred from overt responses or *indicators*. Attitudes as evaluative tendencies manifest themselves in three general classes of indicators: *cognitive*, *affective*, and *behavioral*. This chapter considers how responses belonging to the three classes have been or could be used to measure attitudes.

The chapter is organized into several sections. We begin with a general discussion of basic concepts and ideas about measurement. The next section presents some of the more common ways attitude scales are constructed. Most of these scaling techniques are quite general and may be used to construct attitude measures within any of the three indicator classes. This section is followed by a discussion of attitude measures that have been linked to a particular class of indicators and that generally have not been based on formal scaling models. Finally, we discuss various ways of assessing the reliability and validity of attitude measures and some of the factors that influence reliability and validity.

Basic Concepts and Ideas

Measurement

S.S. Stevens (1946, 1951), one of the founders of modern measurement theory, defined measurement as the assignment of numbers to objects or events according to rules. Measurement, however, requires more than number assignment by some rule. Our real number system has certain properties such as order (e.g., $4 > 2$), difference (e.g., $4 - 3 = 1$), and ratio (e.g., $6/2 = 3$). The aim of measurement is to assign numbers to objects so that the properties of the numbers that are assigned reflect the relations of the objects to each other on the attribute being measured (e.g., attitude). For example, if Person A has twice as much of the relevant attribute as Person B, we would like to assign numbers to A and B that reflect that 2-to-1 relationship.

Levels of Measurement. The relations between the real numbers assigned to objects in the measurement process may or may not reflect the actual relations that exist between the objects on the attribute being measured. This fact led Stevens to the concept of *levels of measurement*, or types of scales. In *nominal scales*, the lowest level, the numbers assigned to objects reflect only *equivalence* versus *difference*. Objects that are the same on the attribute are given the same number, and objects that are

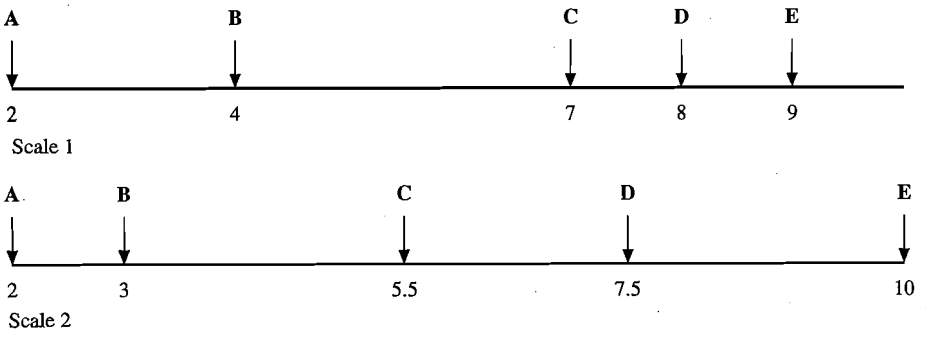

FIGURE 2.1. Two scalings that have the same ordinal properties.

different are given different numbers. For example, different numbers are assigned to ball players to reflect the fact that they are different players. The numbers assigned stand for the players' names (hence, "nominal") and imply nothing about their relative abilities. Similarly, the coding of all males as "1" and all females as "2" would yield a nominal scale of gender that reflects only one property of the number system, equivalence or difference.

The assignment of numbers on the basis of difference versus equivalence is rarely the goal of scale construction. Yet, the categorization of stimuli into same or different classes is fundamental to scaling because we must be able to distinguish between the objects being scaled. When we can also determine whether one object has more or less of the attribute than another, an *ordinal scale* can be constructed. For example, if we can discern that Person A has a less positive attitude than Person B and that Person B has a less positive attitude than Person C, numbers can be assigned to A, B, and C that reflect this ordering. Figure 2.1 illustrates an attitudinal dimension on which five persons have been located and assigned numbers that reflect the ordered relations among their attitudes. The values assigned in Scale 1 of Figure 2.1 are arbitrary and reflect only the ordinal properties of our scale. Persons A through E could have been assigned other values as long as the numbers assigned preserved the ordinal relationship between their attitudes. Such an alternative scaling is also shown in Figure 2.1 (Scale 2). A change from one set of values to another, even the arbitrary changes implemented in Figure 2.1, is called a *monotonic transformation*, if it preserves the ordering among the objects that are assessed. The two scales shown in Figure 2.1 thus preserve the ordinal relations among the persons, but not the distances between them on the attitudinal dimension.

When we can ascertain not only the order but also the exact size of the differences between objects, an *interval scale* of measurement can be constructed. To determine

TABLE 2.1

Measured Distances Between Persons and Alternative Interval Scalings

Person	*Distance between persons*	*Scaling 1*	*Scaling 2*	*Scaling 3*
A	2	0	0.0	-1.0
B	1	2	1.0	0.0
C	2	3	1.5	0.5
D	1	5	2.5	1.5
E		6	3.0	2.0

how much Person A differs from Persons B and C, a *unit of measurement* is required. That is, we need some standard device or unit that can be used to measure the distances between A, B, and C. The size of the unit of measurement can be arbitrary, just as it is arbitrary whether height is measured in inches or centimeters, or temperature in degrees Fahrenheit or Celsius. Suppose we had such a unit and observed the differences shown in column 2 of Table 2.1. From that table, we know that Persons A and B differ by 2 units, that Persons A and C differ by 3 units, and so on. Person A could then be assigned the number 0, B the number 2, and C the number 3, and so on (see Scaling 1 in Table 2.1). Yet, it would also be possible to assign to Persons A through E the numbers indicated in Scaling 2 and Scaling 3 of Table 2.1. Scaling 2 differs from Scaling 1 only in a change of the unit of measurement from the difference of 1 between B and C to the difference of 2 between A and B. The location of the zero point on an interval scale also is arbitrary. The numbers assigned in Scaling 3 differ from those in Scaling 2 in that B instead of A was assigned the value of 0. Because the zero point and unit of measurement in an interval scale are arbitrary, any system of number assignment can be changed to another one by a *linear* transformation.[1] Scalings 1, 2, and 3 differ from each other by linear transformations but preserve the basic distance relationships between persons that are given in column 2 and from which the scalings were derived.

If objects are measured on an interval scale, it is possible to make general statements about the *differences* between objects on the scale. For example, given the numbers assigned in Scaling 1 of Table 2.1, we can say that the difference between D's attitude and C's attitude (2) is twice the difference between B's attitude and C's attitude (1). This statement is true in all three of the scalings (or in any other linear transformation of them). However, it is not possible to say that D's attitude is 2.5 times more favorable than B's attitude because this statement would not be true across different interval scalings of the attitude such as those shown in Table 2.1.[2]

Ratio scale measurement is necessary in order to make statements about the number of times one person's attitude is more favorable or less favorable than another person's attitude. To construct a ratio scale, the numbers assigned must reflect distances from a *unique origin* or *zero point*, a point that is the same for all possible scalings of the objects and independent of their units of measurement. Then statements can be made about the

relative magnitudes of objects on the attribute. Because the size of the unit of measurement of a ratio scale is arbitrary, this unit can be changed without distorting the ratios of the objects to one another on the scale. A change in the unit of measurement without a change in the zero point of the scale is known as a *multiplicative transformation* ($Y = bX$). If two scalings of the same stimuli are on a common ratio scale, they should be linearly related to one another and have the same origin.

Representational Measurement. The ideal measuring instrument assigns numbers to people's attitudes (or other attributes) such that the relations among these numbers mirror aspects of the actual relations that exist among the attitudes of the people measured. When there is a correspondence between an empirical relation system and a numerical relation system, we have *representational measurement* (Dawes & Smith, 1985; Krantz, Luce, Suppes, & Tversky, 1971; Suppes & Zinnes, 1963). The importance of representational measurement is that the numbers assigned to objects allow us to deduce relationships that exist empirically between the objects on the dimension scaled. For example, if we knew that A has an attitude score of 8 and B an attitude score of 2 on a ratio scale, we would know that A is four times as favorable as B toward the attitude object. Because ratio scales of psychological attributes are quite rare, we are seldom in a position to make such statements about attitudes.

To determine whether representational measurement exists at a particular measurement level (e.g., ordinal, interval, or ratio), checks on the consistency of the number assignments should be conducted during the process of scale construction.[3] These consistency checks make use of the properties of the real number system to ascertain whether the numerical relations of the assigned scores mirror the empirical relations among the objects. For example, ordinal scales have the property not only of order, but also of *transitivity*: If $B > A$ and $C > B$, then $C > A$. Thus, transitivity provides a way of assessing whether a true ordinal scale has been constructed. For example, if Person D is judged to have a more positive attitude than Person C, and C a more positive attitude than Persons A and B, then D should be judged to have a more positive attitude than B or A. Intransitivities suggest that the people cannot be ordered consistently on a single dimension.

Interval scales have additional properties that can be used to check whether the scaling has met the basic requirements of an interval scale. For example, if Persons B, C, D, and E have been assigned the numbers 2, 3, 5, and 6, respectively, this implies that the difference between B and C should be judged equal to the difference between D and E. The various properties of different measurement levels are detailed in several useful discussions of measurement (e.g., A. B. Anderson, Basilevsky, & Hum, 1983; Dawes, 1972; Krantz et al., 1971; Suppes & Zinnes, 1963).

Attitude measures that lack representational measurement properties have been labeled *index measurement* (Dawes, 1972) or *nonrepresentational measurement* (Dawes & Smith, 1985). The fact that a particular scale yields "attitude scores" that are not based on representational measurement does not mean that the scale is worthless, however. The scale still may be useful in predicting scores on other variables. Yet nonrepresentational measures do not permit us to deduce the precise relations between

persons from knowledge of their attitude scores or between groups of persons from knowledge of their mean attitude scores. In considering how attitude scales are commonly constructed, we will discuss how they may be checked to determine if they have representational measurement properties.

Levels of Measurement and Statistics. In calculating certain descriptive and inferential statistics in attitude research, researchers typically add and multiply the numbers that represent subjects' attitudes. For example, in calculating the mean attitude of a group of individuals, researchers add the numbers assigned to those individuals and divide by the number of individuals. A person with a score of 8 contributes twice the amount in determining the group mean that a person with a score of 4 does. Yet, if the measurement level of the scale is only ordinal, a score of 8 may only indicate that the person's attitude is more positive than that of the person whose score is 4. It would be entirely consistent with the relations that exist between people's attitudes to transform the assigned values to some other set of numbers that preserves the existing ordinal relationships. Such a transformation would yield a different mean for the group. Moreover, the relationships between the means of different groups could be quite different depending upon the nature of the ordinal transformation. Recognition of this fact led Stevens (1951) and others (e.g., Siegel, 1956) to conclude that common statistical tests that require adding values should not be performed on scales that lack interval scale properties. Ordinal scales, they argued, require statistics such as the median that do not make use of scores' values but only of their order. Such statistics are called *nonparametric*.

Stevens' dictum led to considerable debate in the 1950s and early 1960s about the appropriateness of various statistical methods and tests at different levels of measurement. The debate subsided for a while among psychologists but was renewed in papers by Borgatta and Bohrnstedt (1980), Gaito (1980), and Townsend and Ashby (1984). Critics of Stevens' position argued that the level of measurement is not a problem for statistics but for the *interpretation* of certain statistical results (e.g., N. H. Anderson, 1961; Hays, 1963; F. M. Lord, 1953). After all, it was argued, the calculator or computer does not know where the numbers came from. It is a fact that the mean of the numbers assigned to one group is higher than the mean of the numbers assigned to another group. Given a significant *t*-test for this difference, the fact that the group means differ is likely to reflect a corresponding difference in their population means. These arguments are correct as far as the numbers are concerned. However, they do not resolve the issue of whether we can conclude that the two groups differ on the underlying attribute independent of the scale-specific method of number assignment.

This issue is complex because it is bound up in different theories or paradigms of what constitutes measurement (Michell, 1986). Yet, some progress has been made toward the resolution of this forty-year-old debate. Davison and Sharma (1988, 1990) have shown that, if an observed measured variable is a continuous ordinal variable that is a monotonically increasing function of an underlying latent variable (and the standard assumptions of homogeneity of variance and normality hold), the conclusion to reject or not reject the null hypothesis of no difference between the means on the

basis of a *t*-test or one-way analysis of variance on the measured variable also may be applied to the null hypothesis about the means on the latent variable. The same logic holds for tests about whether a correlation coefficient or multiple correlation coefficient is different from zero. Because it is reasonable to expect that our methods of measuring attitudes ordinarily are at least monotonically related to the true attitudes of our respondents, Davison and Sharma's findings indicate not only that the usual parametric statistical tests performed on our measured attitudes are permissible, but also that the conclusions drawn from them are likely to apply to the underlying attitudes.[4]

Reliability and Validity of Measurement

Any instrument designed to measure attitudes should be both a reliable and valid indicator of the underlying attitude. The *reliability* of a measuring instrument refers to the extent to which that instrument yields consistent scores or values over repeated observations. The *validity* of a measuring instrument refers to the extent to which that instrument measures what it claims to measure. That is, reliability is concerned with whether an instrument—regardless of what it "truly" measures—yields scores that are consistently repeatable. The validity of an attitude measure pertains to whether scores on that scale in fact indicate people's attitudes toward the object.

Errors of Measurement. All measurement is subject to some degree of error. Errors may arise from a variety of sources: The measuring instrument itself may have certain limitations that produce fluctuations; the object measured may vacillate on the attribute from one time or place to another; or, the observer or recording device may produce variability. For example, the measured weight of a person may differ from its true value and from a second or third measurement because of certain physical properties of the scale on which the person stands, because the person's weight fluctuates at different points in the day, or because the observer may read the scale from different visual perspectives and under different lighting conditions. Similarly, variability may be introduced by the electrical apparatus used to measure attitudes physiologically, attitudinal expressions may vary at different points in time, or people may misread an item or check the wrong alternative in responding to an item on a self-administered questionnaire.

Some errors fluctuate randomly; they are just as likely to cause the observed score to be higher as lower than its true value. By definition, such *random errors* have a mean of zero over repeated observations. That is, in the long run, errors in one direction will be balanced by errors in the other direction. *Systematic errors*, on the other hand, are departures from the true value that do not cancel themselves out over repeated observations. A tendency to make socially desirable responses, for example, would repeatedly lead to responses that depart from the true value only in the socially desirable direction. Random errors are the basis of a measuring instrument's unreliability, whereas systematic errors contribute to the instrument's invalidity.

Correlation coefficients typically are used to assess reliability and validity. There are a number of different ways of obtaining an index of reliability, but the basic idea is

to obtain a measure of the extent to which a set of scores on an instrument correlate with themselves on several observations (i.e., how consistent the scores are). In validity assessment, the relevant correlation is between scores on the measuring instrument and on some other variable to which the scores might reasonably be expected to be related if, in fact, the instrument measures what it claims to measure. A more detailed discussion of ways of assessing the reliability and validity of attitude measures is presented after discussion of some of the more common ways that attitudes have been measured.

Models of Measurement

There have been two traditions of measurement in psychology: psychophysical scaling and psychometric assessment. Both have influenced the ways we commonly measure attitudes.

Psychophysical scaling developed in the nineteenth century to examine the relationships between the attributes of physical stimuli and the psychological sensations that these stimuli produced. For example, researchers investigated how changes in the sound pressure of a tone related to sensed changes in loudness. To study this relationship, researchers would manipulate the tone's sound pressure and have perceivers judge how loud the tone was or how much louder it was than another tone. Psychophysical scaling involves mapping a psychological judgment dimension (e.g., loudness) onto the different physical values of a stimulus attribute (i.e., sound pressure).

The *Thurstone judgment* and *magnitude estimation* techniques of attitude measurement that we consider below have historical roots in psychophysical scaling. N. H. Anderson's functional measurement, covered in Chapter 5, fits within the psychophysical tradition as well. Given this heritage, these techniques scale *stimuli* (e.g., statements of belief, affect, or behavior) on a psychological dimension of evaluation, just as psychophysical techniques scale stimuli (e.g., tones) on a psychological dimension (e.g., loudness). However, because attitudes are attributes of persons, a second phase of scaling is used to locate *persons* on the attitude continuum. To recognize these two steps in this type of attitude measurement, these methods are referred to below as *stimulus, then person* scaling techniques. Generally, methods modeled on the psychophysical tradition aspire to some form of representational measurement.

The second measurement tradition, *psychometrics*, has its origins in the methods of mental and psychological testing. In contrast to psychophysical scaling, the attributes measured (e.g., intelligence) usually have no physical stimulus counterpart. On these tests, an individual responds to a series of items, each of which purports to assess the common underlying attribute that the test is designed to measure. Because more precise information about the attribute accumulates as the number of items increases, the sum (or average) of the scores on a number of items provides a good indication of where the person stands on the attribute. In the psychometric tradition, *persons* are located directly on the attribute based upon their total scores on a set of items. The typical multiple-choice course exam is an example of a test based on this psychometric model.

This psychometric heritage is also well represented in attitude measurement. Both *Likert's method of summated ratings* and *Osgood's semantic differential* fit within this approach. Techniques based on the psychometric model are referred to below as *person* scaling techniques. The representational measurement properties of scales derived from these person scaling techniques are generally unknown.

Guttman scaling, which we also discuss, combines aspects of both the psychophysical and psychometric heritages. As we shall see, Guttman scaling locates stimuli and persons simultaneously on the attitude continuum and in this chapter is labeled a *simultaneous stimulus and person* scaling technique. Guttman scaling yields ordinal scales with representational measurement properties.[5]

Scaling models differ in a variety of ways other than whether they scale stimuli, persons, or both. For example, the data used by a model may require judgments of order, while other models require distance or similarity judgments (Coombs, 1964; Coombs, Dawes, & Tversky, 1970; Dawes, 1972). Models also differ in whether they are designed to locate objects on a single dimension or to provide multidimensional representations. Because the most common techniques for measuring attitudes seek to locate people on a single dimension of favorability, this chapter focuses exclusively on *unidimensional* models. Readers may wish to consult other sources for discussion of multidimensional scaling (e.g., Kruskal & Wish, 1978; Schiffman, Reynolds, & Young, 1981; Shepard, Romney, & Nerlove, 1972). In the subsequent three sections we review and illustrate some of the traditional ways that researchers have constructed attitude scales by stimulus, then person scaling, by simultaneous stimulus and person scaling, and by person scaling. Most of these methods are quite general and can be applied across the three classes of indicators (cognitive, affective, behavioral). Other examples of these scaling techniques may be found in Shaw and Wright (1967), Robinson, Rusk, and Head (1968), Robinson and Shaver (1973), and Robinson, Shaver, and Wrightsman (1991).

Attitude Scale Construction: *Stimulus, Then Person Scaling*

Stimulus, then person scales require a two-step process. In the first step, stimuli (e.g., statements describing beliefs, affects, or behaviors) are judged and scaled to determine the location of each stimulus on a favorable-unfavorable dimension. For example, Table 2.2 presents some belief statements from the Attitude toward Capital Punishment Scale (R. C. Peterson & Thurstone, 1933/1970). Next to each item is a *scale value* representing the position of the item on an unfavorable (0) to favorable (11) dimension. The item scale values were derived from judges' ratings by the Thurstone method of equal-appearing intervals that is described subsequently. Once the items have been scaled, then persons (i.e., respondents whose attitudes are to be measured) are located on the same dimension by their endorsements of one or more of the scaled statements.

TABLE 2.2

Some Items and Their Scale Values from Attitude Toward Capital Punishment Scale

Scale value	*Item*
0.0	Capital punishment is absolutely never justified.
1.5	We can't call ourselves civilized as long as we have capital punishment.
2.4	Capital punishment cannot be regarded as a sane method of dealing with crime.
3.4	Life imprisonment is more effective than capital punishment.
3.9	I think the return of the whipping post would be more effective than capital punishment.
5.5	It doesn't make any difference to me whether we have capital punishment or not.
6.2	I think capital punishment is necessary, but I wish it were not.
7.9	Capital punishment is justified only for premeditated murder.
8.5	We must have capital punishment for some crimes.
9.1	Capital punishment should be used more often than it is.
9.6	Capital punishment is just and necessary.
11.0	Every criminal should be executed.

Note: Scale values of the items were obtained by the method of equal-appearing intervals (see text).
Source: This scale was presented by R. C. Peterson and Thurstone (1933/1970, pp. 22–23).

Thurstone Judgment Techniques

Louis L. Thurstone (1927a, 1927b), in two papers on psychophysics and what he called the *Law of Comparative Judgment*, developed a theory of judgment and choice that revolutionized psychophysics as well as psychological measurement. Many psychophysical experiments required subjects to compare a series of stimuli to some standard stimulus and to indicate which of the two was louder, brighter, or heavier. These judgments were then related to their physical stimulus dimensions and psychologically scaled in units of the physical dimension. Earlier psychophysicists saw a physical stimulus as producing a fixed sensation. In contrast, Thurstone theorized that the reaction to or judgment of a stimulus might vary slightly in a random fashion from one presentation to another and follow the shape of the normal curve. Figure 2.2 shows the psychological reactions to three different stimuli, i, j, and k. Thurstone called these distributions *discriminable dispersions*. The most typical psychological reaction, the mean, is the stimulus' *scale value*. Thurstone's great insight was that the extent to which one stimulus is judged to be greater (e.g., louder, more favorable) than another is related to the distance in their scale values on the psychological dimension (e.g., stimulus k should be judged greater than stimulus i more often than stimulus j is judged greater than stimulus i). By assuming that the distributions were

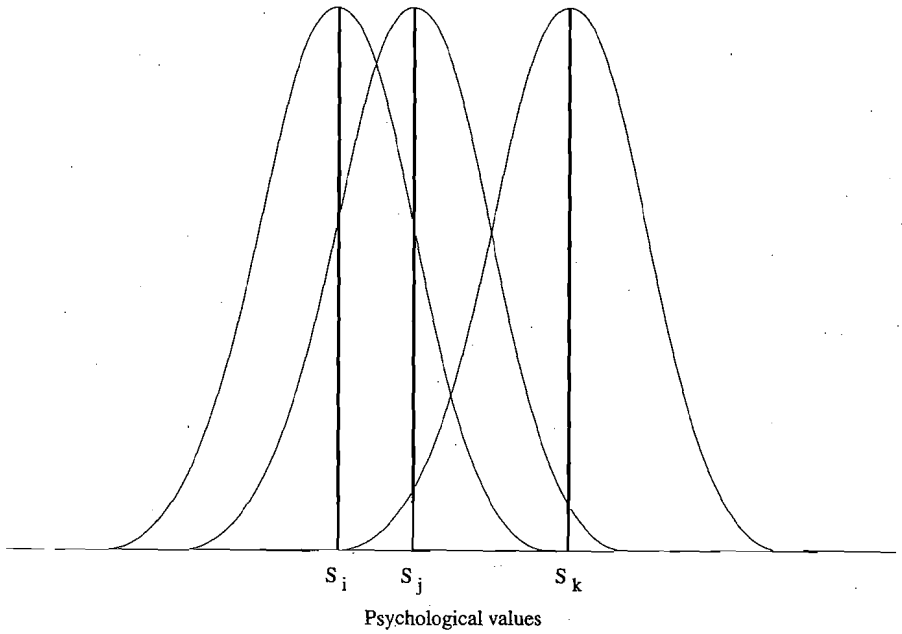

FIGURE 2.2. Discriminable dispersions created by variability in psychological reactions or judgments to three different stimuli *i*, *j*, and *k*. The ordinate indicates the likelihood of a judgment of a given psychological value. The means of the respective distributions (the scale values of the stimuli) are s_i, s_j, and s_k.

normal, Thurstone was able to measure the distances between stimuli in normal curve units of the psychological dimension rather than in units of a physical scale. The theory provided a rationale for the measurement of psychological attributes that did not have an underlying physical dimension.

In 1928, Thurstone published a paper entitled, "Attitudes can be measured." In this paper he demonstrated how his theory and the methods of psychophysical scaling—especially the *method of paired comparisons*—could be extended to attitude measurement. Thurstone and his coworkers subsequently developed the methods of *equal-appearing intervals* (Thurstone & Chave, 1929) and *successive intervals* (Saffir, 1937) as additional judgment techniques for attitude measurement and as approximations to the law of comparative judgment and the method of paired comparisons. This work marked the first applications of formal scaling methods to the measurement of attitudes.

In all of the Thurstone attitude scaling methods, the process of scale construction begins with the writing and assembling of a pool of statements that express varying

degrees of favorability and unfavorability toward the attitude object. The item pool should be large enough to represent as many different points as possible, including neutral points, along the favorable-unfavorable continuum. Once the pool of statements is assembled, the items are presented to a group of judges for the purpose of locating the items' positions on the evaluative dimension.

As noted, a basic assumption common to all Thurstone scaling methods is that each stimulus produces a normal distribution of judgments on the dimension of judgment (see Figure 2.2). For example, the dimension of judgment could represent the degree of favorability toward capital punishment that a belief statement is judged to express. The distribution arises from the fact that the same statement may elicit somewhat different degrees of judged favorableness in different individuals or in the same person (i.e., judge) from one occasion to another. The point of central tendency of the distribution, the mean or median (which are the same in a normal distribution) represents the item's scale value on the evaluative dimension. Below, we consider how the scale values of the items may be obtained from the methods of equal-appearing intervals, successive intervals, and paired comparisons.

Method of Equal-Appearing Intervals. In this method the judges are instructed to place each stimulus into one of a number of rating intervals (usually 11) according to how favorable or unfavorable an evaluation it expresses. For example, in the first application of this method to attitude measurement, Thurstone and Chave (1929) had 300 judges sort 130 belief statements about the church into 11 piles or intervals according to how favorable or unfavorable the item was toward the church (i.e., institutionalized religion). In some applications of the method the judges are instructed to treat all of the intervals as equal, but that instruction is not essential.[6] The method assumes that the judges, even without being told, sort the items into what appear to them to be equal intervals. As in the other Thurstone judgment techniques, judges are told not to express their own views about the attitude object or issue but to judge the favorableness or unfavorableness expressed by the item.

Scale values for the items are easily determined by the method of equal-appearing intervals. Because each interval is assumed to be equal to every other interval, the width of each interval can be arbitrarily set equal to 1. Consecutive scores (e.g., 1 to 11) can then be assigned to each of the intervals. A score can be assigned to each item equal to the value of the interval in which each judge placed the item. For example, if a judge placed the item into the fifth interval, the item would have a score of 5 based on that judgment. The scale value of an item is the median of the scores assigned to the item by all the judges.

Table 2.3 shows another set of items scaled by the method of equal-appearing intervals. The statements, which describe affective reactions, are from a scale that Breckler and Wiggins (1989b) developed to measure attitudes toward donating blood. The scale values were based on 15 judges' sortings of the items into 7 intervals ranging from very unfavorable (1) to very favorable (7) about blood donation.

TABLE 2.3

Affect Items and Their Scale Values from Attitudes Toward Blood Donation Scale

Scale value	*Item*
3	Blood donation makes me feel *uncomfortable.*
6	Blood donation makes me feel *generous.*
2	Blood donation makes me feel *unhappy.*
1	Blood donation makes me feel *ill.*
4	Blood donation makes me feel *bored.*
5	Blood donation makes me feel *assured.*
5	Blood donation makes me feel *relaxed.*
3	Blood donation makes me feel *jittery.*
2	Blood donation makes me feel *bad.*
6	Blood donation makes me feel *useful.*
4	Blood donation makes me feel *indifferent.*
7	Blood donation makes me feel *overjoyed.*

Note: Scale values of the items were obtained by the method of equal-appearing intervals. The items were selected to represent each integer point on the 1–7 scale.
Source: This scale was presented by Breckler and Wiggins (1989b, pp. 401–404).

Measurement of Respondents' Attitudes. The scaling of items merely locates the items on the attitude dimension. The next step is to select from the pool of scaled stimuli a subset of items to be administered to the respondents whose attitudes are to be measured. Items are selected so that collectively they represent, in even gradations, the range of possible scale values from very unfavorable to very favorable toward the attitude object. As explained below, these items should meet other criteria as well (e.g., low variability in their placements by the judges). These items are presented in a random order (without their scale values) to the respondents, who are asked to indicate the items with which they agree. A respondent's attitude score is the mean or median of the scale values of the items that she or he endorses in all Thurstone methods.

Method of Successive Intervals. Research comparing the scale values obtained by the method of equal-appearing intervals and the method of paired comparisons (see below) indicated that the relationship was not perfectly linear. Stimuli tended to be bunched together more at the extremes by the method of equal-appearing intervals (see A.L. Edwards, 1957b; Guilford, 1954). The intervals at the extreme needed stretching, and the middle intervals needed contracting in order for the two methods to be perfectly related. Because Thurstone regarded the method of equal-appearing intervals as yielding only an approximation to the results obtained by the method of paired comparisons, he devised the method of successive intervals as another way of obtaining scale values for the stimuli and improving upon the method of equal-appearing intervals.[7]

TABLE 2.4

Some Behavioral Items and Their Scale Values from Social Distance Scale

Scale value	*Item*
0.00	I would marry this person.
11.11	I would accept this person as an intimate friend.
21.50	I would accept this person as a close kin by marriage.
29.50	I would accept this person as a roommate or I would date this person.
38.70	I would accept this person as a neighbor.
49.40	I would live in the same apartment house with this person.
52.40	I would accept this person as one of my speaking acquaintances.
63.10	I would give asylum to this person, if he were a refugee, but I would not grant him citizenship.
69.70	I would not permit this person to live in my neighborhood.
81.00	I would not permit this person's attendance of our universities.
95.00	I would exclude this person from my country.
97.20	I would be willing to participate in the lynching of this person.

Note: Scale values of the items were obtained by the method of successive intervals. By a linear transformation of the original scale values, marriage was given a scale value of 0, and the item scale values extend over a 100-point range. To use this scale to assess attitude toward a group, the wording should be modified to read "I would _____ a person of Group X."

Source: These items were presented by Triandis and Triandis (1960, Table 1, p. 111).

The method, first reported by Saffir (1937), uses the same sorting or rating procedures for judging the items as the method of equal-appearing intervals. The judges are not told to treat the intervals as equal, but it would not matter if they were. The method assumes that the intervals may not be equal and derives their widths from the judgment data. Thus, data obtained by the method of equal-appearing intervals could be scaled by the method of successive intervals.

Table 2.4 presents 12 items from a scale designed to measure attitudes toward individuals or groups on the basis of statements that describe interpersonal behavior. Triandis and Triandis (1960, 1965) scaled these and other items by the method of successive intervals on the basis of 35 undergraduates' judgments.

To appreciate the specifics of deriving scale values by this method, we must understand the logic by which the method derives the widths of the intervals and locates the items on the resulting scale. For simplicity, assume that each judge sorted a number of items into one of five intervals (categories) according to how unfavorable (1) or favorable (5) the item was toward the attitude object. For the group of judges as a whole, suppose that the proportions of judges who placed item i in the successive intervals 1 through 5 were .023, .136, .341, .433, and .067. As in the other Thurstone methods, the judgments of each item are assumed to be normally

distributed around the mean, which is the item's scale value (s). Figure 2.3(a) shows the discriminable dispersion for item i. The differently shaded patterns shown on this curve demarcate the portions of the area under the curve that correspond to the proportion of times the judges placed item i in each of the intervals 1 through 5. As can be seen, because these demarcated areas vary in size, the widths of the intervals differ.

The widths of the intervals are derived from the assumption that the judgments of an item are normally distributed and from the properties of the normal curve. In a normal curve, 2.3 percent of the scores fall below a z-score of -2. Therefore, if 2.3 percent of the judges placed item i in Interval 1, the upper boundary of Interval 1 would be defined by a z-score of -2. If 13.6 percent of the judges placed item i in Interval 2, the proportion of times that item i was placed in Intervals 1 and 2 would be .159 (.023 + .136). In a normal curve, .159 of the area is below a z-score of -1. Therefore, the value of the upper boundary of Interval 2 is -1. It follows that the width of Interval 2 is 1. More generally, a z-score expresses how much any point (t) on the horizontal axis deviates from the mean (s) in units of the standard deviation (σ) of the distribution. Symbolically,

$$z = (t - s) / \sigma \tag{2.1}$$

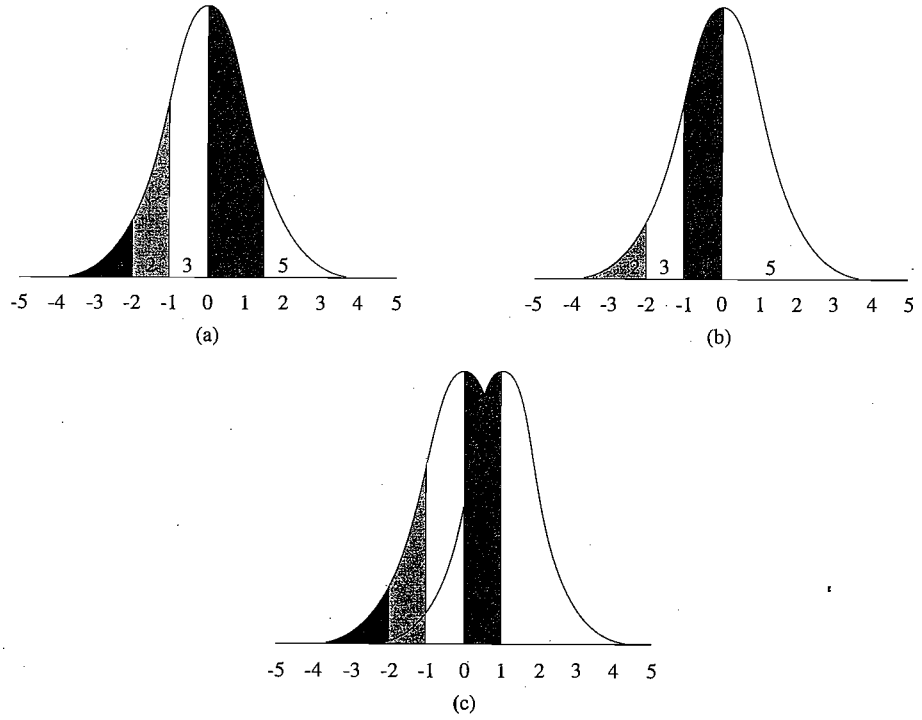

FIGURE 2.3. Discriminable dispersions for stimulus *i*, Panel (a), and stimulus *j*, Panel (b), in relation to the interval boundaries and to each other, Panel (c), in the method of successive intervals. Numbers in the different shaded areas give the interval in which the item was placed, and the size of the area indicates the proportion of times the item was placed in each of the five intervals. Panel (b) assumes that none of the judges placed the item in Interval 1. Panel (c) shows both distributions placed on the same horizontal axis. The horizontal axis of each panel is measured in *z*-score units of the standard normal curve.

Consequently, we know that the upper boundary of Interval 1 is 2 standard deviations below item *i*'s scale value and that the upper boundary of Interval 2 is 1 standard deviation below. Thus, given normality of the judgment distribution of an item, the proportions of judges who locate an item in each category provide estimates of the widths of the intervals and the locations of the interval boundaries relative to the item's scale value.

Figure 2.3(b) shows the normally distributed judgments for a second item, *j*. The area of the curve has been partitioned according to the proportion of times item *j* was placed by the judges in the various intervals. As was the case with item *i*, these proportions yield estimates of the widths of the intervals and the locations of their boundaries expressed in *z*-score units. Accordingly, Figure 2.3(b) shows that the upper boundary of Interval 2 is located 2 standard deviation units below the mean of the judgment distribution for item *j*.

In applications of the method of successive intervals, it is quite common to assume that the distributions of judgments for different items have the same standard deviation. Consequently, Figures 2.3(a) and 2.3(b) were drawn so that both distributions would have standard deviations equal to 1.[8] We know that the upper boundary of Interval 2 was -1 in units of distribution *i*'s standard deviation and -2 in units of distribution *j*'s standard deviation. With both item distributions having a standard deviation of 1, it follows that the scale value of item *i* is 1 unit below the scale value of item *j*. This difference is shown in Figure 2.3(c) where the two distributions can be seen on the same attitude continuum. More generally, in the method of successive intervals, the scale values of the items are determined in relation to the interval boundaries, whose locations are derived from the judgment data.

The specifics of estimating the scale values of the items and the interval boundaries follow the underlying logic outlined above. To estimate the interval boundaries and the scale values of the items we first obtain the proportion of times each item was sorted into each interval category or the categories below it in rank. These *cumulative proportions* are entered into a matrix like that shown in Table 2.5. In this matrix, the rows represent the items, and the columns represent the intervals. With the aid of a table that gives *z*-score values for areas (i.e., proportions) under the standardized normal curve (found in any elementary statistics book), the cumulative proportion matrix of Table 2.5 is then transformed into a matrix of *z*-scores (see Table 2.6). For example, the cumulative proportion of .30 in the upper left cell of Table 2.5 corresponds to the *z*-score of $-.52$ in the upper left cell of Table 2.6 because 30 percent of the area in a normal distribution lies below a *z*-score value of $-.52$. Note that in Table 2.6 the last column from Table 2.5 has been omitted because of the indeterminancy of *z*-scores for proportions of 1.00. Similarly, if any columns on the left of Table 2.5 contained only proportions of 0.00, these columns would have been omitted.[9]

Given the assumption that the standard deviations of the judgments are equal to 1 for all items, the difference between any two *z*-scores in the same row of Table 2.6 provides an estimate of the difference in the locations of the interval boundaries and

TABLE 2.5

Proportion of Times Each of Five Items Was Placed in Each Interval or Intervals Below it in Rank in the Method of Successive Intervals

	Intervals					
Item	*1*	*2*	*3*	*4*	*5*	*6*
1	.30	.55	.75	.85	.95	1.00
2	.20	.50	.85	.90	.96	1.00
3	.20	.45	.80	.85	.95	1.00
4	.10	.20	.40	.70	.85	1.00
5	.05	.15	.30	.50	.90	1.00

TABLE 2.6

Normal Curve *z*-Score Values for Cumulative Proportions of Successive Intervals in Table 2.5

	Interval boundary							
Item	*1*	*2*	*3*	*4*	*5*	*Row sum*	*Row mean*	*Scale value*
1	-.52	.13	.67	1.04	1.64	2.96	.59	-.35
2	-.84	.00	1.04	1.28	1.75	3.23	.65	-.41
3	-.84	-.13	.84	1.04	1.64	2.55	.51	-.27
4	-1.28	-.84	-.25	.52	1.04	-.81	-.16	.40
5	-1.64	-1.04	-.52	.00	1.28	-1.92	-.38	.62
Sum	-5.12	-1.88	1.78	3.88	7.35		1.20	.00
Mean (Interval boundary)	-1.02	-.38	.36	.78	1.47		.24	

thus of the interval width. Each row provides a separate estimate of the differences in location of the corresponding interval boundaries and of the interval width. In addition, the difference between any two *z*-scores in the same column of Table 2.6 provides an estimate of the difference in the scale values of the items in the corresponding rows. For example, the difference between −.52 and −.84 in column 1 is an estimate of the difference in scale values between items 1 and 2. The difference between .13 and .00 in column 2 also is an estimate of the difference in scale values between items 1 and 2. The values in rows 1 and 2 for each of the other columns also yield estimates of the scale values of items 1 and 2.[10]

In actuality, we can avoid calculating all of these differences to estimate the interval boundaries and scale values because the mean of the differences is the same as the difference between the means. Therefore, the differences between the column means reflect the average of the estimated differences between the interval boundaries. Similarly, the differences between the row means reflect the average of the estimated differences between the item scale values.

In order to determine the final scale values of the items, a zero point needs to be set. One simple way to do this is to allow the zero point to be the mean of the assigned scale values (i.e., of the values in the "row mean" column of Table 2.6). This value is .24 in Table 2.6. The final scale values of the items are obtained by subtracting the row means from this value.[11] The locations of the category boundaries are given by the column means in the *z*-score matrix. When some entries in the *z*-score matrix are missing because the obtained proportions are extreme, a somewhat more complex procedure must be used for obtaining interval widths and scale values (see A. L. Edwards, 1957b; B. F. Green, 1954; Torgerson, 1958).

Several consistency checks on scalings by the method of successive intervals were suggested by A. L. Edwards and Thurstone (1952). For example, the assumption that the dispersion distributions are normal can be checked by plotting on normal probability paper the cumulative proportions for an item (i.e., the entries in a given row of Table 2.5) against the interval boundary values obtained from Table 2.6. The plot for each row should be approximately linear. The consistency of the scaling also can be checked by working backwards to generate the predicted cumulative proportions in each of the categories once we have determined the scale values of the items and the category boundaries. A. L. Edwards (1957b) reported average absolute discrepancies of .025 and .021 between the predicted and obtained cumulative proportions for two different scalings. Average errors of these magnitudes appear to be quite minor, but the statistical properties of this discrepancy index are unknown. To our knowledge, an overall statistical test of goodness-of-fit has not been developed.

Once the items have been scaled from judgments by the method of successive intervals, the items can be used to measure the attitudes of respondents. This procedure, by which respondents indicate the items they agree with, is the same as that described for the method of equal-appearing intervals.

Method of Paired Comparisons. The core of Thurstone's initial theoretical development concerned comparative judgments and was designed for data collected by the method of paired comparisons. In this method each stimulus is paired with every other stimulus. For each pair, judges are required to state which of the two stimuli lies above the other on the judgmental dimension. For example, a set of n belief statements about capital punishment may be paired with one another, resulting in $[n(n-1)]/2$ pairs. For each pair, judges indicate which member of the pair is more favorable toward capital punishment. For a group of judges, the proportion of times statement j is judged more favorable than statement i is obtained. The method makes use of the data on the proportion of times the judges view one item as more favorable than another to derive the distances between the items' scale values and to position the items on the attitude dimension.

The details of how the scale values are determined by the method of paired comparisons are not pursued here (see A.L. Edwards, 1957b; B.F. Green, 1954; Torgerson, 1958) because of the limited usefulness of this method for scaling a large number of attitudinal statements. This limitation stems from the requirement that judges compare each stimulus with every other stimulus. As the number of stimuli increases, the number of required pairings and judgments increases more rapidly, and the technique becomes unwieldy. For example, 10 stimuli require 45 pairings, but 20 stimuli require 190 pairings. Yet, in order to construct a scale containing a sufficient number of items located at various points along the evaluative continuum, most investigators would probably want to scale 20 to 25 attitudinal statements, at least. Because of these practical limitations, Thurstone developed the methods of equal-appearing intervals and successive intervals, which require far fewer judgments.[12]

The theory underlying the method of paired comparisons is richly developed in Thurstone's (1927a) paper on the law of comparative judgment. The paper also considers a variety of subcases that make different assumptions about the equality of the standard deviations of the item dispersions and the correlations of the judgment pairs. Many further developments are discussed in Torgerson (1958). As in the method of successive intervals, the method of paired comparisons has associated with it a way of checking the consistency of the scaling. After the scale values of the items are obtained, the formulas can be reversed to generate predicted proportions which can be compared to the obtained proportions. Yet, in contrast to the method of successive intervals, the method of paired comparisons has in addition a statistical test of goodness-of-fit of the scaling (Mosteller, 1951). Consequently, the accuracy of the scaling and the interval scale assumption can be rigorously checked. The method of paired comparisons is highly recommended for scaling stimuli when the large number of judgments required by this method is not a serious limitation.

Item Selection in the Thurstone Techniques. Using the methods of equal-appearing and successive intervals, researchers can readily scale more items than are needed on the final questionnaire to represent the range of the evaluative dimension. However, some of these items might be inappropriate because they are ambiguous or irrelevant. Following Thurstone and Chave (1929), there are two criteria for eliminating inadequate items.

One of these criteria allows researchers to detect *ambiguous items*. With an ambiguous item, some of the judges might see it as favorable toward the attitude object, and others might judge it as considerably less favorable or even as unfavorable. Highly ambiguous items would be distributed by the judges across a wide range of intervals on the evaluative continuum. Therefore, items that have a large spread should be eliminated because their judged favorableness varies considerably with different judges. Thurstone and Chave (1929) suggested the use of Q (the interquartile range) as an index of spread, but the standard deviation of the item would do as well except when the items are quite skewed (Guilford, 1954). Given two or more items of roughly the same scale value, the one with the least spread is preferred for the final scale.

The second criterion is intended to eliminate *irrelevant items*, that is, items that do not differentiate between people with different attitudes on the issue. For example, people with different attitudes toward organized religion did not respond differently to the

statement *I am interested in a church that is beautiful and that emphasizes the aesthetic side of life* (Thurstone & Chave, 1929). People who are favorable to religion generally are interested in beautiful churches, but many atheists evidently are interested too. Such an item would probably not be eliminated because of ambiguity concerning its location on the scale (i.e., it definitely favors churches). Nonetheless, the item is inappropriate because it does not discriminate between people who are favorable and unfavorable toward religion.

To eliminate such items, researchers need to determine how each item relates to the attitudes of the respondents. This relation can be examined by determining the item's *operating characteristic*, which plots respondents' probability of agreement with an item as a function of their attitude scores on the entire scale. To obtain an item's operating characteristic, a large number of respondents are grouped according to their attitude scores on the scale (e.g., all those with a score of 1 are grouped together, those with a score of 2 are grouped together, and so on). Within each score group, the proportion of respondents who agreed with the item is obtained. When these proportions are plotted against their respective attitude scores, the resulting curve should resemble that shown in Figure 2.4. This figure shows an ideal operating characteristic curve for an item with a scale value of 6. As depicted in this figure, people whose attitude scores are in the middle of the distribution should agree with the item because it is close in value to their

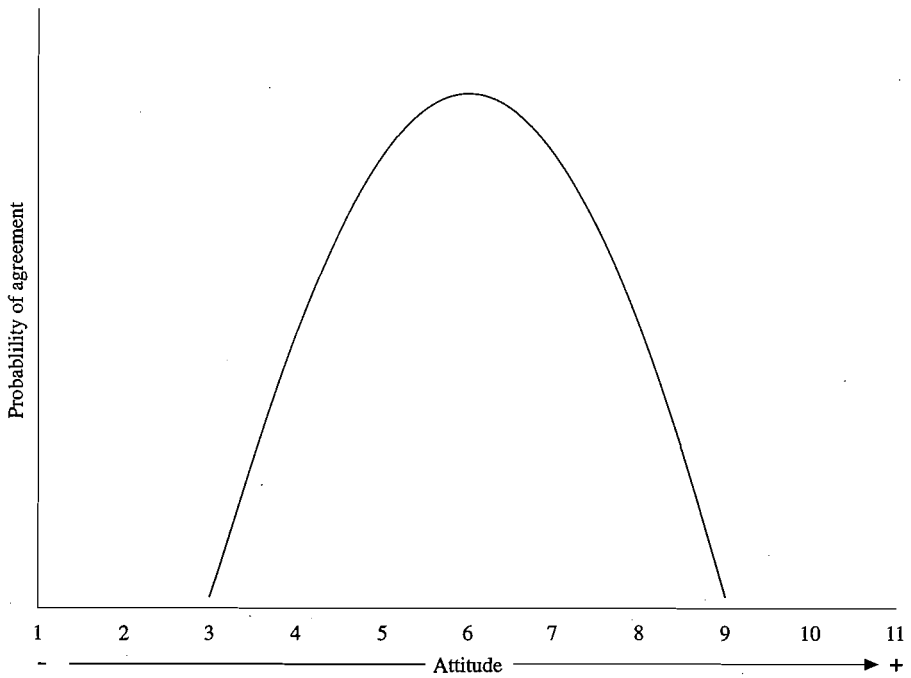

FIGURE 2.4.
Ideal operating characteristic curve for a Thurstone scale item with a scale value of 6. The figure indicates that the probability of agreement should be highest for respondents whose attitudes are close to the item's scale value and should decrease as respondents' attitudes are less or more favorable than that expressed by the item.

attitudes, and those whose attitudes are more extreme in either direction should be less likely to agree. In general, an item scaled by any of the Thurstone methods should have a *nonmonotonic* operating characteristic with a single maximum like that shown in Figure 2.4. The curve should peak somewhere close to the item's scale value. Yet items with extreme scale values would exhibit a monotonic operating characteristic. Items with flat or multipeaked operating characteristics should be discarded as irrelevant to the attitude dimension because they are endorsed by people whose attitudes are at various locations on the scale.

Evaluation of Thurstone Judgment Techniques. The methods of paired comparisons and successive intervals are sophisticated techniques with well established traditions in psychological scaling. Moreover, these two methods have built-in procedures for checking whether the scalings have interval scale properties. However, the method of paired comparisons is typically used only for scaling a small number of stimuli, while the most popular of the Thurstone techniques, the method of equal-appearing intervals, is based on a nontestable and unrealistic assumption that the scale intervals are in fact equal. Thus, attitude researchers have favored the weakest of these three techniques.

A major drawback to the Thurstone methods is that they tend to eliminate items at the extremes of the favorability continuum. Extreme items are generally placed by all of the judges in the same extreme interval, or all of the judges agree that the item is more favorable (unfavorable) than other items. These items are eliminated because they violate the normality assumptions of the methods. Thus, they are not eliminated because they fail to reflect an attitude, but because they do not satisfy the underlying Thurstone theory, which requires that the judgments be subject to random variation among the judges. The techniques therefore are unable to handle items that produce little or no variation in the judgments. Although few people may endorse extreme items as representative of their attitudes, in some research we would like to be able to identify individuals with very extreme attitudes. The Thurstone methods may not permit us to do so.

One key question about the Thurstone techniques for scaling attitudes is whether the scalings of attitudinal stimuli are influenced by the attitudes of the judges from whose responses the scalings are derived. Early research on scaling attitudes toward blacks (Hinckley, 1932), war and peace (Ferguson, 1935), and patriotism (Pintner & Forlano, 1937) concluded that there was little or no influence of the judges' attitudes. However, Hovland and Sherif (1952) noted a methodological problem with some of this early work and presented data that showed systematic biases due to the judges' own attitudes. Indeed, Sherif and Hovland (1961) provided a theoretical account of these judgmental biases (see Chapter 8).

Subsequent research has confirmed that judges' attitudes influence the perceived position of attitudinal statements (e.g., Eiser, 1971; Manis, 1960, 1961b; Selltiz, Edrich, & Cook, 1965; Upshaw, 1962, 1965; Zavalloni & Cook, 1965; see Chapter 12). However, Upshaw (1962, 1965, 1969) presented evidence that the judges' attitudes influence only the origin and unit of measurement of the scale. Because the origin and unit are arbitrary for interval scale measurement, Upshaw claimed that bias due to judges' attitudes does not invalidate the scaling technique. Research by Kelley,

Hovland, Schwartz, and Abelson (1955) suggests that the problem may be most serious for scalings by the method of equal-appearing intervals. They found that a scaling by the method of successive intervals showed less influence of the judges' own attitudes, and a scaling by the method of paired comparisons evidenced no influence at all. Because the method of paired comparisons forces the judges to discriminate between each item pair, that method is likely to be least susceptible to biases from the judges' attitudes.

Because all of the Thurstone techniques require a scaling of items and then of persons, they are often regarded as more tedious and cumbersome than other methods of attitude measurement. Yet all scaling techniques require pretesting and that calculations be performed to select the items for the scale. Although the calculations required for Thurstone's successive intervals and paired comparisons techniques were once regarded as time-consuming, this criticism was relevant only before the advent of modern data-processing techniques. With computers, the scale values of items can be obtained efficiently for all of the Thurstone techniques. Also, reliable scalings of items can be obtained when as few as 15 judges are used (see A. L. Edwards, 1957b, pp. 94–95).

Magnitude Estimation

Psychophysics offers a number of additional techniques that can be used to obtain scale values for attitudinal stimuli. One of the more useful methods is Stevens' *magnitude estimation* task (Stevens, 1956; Stevens & Galanter, 1957). Although Stevens (1966, 1972) and Hamblin (1974) noted the utility of this method for scaling stimuli of interest to social psychologists, the technique has received only limited attention from social scientists (e.g., W. E. Dawson, 1982; W. E. Dawson & Brinker, 1971; Lodge, 1981; Lodge & Tursky, 1982; Wegener, 1982).

In the magnitude estimation method, two stimuli are presented to judges who are required to judge the ratio of the stimuli. Typically, a judge is presented with one stimulus (i.e., attitude item) called the modulus, which is given an arbitrary numerical value, say 100. A second stimulus is provided, and the judge is required to assign a number that reflects the ratio between the two stimuli. For example, in relation to an attitudinal modulus (i.e., a belief statement) located at 100, a judge would assign a stimulus (i.e., a second belief statement) a value of 200 if he perceived it as twice as favorable as the modulus, 150 if he perceived it as one and one-half times as favorable, 50 if he perceived it as half as favorable, and so on. As in the Thurstone methods, the mean numerical judgment of each stimulus is calculated. These means are analogous to the item scale values in the Thurstone techniques. If the judges did make ratio judgments, the item means would differ from the Thurstone scale values because they would be measured on a ratio scale as opposed to the interval scale assumed by the Thurstone techniques.

Once the items have been located on the attitude dimension, the scaling of persons would generally follow the procedure used in Thurstone techniques. The respondents whose attitudes we wish to measure would be presented with the items or a subset of them and would be instructed to indicate which ones represented their position on the issue. The respondent's attitude score would be the mean or median of the scale values she or he endorsed.

Tasks other than number assignment have also been used to obtain estimates of magnitude. For example, judges may be instructed to treat the prestige of a particular occupation as equal to the brightness of a modulus light. Judges would then estimate the prestige of a second occupation by adjusting the brightness of a variable light so that the relative brightness of the two lights indicates the relative magnitudes of the prestige of the two occupations. The variable light would thus be set twice as bright as the modulus light to indicate that the second occupation has twice the prestige as the first occupation. In addition to judgments of the brightness of lights, several other response tasks have been used to obtain estimates of the magnitudes of social stimuli—namely, judgments of the strength of handgrips, the loudness of tones, and the length of lines.

One advantage of magnitude estimation techniques is that multiple modalities (e.g., brightness of lights and loudness of tones) can be used to cross-validate a scaling; that is, a scaling obtained using one modality (e.g., brightness) can be compared with a second scaling using a second modality (e.g., loudness; see W. E. Dawson, 1982).

In addition to scaling stimuli such as occupations for their prestige value, researchers have used magnitude estimation techniques to scale the favorability of adjectives associated with the response categories often used in survey research (Lodge, Cross, Tursky, & Tanenhaus, 1975). For example, ratings of the favorableness of adjectives yielded scale values of 233 for *excellent*, 107 for *good*, and 47 for *neither good nor bad*, averaged over several different policy issues. These results suggest that favorability denoted by a response of *excellent* is approximately twice that of *good*, which, in turn, is approximately twice that of *neither good nor bad*. When these adjectives are used as response categories to measure attitudes, a respondent can be assigned an attitude score equal to the scale value of the adjective that she endorsed. Research on this technique has been limited to responses to single items and has not included multi-item scales.

The magnitude estimation task is suited for the scaling of attitudinal stimuli, and the use of cross-validation techniques in this work is quite sophisticated. Yet, it is not clear whether magnitude estimation judgments yield ratio or interval scales. M. H. Birnbaum (1982) persuasively argued that magnitude estimation judgments do not have the ratio properties that Stevens claimed. Moreover, most of the work on magnitude estimation has focused on the scaling of stimuli, and only a few studies considered the subsequent step of using these scaled stimuli to scale persons' attitudes (e.g., Lodge & Tursky, 1979). Consequently, additional research is needed before the value of these techniques for attitude measurement can be assessed.

Attitude Scale Construction: *Simultaneous Stimulus and Person Scaling*

As noted earlier, Louis Guttman (1941, 1944) developed a scaling technique that simultaneously scales stimuli and persons. This technique orders stimuli and persons on a single dimension that has *cumulative* properties. In attitude measurement, this single cumulative dimension would be an evaluative dimension. To understand what is meant

TABLE 2.7a

Raw Data Matrix for Guttman Scalogram

	Stimuli (rods)				
Persons	*C*	*E*	*B*	*D*	*A*
2	1	1	1	1	0
4	0	1	0	1	0
3	1	1	0	1	0
6	0	0	0	0	0
5	0	1	0	0	0
1	1	1	1	1	1

TABLE 2.7b

Reordered Data Matrix for Guttman Scalogram

	Stimuli (rods)					
Persons	*A*	*B*	*C*	*D*	*E*	*Score*
1	1	1	1	1	1	5
2	0	1	1	1	1	4
3	0	0	1	1	1	3
4	0	0	0	1	1	2
5	0	0	0	0	1	1
6	0	0	0	0	0	0

by a cumulative scale and how such a scale simultaneously orders both stimuli and persons, consider a simple example of scaling along the physical dimension of length. Assume that we have five rods that vary in length between 5 and 7 feet, although the exact length of each rod is unknown. We will use these rods to create an *ordinal* scaling of the height of various persons by comparing each person's height to the length of each rod.

To construct a Guttman scale of length (or height), we begin with a matrix in which (a) the columns represent the stimuli (rods) and (b) the rows represent the persons whose heights we intend to measure (see Table 2.7a). Guttman called this matrix of stimuli by persons a *scalogram*, and his method of scaling is often referred to as *scalogram analysis*. If a person is taller than a particular rod, we place a 1 in the corresponding stimulus-person cell. If a person is not taller than a particular rod, we enter a 0 in the corresponding cell. The results of our measurements might look like those displayed in Table 2.7a. This table shows that Person 2 is taller than Rods C, E, B, and D, but not taller than Rod A. In contrast, Person 4 is taller than Rods E and D, but not taller than Rods C, B, and A.

The next step in scalogram analysis is to reorder the stimulus columns on the basis of how many 1s each column has so that the rightmost column has the most 1s, and the leftmost column has the least 1s. The person rows of the matrix also are reordered so that the top row has the most 1s (i.e., the tallest person is at the top), and the bottom row has the least 1s (i.e., the shortest person is at the bottom). Table 2.7b shows the reordered measurement matrix. Notice that the cell entries follow a pattern in which the 1s form a triangle. This triangular pattern indicates that we have successfully created a Guttman scale in which both the stimuli (rods) *and* the persons have been ordered (i.e., scaled) on a length dimension, even though we never directly compared one person to another person or one rod to another. The reason we could order both the rods and the persons is that the dimension of length has cumulative properties. Length accumulates such that the magnitude of a longer rod includes the magnitude of a shorter rod. Therefore, when we observed that Person 1 was taller than Rods A and B

and that Person 2 was not taller than Rod A but was taller than B, it followed that Person 1 must have been taller than Person 2 and that Rod A must have been longer than Rod B.

In the last column of the matrix in Table 2.7b, the persons have been assigned scores that consist of the number of 1s in their respective rows of the matrix. Because of the properties of the scale, Person 2's score of 4 tells us not only that he surpassed 4 rods in height but also that he is taller than Rods B, C, D, and E. Similarly, Person 3's score of 3 tells us that she is taller than the 3 lowest-ranked rods (i.e., Rods C, D, and E). In general, a person's score tells us not only how many but also which specific rods she or he surpasses in height. Accordingly, in this example, we can reproduce the entire matrix of measurements from knowledge of the persons' scores. When a matrix is reproducible from persons' scores, the scale that has been constructed is said to be unidimensional. The reproducibility of the measurement matrix was thus regarded by Guttman as a way of testing the hypothesis that a stimulus attribute (e.g., height, attitude) is scalable on a single dimension. Indeed, *reproducibility* of the matrix from respondents' scores defines *scalability* and *unidimensionality* in Guttman scaling.

Guttman Attitude Scales. Let us now substitute attitudinal stimuli for the rods in the above example. To illustrate a Guttman attitude scale, Table 2.8 shows the items from the Bogardus (1925, 1959) Social Distance Scale, one of the earliest efforts to measure attitudes toward ethnic groups. On this scale, the items reflect how closely one would be willing to associate with members of a particular ethnic group. Bogardus found

TABLE 2.8

Bogardus' Social Distance Scale

According to my first feeling reactions I would willingly admit members of each race (as a class, and not the best I have known, nor the worst members) to one or more of the classifications under which I have placed a cross (x).

	To close kinship by marriage	To my club as a personal chum	To my street as neighbors	To employment in my occupation	To citizenship in my country	As visitors only to my country	Would exclude from my country
Armenians							
Bulgarians							
Canadians							
Czecko-Slovaks							
Danes							
Dutch							
⋮							

Source: This instrument was presented by Bogardus (1925, p. 301).

that the items formed a hierarchical ordering of social distance. That is, people who indicated their agreement to permit members of an ethnic group to become part of their family also indicated agreement to allow these persons to do things that were less intimate (e.g., be a member of their occupational group). Conversely, if they agreed to exclude members of the group from their country, they did not agree to permit these persons to be neighbors. Although the social distance scale predated Guttman scaling, the behavioral intention items that Bogardus used for this scale appear to be cumulative and fit the requirements of a Guttman scale. Therefore, it would be possible to assign scores to individuals that unambiguously indicate the social distances they were willing to permit for members of a particular group. These scores represent the individuals' attitudes toward the group. Furthermore, knowledge of an individual's score also tells us which other items he endorses and which items he does not.[13]

Properties of Guttman Scales. The ability to reproduce the individuals' patterns of agreement and disagreement to each of the items necessarily follows from the nature of the operating characteristics of the items required for a Guttman scale. As we indicated earlier in relation to the Thurstone techniques, an item's operating characteristic indicates the relationship between respondents' attitudes and the probability that they agree with an item. An ideal Guttman operating characteristic, which takes the form of a step-function, is displayed in Figure 2.5: *All* persons below a certain point on the

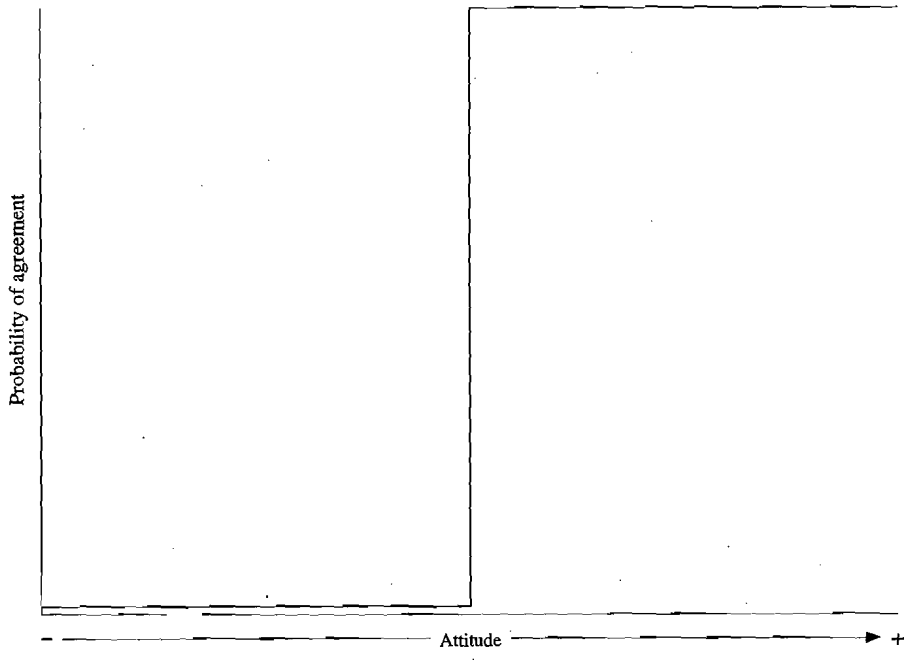

FIGURE 2.5.
Ideal operating characteristic curve for a positive Guttman scale item. The step function shown indicates that the probability of agreement should be zero as respondents' attitudes become more positive up to the point where the item is located. Respondents whose attitudes are equal to or more positive than that expressed by the item should all agree with the item.

attitudinal dimension should disagree with the item, and *all* persons above that point should agree with the item.

Unfortunately, a perfect Guttman scale is rarely obtained. Usually people's endorsements of attitude items do not form the triangular pattern of a perfect scale. To assess how common deviations from this pattern are, Guttman proposed a *coefficient of reproducibility*. This coefficient measures the extent to which the respondents' endorsements of the items can be reproduced from the triangular relationship that defines a perfect scale. The coefficient of reproducibility, *R*, is equal to 1 minus the proportion of responses that must be changed (i.e., *errors*) to produce a perfect scale (i.e., a perfect triangular pattern for all respondents). Guttman (1950) suggested that a coefficient of at least .90 is desirable to indicate scalability of stimuli on a single dimension.[14]

Guttman's coefficient of reproducibility has proven to be less informative than originally thought. High coefficients can sometimes be obtained from random patterns, and the value of the coefficient depends upon the proportions of respondents who endorse items. For example, Nunnally (1978) noted that a three-item Guttman scale with almost perfect reproducibility could easily be constructed by choosing one item endorsed by 10 percent of the respondents, a second endorsed by 50 percent, and a third endorsed by 90 percent, regardless of whether the items related even to the same content area. Corrections for this problem have been proposed (A.L. Edwards, 1957b; B.F. Green, 1956). Furthermore, there is disagreement about the counting of errors and the assignment of scale scores. These problems, some proposed solutions, and alternative measures of reproducibility are discussed in more detail elsewhere (Dawes & Smith, 1985; Dotson & Summers, 1970; A.L. Edwards, 1957b; McIver & Carmines, 1981).

Because a Guttman scale is an ordinal scale, a zero point is unnecessary and, at best, arbitrary. Nonetheless, Guttman (1947a) and Suchman (1950) suggested that it would be useful to distinguish between favorable and unfavorable attitudes. They further suggested that the zero point of the scale be located at the point where intensity of feeling about the issue is lowest. One way to determine this point is to ask the respondents "How strongly do you feel about this?" after eliciting their agreement or disagreement with the item. Suchman (1950) found a U-shaped relationship between intensity and Guttman scale scores such that people at both extremes of the scale felt more intensely about the issue. The low point of the U-shaped relationship—that is, a point of indifference—became the zero point of the scale.

Evaluation of Guttman Scaling. Guttman succeeded in constructing a number of attitude scales during World War II (Stouffer et al., 1950), and other scales have been constructed with the technique since then (see Robinson, Rusk, & Head, 1968; Robinson & Shaver, 1973; Robinson, Shaver, & Wrightsman, 1991; Shaw & Wright, 1967). Nonetheless, constructing a scale by Guttman's method is not an easy task. Several revisions generally are required. One problem is that the initial selection of items that may meet Guttman scaling criteria remains intuitive (A.L. Edwards, 1957b). Items often are discarded, rewritten, rescored, or otherwise manipulated in order to obtain a scale that meets satisfactory reproducibility criteria (see A.L. Edwards, 1957b;

McIver & Carmines, 1981). Scales exceeding 6 to 10 items rarely meet these criteria. Yet, since the number of items determines the number of different attitude scores that can be assigned to persons, short scales provide less discrimination between individuals' attitudes.

Some of the features of Guttman scaling are illustrated in Table 2.9, which lists 13 items used by Teske and Hazlett (1985) to construct a Guttman scale to measure attitudes toward handgun control. This scaling was based on data from a large sample of Texans who responded to an annual mail survey on crime. Respondents indicated whether they (a) strongly favor, (b) somewhat favor, (c) do not favor, or (d) have no opinion about the proposal expressed in each item. To construct a Guttman scale, Teske and Hazlett collapsed the first two alternatives into an agreement category, and the remaining two into a no agreement category.

The first nine items in Table 2.9 form a Guttman scale with a coefficient of reproducibility of .915, which far exceeds chance reproducibility. The remaining four items were eliminated from the scale because they lowered the reproducibility of the scale. In order to achieve the reported reproducibility, a step-by-step process was

TABLE 2.9

Items from a Guttman Scale of Attitudes Toward Handgun Control

1. Institute a waiting period before a handgun can be purchased, to allow for a criminal records check.
2. Require all persons to obtain a police permit before being allowed to purchase a handgun.
3. Require a license for all persons carrying a hundgun outside their homes or places of business (except for law enforcement agents).
4. Require a mandatory fine for all persons carrying a handgun outside their homes or places of business without a license.
5. Require a mandatory jail term for all persons carrying a handgun outside their homes or places of business without a license.
6. Ban the future manufacturing and sale of non-sporting-type handguns.
7. Ban the future manufacture and sale of all handguns.
8. Use public funds to buy back and destroy existing handguns on a voluntary basis.
9. Use public funds to buy back and destroy existing handguns on a mandatory basis.

Discarded items

A. A crackdown on *illegal* handgun sales.
B. Strengthen the rules for becoming a commercial handgun dealer.
C. Require a mandatory prison sentence for all persons using a handgun to commit a crime.
D. Ban the manufacturing and sale of small, cheap, and low-quality guns like the "Saturday Night Special."

Source: These items were presented by Teske and Hazlett (1985, p. 375).

followed in which an item was eliminated, the reproducibility of the remaining items recomputed to see if it met acceptable levels, another item was eliminated, the reproducibility of remaining items recomputed, and so on. One of the difficulties with this procedure is that the results may capitalize on chance variations among item frequencies. In order to be more certain that such an outcome has not occurred, the final scale should be cross-validated on a second sample to see if it again yields an acceptable level of reproducibility.

Guttman scaling is referred to as an *interlocking* technique (Dawes & Smith, 1985) because the resultant scale is a joint product of both stimuli and the persons scaled. In order to be successfully applied, the technique requires a particular relationship between the stimuli and the persons—namely, people who agree with an item also agree with items of lesser rank. To meet this requirement, successful Guttman scales often incorporate in the wording of the items certain circumstances or policies that make agreement with an item of a particular rank imply assent to the circumstances or policies described in the items of lesser rank. Thus, the items used by Teske and Hazlett are worded to have implications for one other. For example, someone who favors the most extreme item, which dictates mandatory destruction of all existing handguns, should also agree to less extreme items—for example, the item that calls for banning the future manufacture and sale of handguns. In general, the chances of creating a scale with Guttman properties are enhanced by wording the items so that acceptance of more extreme items has logical implications for acceptance of less extreme items.

Certain reactions, behaviors, and experiences occur in an orderly progression that make them amenable to Guttman scaling. For example, fear reactions in combat often progress from a pounding heart to urinating involuntarily (Stouffer et al., 1950). Sexual behavior between opposite sex college students appears to follow an orderly progression as well (Bentler, 1968a, 1968b; Podell & Perkins, 1957). So does social distance (see Table 2.8). Guttman scalings of attitudes are more likely to be successful if the items on the scale represent a clear progression from one to another. Conversely, the less they represent an orderly progression, the less responses to them are likely to be reproducible and fit Guttman's criterion of scalability.

To illustrate this issue, Guttman (1944) gave the example of a three-item scale of mathematical ability consisting of a problem on finding the area of a circle, a problem requiring the solution of a quadratic equation, and a problem in differential calculus. He noted that "there is no necessary logical reason why a person must know the area of a circle before he can know what a derivative is... The reason for a scale emerging in this case seems largely cultural. Our educational system is such that the sequence with which we learn mathematics... is first to get things such as areas of circles, then algebra, and then calculus" (p. 149). Elsewhere, he wrote that "If a population is not subjected to the same social stimuli with respect to the attitude, it might be expected that it will prove unscalable for them" (Guttman, 1947b, p. 461). In this respect, Guttman scaling can provide a useful technique for ascertaining whether stimuli have the cumulative, progressive, stepwise structure required by the scaling model.

The issue of whether a Guttman scale of a particular set of attitudinal expressions is achievable relates to yet another aspect of Guttman's theorizing. He viewed his scaling

technique as more than a method of constructing ordinal scales. He saw it as a way of testing whether attitudes toward some object or issue—"content universe," as he termed it—fell on a single dimension. Unidimensionality, of course, was defined by the successful construction of a Guttman scale. Determining whether various attitudinal responses toward some object form a unidimensional ordinal scale requires a somewhat different research strategy than determining whether a set of items and a set of respondents can be made to interlock to form a Guttman scale. To investigate unidimensionality of the content universe, a large, representative sampling of attitudinal responses toward the object should be subjected to the scaling technique. Moreover, some of the tactics that researchers often use to attain a successful scaling (i.e., the discarding and rewriting of items) would no longer be appropriate, just as discarding data from an ordinary study is not appropriate to coerce the data to fit the hypothesis.

Guttman was pessimistic that people's beliefs about most attitude objects would prove to be unidimensional. Early in the development of his scaling method, he recognized that "scalable universes may be the exception rather than the rule" (Guttman, 1947b, p. 461). Therefore, Guttman devoted much of the latter part of his career to the development of partial order scalogram analysis (e.g., Shye, 1978), multidimensional scalogram analysis, and other techniques of multidimensional scaling (e.g., Guttman, 1959, 1968; Lingoes, 1963; Zvulun, 1978).

Attitude Scale Construction: *Person Scaling*

In the scaling techniques discussed thus far the persons whose attitudes we wish to measure are positioned on the evaluative dimension in relation to the locations of the stimuli they have endorsed. The locations of the stimuli were either determined as a first step (e.g., Thurstone scaling) or simultaneously with locating the persons on the dimension (Guttman scaling). In contrast, in the methods considered in this section, there is no attempt to locate the stimuli at different points on the evaluative dimension. Stimuli are classified *a priori* as either favorable or unfavorable toward the attitude object, and the locations of persons on the attitude dimension are determined by the number of stimuli with which they agree and the extent of their agreement. As indicated earlier in this chapter, these scaling methods are derivatives of the *psychometric model* tradition in which responses to items are viewed as indicators of a common latent variable.

In this section we consider two such scaling techniques: *Likert scaling* and the *semantic differential*. Like the other scaling techniques we have considered, Likert's method is a general scaling technique that may be applied to any of the three classes of attitudinal responding. In contrast, the semantic differential does not apply across all three classes of indicators. It is instead based on ratings of the attitude object on adjective scales that present generalized evaluative beliefs (e.g., good vs. bad). The semantic differential is discussed in this section because the underlying measurement model is similar to that of Likert scaling.

Likert Scaling

Rensis Likert (1932) developed his *method of summated ratings* because he believed that Thurstone's techniques were too cumbersome and time-consuming. He set out to develop a simpler method of scaling that would be at least as reliable and valid as Thurstone's method of equal-appearing intervals.

Likert's scaling technique, like Thurstone's, begins with a large pool of items that are chosen intuitively for their relevance to the attitude object. Although in most applications of the technique these items consist of statements of belief, statements about behaviors or affective reactions toward the attitude objects have been used (e.g., Fishbein & Ajzen, 1974; Kothandapani, 1971; Ostrom, 1969). Unlike Thurstone items which are written to represent a variety of points along the evaluative continuum,

TABLE 2.10

Some Items from the Short Form of the Attitudes Toward Women Scale

The statements listed below describe attitudes toward the role of women in society that different people have. There are no right or wrong answers, only opinions. You are asked to express your feeling about each statement by indicating whether you (A) agree strongly, (B) agree mildly, (C) disagree mildly, or (D) disagree strongly. Please indicate your opinion by blackening either A, B, C, or D on the answer sheet for each item.

1. Swearing and obscenity are more repulsive in the speech of a woman than of a man.
2. Women should take increasing responsibility for leadership in solving the intellectual and social problems of the day.
3. Both husband and wife should be allowed the same grounds for divorce.
4. Intoxication among women is worse than intoxication among men.
5. Under modern economic conditions with women being active outside the home, men should share in household tasks such as washing dishes and doing the laundry.
6. There should be a strict merit system in job appointment and promotion without regard to sex.
7. Women should worry less about their rights and more about becoming good wives and mothers.
8. Women earning as much as their dates should bear equally the expense when they go out together.
9. It is ridiculous for a woman to run a locomotive and for a man to darn socks.
10. Women should be encouraged not to become sexually intimate with anyone before marriage, even their fiancés.
11. The husband should not be favored by law over the wife in the disposal of family property or income.
12. The modern girl is entitled to the same freedom from regulation and control that is given to the modern boy.

Source: These items were presented by Spence, Helmreich, and Stapp (1973, pp. 219–220).

Likert items are written and selected so that agreement with the item represents either a favorable or unfavorable attitude toward the object. The degree of favorability or unfavorability is ignored. Each item is presented to respondents in a multiple-choice format such as the following: A. Strongly Disagree; B. Disagree; C. Undecided; D. Agree; E. Strongly Agree. Respondents choose the alternative that best represents their degree of agreement or disagreement with the item. Each alternative on a Likert scale receives a score from 1 to 5 depending on the respondent's degree of disagreement or agreement with it. If, as is conventional, strong agreement with favorable items receives a high score (5), the scoring is reversed for unfavorable items so that strong disagreement receives a high score (5). Sometimes items are scored −2 to +2; both the scoring direction and the number assignments are arbitrary. Additional variations on the Likert procedures include provisions of more or fewer than five alternatives of agreement and disagreement as well as omission of the neutral or undecided alternative. For example, Table 2.10 reproduces some of the items from the short form of the *Attitudes toward Women Scale* (Spence, Helmreich, & Stapp, 1973), which assesses attitudes toward equal rights for women. These items have four alternatives and no neutral alternative. Likert's technique is referred to as the method of summated ratings because the scores received on each item are summed to obtain the respondent's total score on the attitude scale.

Item Analysis. In order to establish a Likert scale, the initial pool of items must be pilot tested on a group of respondents to eliminate ambiguous and nondiscriminating items. One frequently used technique in precomputer days for assessing whether an item was properly discriminating was to select those people in the top and bottom 27 percent of the total scale score distribution and test whether there was a statistically significant difference between the two groups' mean scores on the item.[15] The preferred contemporary procedure is to examine the *item-total score correlations*, each of which correlates the respondents' scores on an item with their scores summed over all the items.[16] A good item will have a positive item-total score correlation. Generally speaking, higher correlations indicate better items. Items with low or no correlation with the total score are discarded.[17]

In a complete item analysis, the researcher also examines the operating characteristic of each item, the relation between probability of agreement with an item to attitude scores on the total scale. Because Likert scales usually have several alternatives that reflect degrees of agreement, the frequencies of responses to the various agreement alternatives would have to be combined to determine the proportion of respondents who agree with an item. A much easier and equally valid way of examining how the item operates is to plot the item scores against total scale scores. The ideal operating characteristic for a Likert scale item is a monotonic function with probability of agreement or item scores increasing with increasing favorability of attitude for favorable items. Figure 2.6 illustrates several item operating characteristic functions consistent with the ideal. The exact shape of the function will depend upon the distributions of scores on the item and the total scale and on the favorability of the item. More critical is the slope of the function: relatively flat operating characteristics suggest

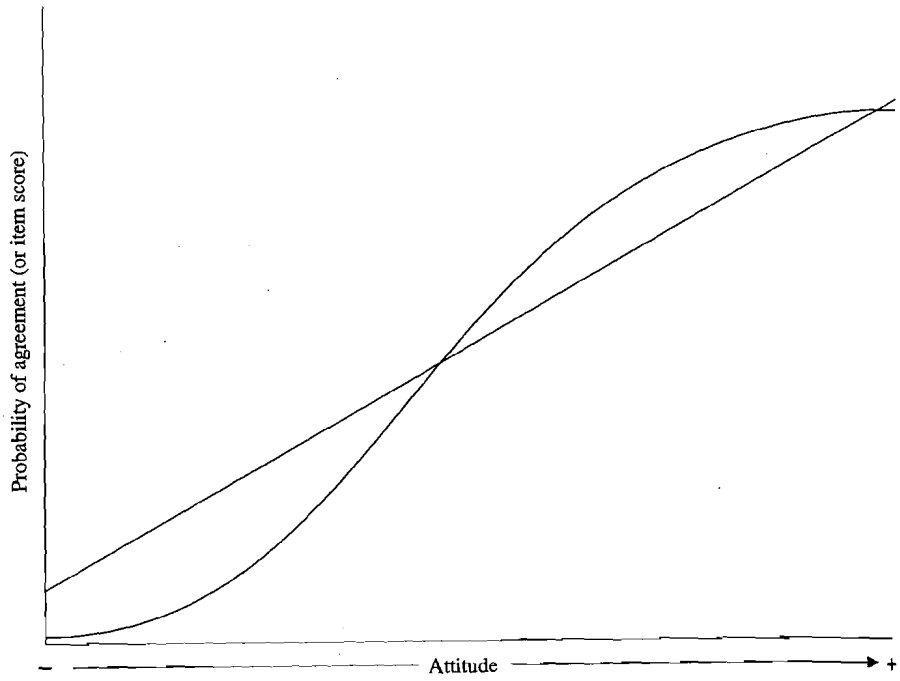

FIGURE 2.6. Ideal operating characteristic curves for a positive Likert scale item. Probability of agreement or the degree of agreement (item score) should increase as respondents' attitudes become more positive.

that the item is ambiguous or irrelevant because it is endorsed by persons with quite different attitudes toward the object.

Because the underlying measurement assumptions of Likert scaling are similar to those of other psychometric tests (e.g., achievement tests), the same item selection criteria used to construct these other tests are valid for maximizing the discriminatory power, reliability, and validity of a Likert scale. We will consider some of these criteria, including Cronbach's (1951) alpha, later in the chapter. More extensive coverage of these criteria may be found in most books on psychometric methods (e.g., M. J. Allen & Yen, 1979; L. Crocker & Algina, 1986; Nunnally, 1978).[18]

Evaluation of Likert Scaling. Generally, Likert accomplished his goal of developing an attitude scaling method that is as reliable and valid as Thurstone's technique but less time-consuming to construct than Thurstone's successive intervals and paired comparisons techniques. However, claims about efficiency gains (Barclay & Weaver, 1962) have been negated in recent years by the widespread availability of computers and research that indicates that reliable Thurstone scalings can be obtained from a much smaller group of judges than Thurstone initially suggested. Careful pretesting of items, item analyses, and item culling are time-consuming features of good scale construction that are required by both the Likert and the Thurstone methods.

Alternative form reliabilities of Likert scales have frequently been found to be greater than those of Thurstone scales when the two methods are compared or when respondents answer Thurstone scale items that are presented in the Likert format (see Seiler & Hough, 1970). However, direct comparisons between the two methods using the same items are problematic (e.g., Likert, Roslow, & Murphy, 1934; Poppleton & Pilkington, 1963). As B. F. Green (1954) has noted, the two methods require items with different operating characteristics, so that items appropriate for one type of scale should not ordinarily be used in constructing a scale of the other type.

The main disadvantage of Likert scales is that the exact level of measurement of the resulting scale scores is unknown. Unlike the Guttman and some of the Thurstone scaling techniques, Likert scaling does not have any internal checks for its representative measurement properties. Therefore, it is difficult to say whether it yields interval or ordinal level measurement. However, recent developments in item response theory (e.g., A. Birnbaum, 1968; Rasch, 1960) appear to provide a basis for assigning metric properties to various psychological tests (Weiss & Davison, 1981). Although these innovations could be applied to attitude scaling, researchers have not taken much advantage of them (but see Reiser, 1980).

Another disadvantage of Likert scaling is that, unlike Guttman's method, there are no built-in tests of dimensionality. Although Likert scaling attempts to locate people on a single dimension of favorability, it is impossible to make statements about the underlying dimensionality of Likert scales without further statistical analysis. As a means of assessing the dimensionality of tests, investigators often employ factor analysis as an adjunct to item analysis, particularly confirmatory factor analysis.[19] Indeed, when factor analyzed, they frequently yield more than one dimension.

Semantic Differential

Osgood, Suci, and Tanenbaum's (1957) semantic differential is the most popular way of measuring attitudes in contemporary research. The semantic differential consists of a series of bipolar adjective scales, each of which is conventionally separated into seven categories, as shown in Figure 2.7. The attitude object is placed at the top of the

Americans

Beautiful	---:---:---:---:---:---:---	Ugly
Bad	---:---:---:---:---:---:---	Good
Pleasant	---:---:---:---:---:---:---	Unpleasant
Dirty	---:---:---:---:---:---:---	Clean
Wise	---:---:---:---:---:---:---	Foolish

FIGURE 2.7. Several semantic differential bipolar scales that connote evaluative meaning.

Likert Scaling

Rensis Likert (1932) developed his *method of summated ratings* because he believed that Thurstone's techniques were too cumbersome and time-consuming. He set out to develop a simpler method of scaling that would be at least as reliable and valid as Thurstone's method of equal-appearing intervals.

Likert's scaling technique, like Thurstone's, begins with a large pool of items that are chosen intuitively for their relevance to the attitude object. Although in most applications of the technique these items consist of statements of belief, statements about behaviors or affective reactions toward the attitude objects have been used (e.g., Fishbein & Ajzen, 1974; Kothandapani, 1971; Ostrom, 1969). Unlike Thurstone items which are written to represent a variety of points along the evaluative continuum,

TABLE 2.10

Some Items from the Short Form of the Attitudes Toward Women Scale

The statements listed below describe attitudes toward the role of women in society that different people have. There are no right or wrong answers, only opinions. You are asked to express your feeling about each statement by indicating whether you (A) agree strongly, (B) agree mildly, (C) disagree mildly, or (D) disagree strongly. Please indicate your opinion by blackening either A, B, C, or D on the answer sheet for each item.

1. Swearing and obscenity are more repulsive in the speech of a woman than of a man.
2. Women should take increasing responsibility for leadership in solving the intellectual and social problems of the day.
3. Both husband and wife should be allowed the same grounds for divorce.
4. Intoxication among women is worse than intoxication among men.
5. Under modern economic conditions with women being active outside the home, men should share in household tasks such as washing dishes and doing the laundry.
6. There should be a strict merit system in job appointment and promotion without regard to sex.
7. Women should worry less about their rights and more about becoming good wives and mothers.
8. Women earning as much as their dates should bear equally the expense when they go out together.
9. It is ridiculous for a woman to run a locomotive and for a man to darn socks.
10. Women should be encouraged not to become sexually intimate with anyone before marriage, even their fiancés.
11. The husband should not be favored by law over the wife in the disposal of family property or income.
12. The modern girl is entitled to the same freedom from regulation and control that is given to the modern boy.

Source: These items were presented by Spence, Helmreich, and Stapp (1973, pp. 219–220).

Likert items are written and selected so that agreement with the item represents either a favorable or unfavorable attitude toward the object. The degree of favorability or unfavorability is ignored. Each item is presented to respondents in a multiple-choice format such as the following: A. Strongly Disagree; B. Disagree; C. Undecided; D. Agree; E. Strongly Agree. Respondents choose the alternative that best represents their degree of agreement or disagreement with the item. Each alternative on a Likert scale receives a score from 1 to 5 depending on the respondent's degree of disagreement or agreement with it. If, as is conventional, strong agreement with favorable items receives a high score (5), the scoring is reversed for unfavorable items so that strong disagreement receives a high score (5). Sometimes items are scored −2 to +2; both the scoring direction and the number assignments are arbitrary. Additional variations on the Likert procedures include provisions of more or fewer than five alternatives of agreement and disagreement as well as omission of the neutral or undecided alternative. For example, Table 2.10 reproduces some of the items from the short form of the *Attitudes toward Women Scale* (Spence, Helmreich, & Stapp, 1973), which assesses attitudes toward equal rights for women. These items have four alternatives and no neutral alternative. Likert's technique is referred to as the method of summated ratings because the scores received on each item are summed to obtain the respondent's total score on the attitude scale.

Item Analysis. In order to establish a Likert scale, the initial pool of items must be pilot tested on a group of respondents to eliminate ambiguous and nondiscriminating items. One frequently used technique in precomputer days for assessing whether an item was properly discriminating was to select those people in the top and bottom 27 percent of the total scale score distribution and test whether there was a statistically significant difference between the two groups' mean scores on the item.[15] The preferred contemporary procedure is to examine the *item-total score correlations*, each of which correlates the respondents' scores on an item with their scores summed over all the items.[16] A good item will have a positive item-total score correlation. Generally speaking, higher correlations indicate better items. Items with low or no correlation with the total score are discarded.[17]

In a complete item analysis, the researcher also examines the operating characteristic of each item, the relation between probability of agreement with an item to attitude scores on the total scale. Because Likert scales usually have several alternatives that reflect degrees of agreement, the frequencies of responses to the various agreement alternatives would have to be combined to determine the proportion of respondents who agree with an item. A much easier and equally valid way of examining how the item operates is to plot the item scores against total scale scores. The ideal operating characteristic for a Likert scale item is a monotonic function with probability of agreement or item scores increasing with increasing favorability of attitude for favorable items. Figure 2.6 illustrates several item operating characteristic functions consistent with the ideal. The exact shape of the function will depend upon the distributions of scores on the item and the total scale and on the favorability of the item. More critical is the slope of the function: relatively flat operating characteristics suggest

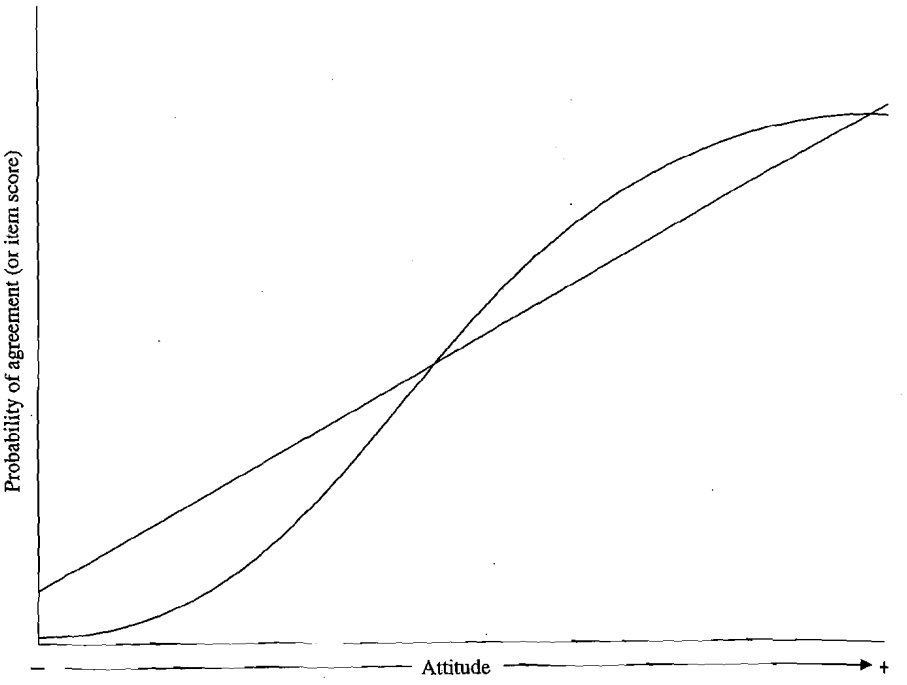

FIGURE 2.6. Ideal operating characteristic curves for a positive Likert scale item. Probability of agreement or the degree of agreement (item score) should increase as respondents' attitudes become more positive.

that the item is ambiguous or irrelevant because it is endorsed by persons with quite different attitudes toward the object.

Because the underlying measurement assumptions of Likert scaling are similar to those of other psychometric tests (e.g., achievement tests), the same item selection criteria used to construct these other tests are valid for maximizing the discriminatory power, reliability, and validity of a Likert scale. We will consider some of these criteria, including Cronbach's (1951) alpha, later in the chapter. More extensive coverage of these criteria may be found in most books on psychometric methods (e.g., M. J. Allen & Yen, 1979; L. Crocker & Algina, 1986; Nunnally, 1978).[18]

Evaluation of Likert Scaling. Generally, Likert accomplished his goal of developing an attitude scaling method that is as reliable and valid as Thurstone's technique but less time-consuming to construct than Thurstone's successive intervals and paired comparisons techniques. However, claims about efficiency gains (Barclay & Weaver, 1962) have been negated in recent years by the widespread availability of computers and research that indicates that reliable Thurstone scalings can be obtained from a much smaller group of judges than Thurstone initially suggested. Careful pretesting of items, item analyses, and item culling are time-consuming features of good scale construction that are required by both the Likert and the Thurstone methods.

Alternative form reliabilities of Likert scales have frequently been found to be greater than those of Thurstone scales when the two methods are compared or when respondents answer Thurstone scale items that are presented in the Likert format (see Seiler & Hough, 1970). However, direct comparisons between the two methods using the same items are problematic (e.g., Likert, Roslow, & Murphy, 1934; Poppleton & Pilkington, 1963). As B. F. Green (1954) has noted, the two methods require items with different operating characteristics, so that items appropriate for one type of scale should not ordinarily be used in constructing a scale of the other type.

The main disadvantage of Likert scales is that the exact level of measurement of the resulting scale scores is unknown. Unlike the Guttman and some of the Thurstone scaling techniques, Likert scaling does not have any internal checks for its representative measurement properties. Therefore, it is difficult to say whether it yields interval or ordinal level measurement. However, recent developments in item response theory (e.g., A. Birnbaum, 1968; Rasch, 1960) appear to provide a basis for assigning metric properties to various psychological tests (Weiss & Davison, 1981). Although these innovations could be applied to attitude scaling, researchers have not taken much advantage of them (but see Reiser, 1980).

Another disadvantage of Likert scaling is that, unlike Guttman's method, there are no built-in tests of dimensionality. Although Likert scaling attempts to locate people on a single dimension of favorability, it is impossible to make statements about the underlying dimensionality of Likert scales without further statistical analysis. As a means of assessing the dimensionality of tests, investigators often employ factor analysis as an adjunct to item analysis, particularly confirmatory factor analysis.[19] Indeed, when factor analyzed, they frequently yield more than one dimension.

Semantic Differential

Osgood, Suci, and Tanenbaum's (1957) semantic differential is the most popular way of measuring attitudes in contemporary research. The semantic differential consists of a series of bipolar adjective scales, each of which is conventionally separated into seven categories, as shown in Figure 2.7. The attitude object is placed at the top of the

Americans

Beautiful ___:___:___:___:___:___:___ Ugly

Bad ___:___:___:___:___:___:___ Good

Pleasant ___:___:___:___:___:___:___ Unpleasant

Dirty ___:___:___:___:___:___:___ Clean

Wise ___:___:___:___:___:___:___ Foolish

FIGURE 2.7. Several semantic differential bipolar scales that connote evaluative meaning.

page and respondents are asked to rate this object by checking a category on each of the bipolar scales (e.g., good-bad). Typically the instructions tell the respondent to check the middle category if neither adjective describes the object better than the other or if both are irrelevant to it. Respondents are told to check further along the scale to the extent that the object is described by either of the two adjectives. These category ratings are usually scored −3 to +3. Scores on the individual bipolar scales are summed or averaged to arrive at a total attitude score for each respondent.

As noted in Chapter 1, the semantic differential was developed to measure the connotative meaning of concepts. In numerous studies, Osgood and his colleagues (1957) had a large number of people within a number of cultures rate many concepts on many bipolar adjective scales. These ratings then were factor analyzed to determine whether the interrelations among the scales could be accounted for by a smaller number of underlying dimensions or factors. These analyses generally yielded three factors, which were labeled *evaluation*, *potency*, and *activity*. The evaluative factor ordinarily accounted for the largest amount of variability among scale ratings and was identified by Osgood and his colleagues as synonymous with *attitude*. Consequently, bipolar adjective scales that load on the evaluative dimension (e.g., those shown in Figure 2.7) are used to measure attitudes in the semantic differential technique.

Item Analysis. Despite Osgood and his associates' extensive research showing that certain adjectives generally indicate evaluative meaning, such adjectives may have more specific meaning in relation to particular attitude objects and issues. For example, the adjective pair *warm-cold* generally indicates evaluative meaning in rating people, but would convey meaning that is less evaluative and more denotative in ratings of the Mojave Desert or Alaska. Such tendencies for particular scales to convey specialized meanings in the context of particular concepts were dubbed *concept-scale interactions* by Osgood and his colleagues. Because of the possibility of such interactions, it is wise to assess the extent to which individual bipolar scales in any particular investigation can, in fact, be treated as forming a common evaluative scale. As in Likert scaling, this assessment can be performed by examining item operating characteristic curves or analyzing the correlations between respondents' scores on the individual scales and their scores summed or averaged across the scales (i.e., their total scores). The ideal item operating characteristic curve would be the same as that for a Likert scale item: Increasing favorability of respondents' total scores on the set of items should be accompanied by increasing favorability on the item (see Figure 2.6). In addition, the factor structure of the bipolar scales can be analyzed more formally through factor analysis (see note 19).

Heise's (1970) review of attitude research that has used the semantic differential suggested that the intercorrelations among the various bipolar adjective scales usually are sufficiently high that four or five bipolar scales yield adequate reliability for most purposes. Generally, for a given attitude object, evaluative scores from the semantic differential correlate highly with scores produced by other attitude scaling techniques (e.g., Breckler, 1984a; Fishbein & Ajzen, 1974; Jaccard, Weber, & Lundmark, 1975; Osgood et al., 1957).

Evaluation of the Semantic Differential. Unlike the other techniques discussed in this section, the semantic differential cannot be applied across all classes of evaluative responding. Despite this limitation, the semantic differential has become the most popular method of measuring attitudes. Its popularity stems from the ease with which it permits researchers to obtain an attitudinal index. Because the semantic differential uses adjectives (i.e., beliefs) that are very general and heavily saturated with evaluative meaning, specific belief items do not have to be prepared in advance and scaled. Therefore, the bipolar scales of the semantic differential have been described as the attitude researcher's "ever-ready batteries." In contrast, the indicators of evaluation used by other techniques (e.g., Thurstone, Likert, Guttman) typically must be inferred from the person's endorsements of favorable or unfavorable beliefs, affects, or behaviors that have been selected for their relevance to a particular attitude object.

Because the semantic differential does not depend upon items specific to a particular attitude object, it has the advantage of allowing comparisons of attitudes across different attitude objects (e.g., social groups, social policies). Using the technique, a researcher might find out, for example, whether respondents are more favorable toward Republicans than Democrats or toward affirmative action in college admissions than toward affirmative action in employment. Although attempts to construct such generalized attitude scales date back to the 1930s (Remmers, 1934; Remmers & Silance, 1934), the semantic differential is the most successful "Master Scale" developed thus far.

The main disadvantage of the semantic differential is that its representational measurement properties are essentially unknown. Consequently, it is difficult to know what level of measurement is obtained or what properties the obtained attitude scores have. However, as noted in the discussion of Likert scaling, recent advances in item response theory may provide a measurement metric for scales that, like the semantic differential, are based on the psychometric tradition.

Attitude Measures Linked to Specific Indicator Classes

Within the conception of attitudes adopted in Chapter 1, virtually any response can serve as an indicator of an attitude, provided that it is reliably associated with respondents' tendencies to evaluate the attitude object. The previous section discussed methods of attitude measurement based on standard scaling techniques that, with the exception of the semantic differential, are applicable to any of the cognitive, affective, and behavioral classes of indicators. In this section we consider methods that are linked to one specific class of indicators and that are not scaled by any of the general scaling techniques we have discussed. This presentation also omits discussion of methods that, like the lost letter technique, yield an attitude measure for populations but not for individuals (see Sechrest & Belew, 1983; Webb, Campbell, Schwartz, & Sechrest, 1966; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Projective techniques are also omitted because they have not proven to be more valid than standard questionnaires (see Kidder & Campbell, 1970, p. 369).

Cognitive Indicators

A number of techniques are based on the assumption that attitudes lead to systematic distortions in thoughts and judgments (see Chapter 12). To the extent that attitudes do exert selective effects at various stages of information processing, these systematic distortions can be used as indicators of attitudes.

One of the earliest measures of attitudes based on the assumption that attitudes bias judgments was Hammond's (1948) *error-choice method*. In this technique, respondents are presented with questions that they are to answer by selecting one of two alternatives that are provided. Although the respondents are led to believe that the questions test their factual knowledge, neither of the alternatives is actually correct. Instead, the alternatives embody either errors in opposite directions from the correct answer or opposing responses to questions that have no determinable answers. Hammond assumed that respondents' choice of one error or answer over another reflects their attitudes. For example, one of the questions Hammond used to measure attitudes toward labor versus management was the following:

The average weekly wage of the war worker in 1945 was:

a. $37
b. $57

The correct answer, $47, was not given as an alternative. Yet, forced to choose one of the two erroneous alternatives, members of businessmen's luncheon clubs were more likely to choose alternative *b*, and people working for a major labor organization were more likely to choose *a*. Hammond also found that alternatives unfavorable toward the Soviet Union were more likely to be chosen by the businessmen than by labor union workers.

Working along similar lines and with similar assumptions, Thistlethwaite (1950) investigated *distortions in logical reasoning* as indicative of prejudice and ethnocentrism. Students at northern and southern colleges were asked to judge whether a conclusion was true or false given certain premises. Both neutral and more "emotional" arguments were used. One of these presumably emotional arguments was the following:

> *Given:* If production is important, then peaceful industrial relations are desirable. If production is important, then it is a mistake to have Negroes for foremen and leaders over Whites.
>
> *Therefore:* If peaceful industrial relations are desirable, then it is a mistake to have Negroes for foremen and leaders over Whites (p. 444).

Thistlethwaite found that white students at southern colleges were more likely to make logical errors in the judged truth value of such emotional items that supported their prejudices (in comparison with more neutrally framed arguments) than were students at northern colleges.

In addition to judgments of logical conclusions, judgments of the *plausibility* of arguments have been examined as a measure of attitudes by Stuart W. Cook and his

colleagues (Selltiz & Cook, 1966; Waly & Cook, 1965). Under the guise of taking a logical reasoning test that required judging arguments from a debate, students rated the effectiveness of various arguments labeled as pro-segregation or pro-integration. Presumably, subjects would rate arguments consistent with their attitudes as more effective than arguments opposed to their attitudes. Correlations between plausibility scores and self-report measures of attitude ranged from .54 to .88 for students from various colleges in these studies, with the higher values associated with students at southern colleges.

Another judgmental phenomenon that has been applied to attitude measurement is the *contrast effect*, a tendency for persons at one end of an attitudinal continuum to displace statements that are distant from their position toward the opposite end of the continuum (see C. W. Sherif & Sherif, 1967; C. W. Sherif, Sherif, & Nebergall, 1965; M. Sherif & Hovland, 1961; M. Sherif & Sherif, 1967). That is, if people are asked to sort opinion items into various categories according to how favorable or unfavorable the items are toward the attitude object, people with favorable attitudes tend to judge unfavorable items as more unfavorable than do people whose own attitudes are not as favorable. Conversely, people with unfavorable attitudes tend to judge favorable items as closer to the favorable end of the continuum. Indeed, some of the influence of judges' attitudes on their judgments of items, a topic we discussed earlier in conjunction with Thurstone scaling, takes the form of contrast effects (see also Chapters 8 and 12).

The Sherifs based their *own categories method* of assessing attitudes on this judgmental contrast effect. When subjects were told to sort opinion items into categories according to how they "belong together," those who had extreme attitudes (a) sorted the items into fewer categories than those who were less extreme and (b) placed more items into the categories at the opposite end of the attitude continuum from their own position (C. W. Sherif & Sherif, 1967). Thus, the Sherifs argued that the number of categories respondents use and their placement of a disproportionate number of items in extreme categories can serve as another measure of attitudes (see Chapters 3 and 8).

Over the years, a variety of other cognitive measures (e.g., the learning and retention of arguments) have been explored as possible mediators of attitude change, particularly in response to persuasive messages. In fact, contemporary research has identified a variety of cognitive responses that are correlated with attitudes and are potential indicators of attitudes (see Chapters 3, 4, 6, and 7). For example, the number of positive or negative thoughts obtained in the *thought-listing* procedures used to investigate the mediational role of cognitive responses in persuasion experiments might serve as an indicator of attitudes (Brock, 1967; A. G. Greenwald, 1968; Petty, Ostrom, & Brock, 1981; see Chapter 6). The sum of these self-generated cognitions might have psychometric properties similar to other summative scales (e.g., Likert scales).

Although thought listing and other cognitive responses have not been systematically investigated as attitude measures, cognitive measures of the strength and favorability of beliefs have been studied. In the expectancy-value model of attitudes, attitudes are viewed as a function of the person's beliefs or expectancies that the attitude object has certain characteristics or attributes and the values attached to these characteristics (e.g., Fishbein, 1963; Rosenberg, 1956; see Chapter 3). For example, in Fishbein's research, the expectancy or strength of association between the attitude object and a

characteristic is measured on probabilistic scales (e.g., likely-unlikely, possible-impossible, etc.), and the evaluation of each characteristic is measured on semantic differential scales. The product of these two ratings is obtained for each characteristic associated with the attitude object, and these products are summed over all characteristics. The sum of these cross-products can be regarded as an index of attitude toward the object. This sum has been shown to correlate positively with a semantic differential measure of the attitude (see Chapter 5). Such summed Expectancy × Value products can be viewed as a respondent-weighted summative scoring system that fits at least informally within the psychometric measurement tradition. A number of methodological issues associated with expectancy-value techniques are discussed in Chapter 5. Finally, it is worth noting that Fishbein and Ajzen (1975) argued that all of the standard methods of measuring attitudes (i.e., Thurstone, Likert, Guttman, and semantic differential) can be regarded as deriving attitude scores from the product of a person's beliefs and the evaluations of associated characteristics or attributes.

Affective Indicators

According to the conceptualization of attitude in this book, attitudes, considered as evaluative tendencies, can be expressed in terms of affect responses (e.g., feelings, emotions) and can originate in affective experiences (see Chapter 1). Therefore, it is reasonable that social psychologists would attempt to measure attitudes through physiological responses that may be linked to emotional processes. In the following pages we briefly review and assess the status of physiological and other affective indicators of attitudes.

Galvanic Skin Response. The galvanic skin response (GSR) is a measure of skin resistance, the ability of the skin to conduct electricity. This response is under the control of the sympathetic nervous system and is related to activity of the sweat glands. Typically, GSR is measured by placing electrodes across the palm of the hand. Because sweating is often a response to stress or emotionality, strongly held attitudes may elicit sweat secretions that can be detected by a galvanometer or voltmeter.

Rankin and Campbell (1955) are generally credited with the first successful demonstration that attitudes may be related to galvanic skin responses. White male subjects had their right arms strapped to a board and GSR electrodes attached to their right palms. A set of dummy electrodes was placed on their left wrists. These subjects were given a word-association test that included words that might evoke emotional responses. The experiment was conducted by an experimenter and assistant, one of whom was black and the other white. On separate occasions, the experimenter and his assistant each made physical contact with the subject by adjusting the dummy electrodes. Rankin and Campbell found that the mean GSR was higher when the individual who made contact with the subject was black.

Porier and Lott's (1967) subsequent replication of this study failed to obtain differential GSRs to black and white experimenters but did find a correlation between the ethnocentrism scores (a measure of prejudice) of their white subjects and the degree

of differential GSRs to their black and white experimenters. Westie and DeFleur (1959) and Vidulich and Krevanick (1966) presented pictures of blacks and whites in interaction and found that white subjects with negative attitudes toward blacks, as measured by standard self-report methods, exhibited higher GSRs to these pictures than did subjects with more positive attitudes toward blacks. J. B. Cooper and his colleagues obtained higher GSRs to the names of negatively valued as opposed to positively valued ethnic groups (J. B. Cooper & Siegel, 1956). They also found higher GSRs when names of negatively valued groups were inserted in complimentary statements (J. B. Cooper & Pollock, 1959; J. B. Cooper & Singer, 1956). However, the GSRs in the latter studies may reflect responses to the inconsistency or unexpectedness of the stimuli rather than attitudinal responses toward the groups.

Despite these early positive findings, there is general agreement that the galvanic skin response is inadequate as a physiological measure of attitudes in several respects (Cacioppo & Sandman, 1981; S. W. Cook & Selltiz, 1964; Mueller, 1970; Petty & Cacioppo, 1983; Shapiro & Crider, 1969). First, large GSRs can be triggered by both negative and positive emotional reactions. As a measure of attitude, therefore, the GSR lacks the important property of directionality. Second, GSR appears to reflect not only arousal, activation, or emotionality, but also the orienting response that is triggered by surprise, change, novelty, inconsistency, or by the unexpected. As such, it may not be a very good measure of an attitude, as reflected in the positive or negative affect attached to an attitude object.

Pupillary Response. The pupils of the eye dilate and constrict and therefore have the potential to yield the bidirectional indicator of attitude that is lacking in skin conductance measures. Although the relation of pupillary response to affect-arousing stimuli had been noted as early as 1920 (Löwenstein, 1920), the potential of these responses to serve as an index of attitudes was first stimulated by the work of Hess and Polt (1960), who suggested that pupil size is related to the interest value of visual stimuli. These investigators exposed male and female subjects to photographs of male and female figures as well as to other stimuli; they found that male subjects showed greater pupil dilation to a photograph of a female nude relative to a control stimulus than did female subjects. In contrast, female subjects exhibited greater dilation to photos of a partially clothed man, a mother and baby, and a baby alone than did male subjects. A subsequent study by Hess, Seltzer, and Shlien (1965) found that male homosexuals showed greater dilation to photos of men than did heterosexual men.

More critical to the potential use of pupillary response as a bidirectional indicator of attitudes is Hess's (1965) report that disliked or aversive stimuli (e.g., a photo of a shark or several emaciated concentration camp victims) initially produced pupillary dilation but with repeated exposures led to constriction. Although some subsequent research has confirmed Hess's work (e.g., Atwood & Howell, 1971; Barlow, 1969), many other studies have failed to find *both* dilation to positive stimuli and constriction to negative or disliked stimuli (e.g., B. E. Collins, Ellsworth, & Helmreich, 1967; Nunnally, Knott, Duchnowski, & Parker, 1967; Woodmansee, 1970). Reviews of this literature suggest (a) that the least reliable aspect of this research is pupillary constriction to aversive or

negative stimuli, and (b) that dilation, like the GSR, may occur as part of an orienting reflex and may therefore be a better measure of attentiveness to stimuli than affect toward them (Petty & Cacioppo, 1983; Woodmansee, 1970).

Facial Electromyographic Activity. Darwin (1872) theorized that different emotions were linked to different overt facial expressions. Recent attempts to develop a bidirectional physiological measure of attitudes have centered on facial muscle contractions. In current thinking, different emotional or affective states give rise to electrical activity in different facial muscle groups even when the person's face remains relatively passive and expressionless. These covert responses are detectable by modern electromyographic (EMG) techniques (see Cacioppo & Petty, 1979c; Cacioppo, Petty, & Geen, 1989; Petty & Cacioppo, 1983).

Evidence that EMG activity can detect positive and negative emotional states was obtained in a number of studies by Schwartz and his colleagues (G.E. Schwartz, Ahern, & Brown, 1979; G.E. Schwartz, Fair, Salt, Mandel, & Klerman, 1976; Sirota & Schwartz, 1982). When subjects were told to imagine positive events, they showed more EMG activity in the zygomatic (smiling) muscles and less in the corrugator (frowning) muscles than when they imagined negative events (G.E. Schwartz et al., 1976).

Extending this work to attitudes, Cacioppo and Petty (1979a) showed that the presentation of a counterattitudinal message, which presumably evoked negative affect and thoughts, elicited less zygomatic muscular activity than a proattitudinal message. Subjects also exhibited more activity in the corrugator muscles when confronted with a counterattitudinal message compared with a proattitudinal message. This pattern of muscular activity also occurred, although more weakly, when subjects were warned about the topic and position of the message but had not yet received it.

Obviously, attitude measurement through electrophysiology has practical limitations because it requires elaborate instrumentation and respondent cooperation. Although having potential for the study of attitudes in the laboratory, research has focused, not on attitude assessment *per se*, but on the use of these techniques for inferring various cognitive mediators of attitude change in reaction to persuasive messages (see Petty & Cacioppo, 1983; and Chapter 6). Greater exploration of EMG techniques in settings in which respondents confront only questionnaire items or the name of an attitude object would be desirable.

Self-Reports of Affect. In addition to physiological measures, a number of researchers have used self-report questionnaires to measure affective reactions to attitudinal objects. Various investigators have constructed affective measures using Thurstone, Likert, and Guttman scaling techniques (Breckler, 1984a; Kothandapani, 1971; Ostrom, 1969). In addition, Breckler and Wiggins (1989a) had subjects rate their affective responses to attitude objects on the same scales that are commonly used in the semantic differential measure of attitudes (e.g., good vs. bad). Among efforts not involving standard scaling models, Nowlis's (1965) Mood Adjective Check-List (MACL) has been popular. Breckler (1984a, Experiment 1) found that scores on

both positive and negative adjective lists from the MACL correlated moderately highly with his Thurstone measure of affect.

This work on self-report measures of affect appears quite promising, but more research is needed before we can assess their general value as indicators of affect. The practical advantages of self-report questionnaire measures are obvious, especially when compared to the laboratory apparatus required to measure affect physiologically. Nonetheless, self-report measures of affect no doubt share many of the biases of other self-report measures (see section below on Response Distortions).

Behavioral Indicators

Behavioral responses may also serve as indicators of evaluation (see Chapter 1). To serve as an indicator of attitude, a behavior must relate to the dimension of favorability-unfavorability toward the attitude object. Because there are other determinants of any action besides attitude toward the object, whether a person performs an act, in and of itself, cannot necessarily be regarded as a valid indicator of attitude. Whether one attends church on a given Sunday does not necessarily indicate a favorable attitude toward religion in general or toward a specific religion or church. As Fishbein and Ajzen (1974, 1975) noted, indexes of behavior aggregated over multiple acts (or repeated observations) are potentially valid measures of attitude if the various actions have in common some degree of favorableness or unfavorableness toward the attitude object (see Chapter 4). Just like a single belief, a single behavior may not provide a reliable or valid indicator of the attitude. Behavioral responses ordinarily become more indicative of an underlying attitude when aggregated across a variety of attitude-relevant behaviors.

As our discussion of formal scaling models has already shown, items describing behaviors and intentions to act have been used to construct attitude measures by the standard attitudinal scaling techniques (e.g., DeFleur & Westie, 1958; Fishbein & Ajzen, 1974; Kothandapani, 1971; Ostrom, 1969; Rosander, 1937; Triandis & Triandis, 1960, 1965; see Table 2.4). Behavioral instruments not derived from standard scaling models have also been constructed. Triandis (1964) created what he called a "behavioral differential" in which subjects rate on 9-point scales whether they would or would not engage in particular behaviors with the stimulus person. Beginning with 700 descriptions of interpersonal behaviors sampled from novels, Triandis reduced these by eliminating redundancies and low frequency behaviors to 61 socially important and diverse behaviors. Triandis had subjects indicate their willingness to engage in these 61 behaviors with respect to 34 stimulus persons who varied in race and other attributes. A factor analysis of the intercorrelations among mean ratings of the behaviors yielded five meaningful, relatively independent social distance factors. Although Triandis concluded that social distance was not unidimensional, the factors he derived can be regarded as measures of attitude because they express evaluation of social groups.

In most of the behavioral studies we have noted and in many of those discussed in Chapter 4, the investigators did not observe actual behavior but relied on respondents' self-reports of behavior or intentions to behave. These measures, therefore, can suffer

from the same response distortions and biases as other indicators measured by questionnaires (see section below on Response Distortions). That such biases occur in reports of behavior and behavioral intentions is illustrated in a study by Linn (1965), in which female (presumably, white) subjects indicated on a questionnaire their willingness to pose for a photograph in which they would be portrayed as part of an interracial mixed-sex couple. These subjects indicated the acceptability of various possible uses of the photograph, ranging from display to a very limited audience (professional research sociologists) to display to a very wide audience (people who would be exposed to a nationwide campaign advocating racial integration). Four weeks later, the same subjects were confronted with a face-to-face request to pose for the photograph and to indicate the uses they were willing to permit. Although the measure of behavioral intention and the subsequent measure of actual behavior referred to the same behavior, subjects were willing to permit only a more limited display of the photograph than the level they had stated in the questionnaire, probably because social desirability pressures had biased their earlier questionnaire responses. Moreover, the questionnaire scores and these subsequent behavioral scores did not correlate significantly.

In some studies, attitude measures based on overt behaviors have been constructed by aggregating behaviors over acts (see Chapter 4). For example, Tittle and Hill (1967) constructed several behavioral indexes of student political participation in student government: A single-act measure based on documented voting in a previous student election, an index based on a count of the number of times the student had reported voting in the previous four elections, a Guttman scale of participation in eight student political activities, and a Likert scale of participation in ten activities. These indexes were significantly correlated with one another and with measures of attitude based on belief statements that were constructed by several different scaling methods.

Studies that have used indexes of aggregated behaviors typically have merely summed the number of acts that have occurred or were reported (e.g., Weigel & Newman, 1976). Yet indexes of behavioral acts that are intended to serve as measures of attitude would ideally be subjected to the item analysis procedures associated with the traditional attitudinal scaling techniques considered earlier in this chapter.

Reliability and Validity of Attitude Measures

Reliability

Earlier in this chapter, we defined the reliability of a measuring instrument as the extent to which it yields *consistent* results over repeated observations. Another way of thinking about the reliability of a measure is the extent to which it is free from *random error*. As also noted earlier, the reliability of a measure generally is assessed by determining how well scores on the measuring instrument correlate with themselves. This section explains why the correlation between two sets of observations using the same or an equivalent measure provides an estimate of the scale's reliability. We also consider how those two sets of observations may be obtained in practice.

If you suspected that your bathroom scale was somewhat unreliable, you probably would weigh yourself several times and average the weight scores that you obtained. You would probably regard that average as a good estimate of your "true" weight. To assess the unreliability of the scale, you might note how much the obtained weights fluctuated over observations. To be more rigorous, you might even compute a statistic that measured the variability in the observed weights (e.g., the standard deviation, σ). If you thought about this statistic for a moment, it might occur to you that this estimate could be specific to your "true" weight. Therefore, a more general estimate of unreliability could be obtained by sampling other people who varied considerably in their "true" weights, taking repeated observations of their weights on the scale, and computing the standard deviation over all the observations.

Without knowing it, you would have done what *classic true score theory* says one ought to do in order to assess the unreliability of a scale. In classic true score theory, each observed value, X, is viewed as a combination of true score, T, plus random error, e. Symbolically, the relationship is expressed as follows:

$$X = T + e \tag{2.2}$$

T, the true score, is assumed to be a constant for an individual within a given time frame, so Equation 2.2 indicates that the reason that X varies from one observation to the next is because of e. Classic true score theory further assumes that errors and true scores are independent of one another (i.e., they are uncorrelated) and that the expected value of the errors (i.e., their average over many observations) is zero. Therefore, for any individual the average of his or her X scores over many repeated observations is T. That is, a person's true score is the mean of his or her observed scores. Conceptualized in this way, taking the average of the values of your weight as an estimate of your "true" weight makes good sense.

Classic true score theory assumes that the errors in the observed scores of one person are independent of the errors in the observed scores of another person and that true scores are independent of error scores both within or between persons. Then, the variance of the observed scores, σ_X^2, is given by the following expression:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \tag{2.3}$$

This equation states that the variation in observed scores is in part due to the fact that they are based on different individuals with differing true scores and in part due to random error. If the terms in Equation 2.3 are rearranged to express the fact that true score variance is equal to observed score variance minus error variance and both sides of the equation are divided by σ_X^2, we arrive at the following theoretical definition of reliability:

$$\text{Reliability} = \sigma_T^2/\sigma_X^2 = 1 - (\sigma_e^2/\sigma_X^2) \tag{2.4}$$

Equation 2.4 states that the reliability of a measure is the proportion of observed score variance that is true score variance. This proportion is equal to 1 minus the proportion

of observed score variance that is error variance. Reliability would be equal to 1 if all the observed score variance were due to true score variance. As the proportion variance that is due to error increases, reliability decreases.

Suppose that instead of measuring each person repeatedly we only measured each person twice; that is, we used what is known as a *parallel measurement*. Because each person's true score, T, is constant, his or her first observed score, X, would differ from the second observed score, X', only because of random error.[20] Given the assumptions of true score theory about the independence of errors and true scores, the above equations would be applicable to X' as well. Parallel measurement is important because it can be shown that the correlation between parallel measures, $\rho_{XX'}$, yields the proportion of observed score variance that is true score variance (M.J. Allen & Yen, 1979; F.M. Lord & Novick, 1968).[21] That is, defined as the proportion of observed score variance that is true score variance, the reliability of an instrument is given by the correlation between two parallel measurements. Therefore, an estimate of the reliability of an instrument can be obtained by correlating individuals' observed scores on two parallel measures.[22]

Estimates of the correlation between parallel measures can be obtained in several ways. One obvious way is to obtain a measure of each person's attitude twice using the same scale and items. The correlation between scores obtained at Time 1 and Time 2 is known as *test-retest* reliability. Although this method of assessing reliability is straightforward, it has some serious drawbacks. If the test-retest interval is short, people may remember their previous responses and thereby produce a higher estimate of reliability than would be obtained with independent administrations of the scale. Yet, if the test-retest interval is long, differences between the two administrations might reflect changes in the underlying attitude rather than mere random error.

To circumvent the problem of choosing an appropriate interval for retesting, several other ways of estimating reliability have been devised. One such method is to develop *equivalent* or *alternative forms* of the same scale, administer both forms to the same people at the same time, and correlate the scores on the two forms. Alternative forms should have the same mean and standard deviation but differ in their items so that respondents' recall of previous responses would no longer be a problem. The reliabilities of Thurstone scales have frequently been assessed by alternative forms (e.g., Likert, 1932; Seiler & Hough, 1970; Thurstone & Chave, 1929).

Another method of determining the reliability of an attitude scale is to split the items into two parts of equal size (e.g., odd-numbered vs. even-numbered items) and correlate the scores across the two parts. However, splitting the scale into two parts results in a scale that is half as long as the original scale. Because the reliability of a scale increases with the number of items on the test and conversely is reduced by decreasing the number of items, the *split-half* correlation will be lower than that of the original test. The appropriate correction for the lowered reliability correlation between halves is the *Spearman-Brown prophecy formula*, which gives the value of the reliability of a test that is N times longer than the original test. The Spearman-Brown formula is:

$$r_{XX'} = \frac{Nr_{YY'}}{1 + (N-1)\,r_{YY'}}$$

where $r_{XX'}$ = the reliability of an entire scale composed of several components or parts and $r_{YY'}$ = the reliability (correlation) of the parts. In the case of split-half reliability assessment, $r_{YY'}$ would be the correlation between the halves, $r_{XX'}$ would be the reliability of the full scale that is twice as long, and N would equal 2. More generally, Equation 2.5 could be used to determine the effect that lengthening or shortening a scale by a certain amount has on the scale's reliability.

The correlation obtained between two halves of an attitude scale depends on the items that compose the two halves. Splitting a scale by an odd-even or any other method is arbitrary and does not necessarily guarantee that the two parts are equivalent or parallel. Might not a measure based on the average of all possible splits of the items provide a better estimate of the scale's reliability? Why not split the test into many parts, intercorrelate each of the parts with the other parts, and base an estimate of reliability on the average of the correlations between all the various parts?

A measure of reliability that has these properties is Cronbach's (1951) *alpha* (α), which is given by the formula:

$$\alpha = \left[\frac{N}{N-1}\right]\left[1 - \frac{\sum\sigma^2_{Y_i}}{\sigma^2_X}\right] \tag{2.6}$$

Alpha yields an estimate of the reliability of a composite, X, made of N parts, Y_i. In most applications, these parts are considered to be single items. In such cases, $\sigma^2_{Y_i}$ is the variance of the respondents' scores on item i, $\Sigma\sigma^2_{Y_i}$ is the sum of the item variances, σ^2_X is the variance of the respondents' total scores (i.e., each respondent's item scores summed over all of the items), and N is the number of items. When the item variances are equal, the formula for alpha becomes:

$$\alpha = \frac{N\bar{r}_{ij}}{1 + (N-1)\,\bar{r}_{ij}} \tag{2.7}$$

where $\bar{r}_{ij}$ is the average correlation between the N items. Equation 2.7 clearly shows the dependence of α on the intercorrelations among the items.

Alpha is the current standard statistic for assessing the reliability of a scale composed of multiple items (but see Greene & Carmines, 1980, for alternative measures). It is the most appropriate reliability measure to use for Likert and semantic differential scales because these methods assume that the items are parallel sample measures of the same attitude content domain. Alpha is not an appropriate reliability measure for Thurstone and Guttman attitude scales because in those methods the items are regarded as representative of *different* points along the evaluative continuum. The calculation of alpha is ordinarily part of the item analysis procedures discussed earlier in connection with these two attitude measurement techniques. Because alpha considers the degree to which items on a scale intercorrelate with one another, it is often referred to as a measure of *internal consistency* (or *homogeneity* or *equivalence*). Reliability measures of internal consistency (e.g.,

alpha, split-half reliability) are appropriately differentiated from measures of *stability* (i.e., test-retest) because the latter may include changes in true scores over time in addition to random error.

Finally, we note that a high value of alpha is often erroneously assumed to indicate that a scale has a single factor structure. Because alpha, like other measures of reliability, is a function of the number of items on the scale, high alphas can be obtained for scales with many items even when the average intercorrelation between items is only moderate. Moreover, a scale composed of several factors might still yield a high alpha. Because of this fact, investigators constructing Likert scales commonly subject their preliminary scales to factor analysis procedures, as we explained earlier in this chapter.

Validity

The validity of an attitude scale refers to the extent to which the scale truly measures the attitude it is intended to assess. Beneath the superficial simplicity of this definition lurks a problem of considerable complexity. If, like other psychological constructs, attitudes cannot be observed directly and can only be inferred from indicators or instruments designed to measure them, how can we determine whether a particular measure really measures the attitude it claims to measure?

Someone once said that "if something looks like a duck, walks like a duck, and quacks like a duck, it must be a duck." The validation of an attitude measure is much like the validation of whether something is a duck. We must determine whether the measure looks like and behaves like a measure of the presumed attitude. This sort of validation, which is known as *construct validity*, is an ongoing process that is based on theory. That is, either on the basis of specific theory or more general assumptions about attitudes, a valid measure of the underlying attitude should enter into certain relationships and not into other relationships. Thus, construct validity of a scale is determined by certain theoretically based predictions about how the scale should behave in relation to other measures of the same construct and other constructs.

The predictions underlying the construct validation of a particular scale do not necessarily require elaborate theoretical development, however. In many instances these predictions are based on certain generally accepted ideas about the nature and functioning of attitudes. For example, we would expect that a scale designed to measure pro-abortion versus anti-abortion attitudes would yield different average scores for members of a right-to-life group compared with members of a pro-choice group. In fact, this *known groups* method of validation has frequently been used to validate and refine attitude scales. The idea that right-to-life and pro-choice groups should differ in their attitudes toward abortion is so fundamental that investigators would probably discard any scale or items designed to assess attitudes toward abortion that did not differentiate between these groups.

Another common method of assessing the validity of a scale is to see how well it correlates with alternative measures of the same attitude. Donald T. Campbell and Donald W. Fiske (1959) have termed this kind of validation *convergent validity*

However, alternative measures of a construct may correlate with each other not only because they measure the same construct but also because they share common sources of bias or method variance (D. T. Campbell & Fiske, 1959). For example, scores on the original Fascism, Ethnocentrism, Anti-Semitism, and Anti-Negro Scales of the classic study of the authoritarian personality were highly correlated with one another as predicted by theory (Adorno, Frenkel-Brunswick, Levinson, & Sanford, 1950). However, all of the items were worded so that agreement indicated a high level of prejudice (see Chapter 12 for some illustrative items). Subsequent research suggested that these measures were correlated with each other at least in part because they all may have assessed two types of bias—namely, a tendency to agree with items (i.e., acquiescence; Couch & Keniston, 1960), and a tendency to agree with items that express socially undesirable views (A. L. Edwards, 1957a; J. B. Taylor, 1961). Thus, the high relationships observed between these scales could have reflected common sources of bias or systematic error.

The authors of the authoritarian personality study argued that a fascist personality as well as ethnocentric, anti-Semitic, and anti-Negro attitudes, although related, were different constructs. However, the very high correlations obtained between these scales called into question the assumption that the instruments really measured different concepts. It was thus possible that these measures lacked *discriminant validity*, the ability to distinguish themselves as measures of unique constructs.

More generally, Campbell and Fiske (1959) proposed that convergent validity as well as discriminant validity are essential components of construct validity. To demonstrate convergent validity, an instrument designed to measure a particular construct should correlate highly or converge with other measures of that construct. In addition, for discriminant validity, the instrument should not correlate too highly with measures of different constructs. Campbell and Fiske suggested that the convergent and discriminant validity of various measures could be examined in what they called a *multitrait-multimethod matrix*.

Table 2.11 provides a hypothetical illustration of a multitrait-multimethod matrix in which there are three different attitudes each measured by three different methods. The entries in the matrix are correlations between measures. The *r*s in the main diagonal represent reliabilities (e.g., alphas), which express the extent to which a measure correlates with itself. The *v*s in the other diagonals represent validity coefficients, which indicate the extent to which different measures of the same construct correlate or converge with each other. Campbell and Fiske argued that the magnitude of these validity coefficients should not be judged in terms of their statistical significance or in absolute terms but in relation to the reliability coefficients and the other correlation coefficients in the matrix. Judgment of the validity coefficients in relation to the reliabilities is important because the reliabilities of the various measures set an upper bound for the validity coefficients. According to classic true score theory, a measure's validity coefficient cannot be greater than the square root of its reliability coefficient (F. M. Lord & Novick, 1968).[23] The observed reliabilities of the measures in the multiattitude-multimethod matrix thus provide some indication of the magnitudes of validity coefficients that are attainable.

TABLE 2.11

Hypothetical Multiattitude-Multimethod Matrix

	Method 1			*Method 2*			*Method 3*		
	A1	*A2*	*A3*	*A1*	*A2*	*A3*	*A1*	*A2*	*A3*
Method 1									
A1	r								
A2	m	r							
A3	m	m	r						
Method 2									
A1	v	h	h	r					
A2	h	v	h	m	r				
A3	h	h	v	m	m	r			
Method 3									
A1	v	h	h	v	h	h	r		
A2	h	v	h	h	v	h	m	r	
A3	h	h	v	h	h	v	m	m	r

Note: A1, A2, and A3 are three different attitudes; r = reliability coefficient; v = validity coefficient; m = monomethod-heteroattitude coefficient; h = heteromethod-heteroattitude coefficient.

The validity coefficients in the matrix should also be examined in relation to two other sets of correlations: correlations between different attitudes measured by *different* methods (*h*s in Table 2.11) and correlations between different attitudes measured by the *same* method (*m*s in Table 2.11). In practice, neither of these sets of coefficients can necessarily be expected to be equal to zero. The magnitude of these correlations depends on how different the attitudes toward the various objects are and how different the methods are from one another. If the measures share a common source of bias (e.g., because they are all questionnaire measures or were measured by a particular kind of scale), the validity coefficients may be quite high because of the common method variance. The correlations between attitudes toward different objects assessed by either the same or different methods could also be high because the respondents do not discriminate between attitude objects and thus the measures assess the same attitude (e.g., toward minority groups in general rather than toward different groups). By specifying that the validity coefficients in the matrix must exceed both the correlations between different attitudes measured by the same methods and the correlations between different attitudes measured by different methods, Campbell and Fiske (1959) required that the measures exhibit both convergent and discriminant validity.

The multitrait-multimethod approach has been used by a number of attitude researchers to determine the convergent validity of scales constructed by various scaling techniques (e.g., Jaccard et al., 1975; Kothandapani, 1971; Ostrom, 1969). However, the approach as formulated by Campbell and Fiske did not include a way of statistically evaluating the relationships observed in the matrix. Among contemporary

investigators there is general agreement that the best approach is through structural modeling and confirmatory factor analysis (Alwin, 1974; Bohrnstedt, 1983; Kenny, 1979; Marsh & Hocevar, 1988; Schmitt & Stults, 1986; Widaman, 1985; see note 19).

Another important aspect of a measure's construct validity is that it enter into relationships that are theoretically expected. A study by Hendrick and Seyfried (1974) that illustrates this point was built on the finding that people like others who are attitudinally similar (Berscheid, 1985; Byrne, 1971; see Chapter 9). In this experiment, subjects were presented with a persuasive message designed to produce attitude change and had their attitudes measured immediately thereafter. A day later they were asked to indicate their liking for two persons after seeing these persons' ostensible responses to an attitude questionnaire. One person's questionnaire responses corresponded to the attitude expressed by the subjects prior to having their attitudes changed and the other person's responses corresponded to the presumably changed attitude. The logic of Hendrick and Seyfried's test of validity was that if the subjects had truly changed their attitudes as a result of the persuasive message, they should indicate greater liking for the person who exhibited an attitude corresponding to their changed attitudes. If the change was fleeting or non-genuine, subjects should prefer the person who held an attitude corresponding to their old position. The results of the study indicated that subjects preferred the person whose attitude on the issue corresponded to their newly changed attitudes. Therefore, the study suggested that the attitude changes observed on the attitude scale were real and somewhat enduring.

Although Hendrick and Seyfried were not interested primarily in testing the validity of their scale but in demonstrating the validity of the observed changes on the scale, their experiment established the validity of their attitude measurement. Here, though, the construct validity of their scale of measurement was not established by its convergent and discriminant validity in relation to other methods of measurement but by its ability to reflect a known relationship between attitude similarity and liking.

Various expositions often mention an additional form of construct validity known as *criterion validity*. This form of validity refers to the extent to which scores on the measuring instrument are correlated with some external criterion. For example, do scores on an attitude scale predict behavior? When scores on the criterion measure are obtained within the same time frame as scores on the instrument to be validated, this form of criterion validity is known as *concurrent validity*. When scores on the criterion are obtained at a subsequent point in time, the form of criterion validity is known as *predictive validity*.

Consideration of criterion validity immediately raises the question of what should serve as a criterion for validating an attitude measure. In applied contexts the answer to this question follows from the reasons for creating the attitude measure. That is, attitude measures often are created to predict some aspect of behavior (e.g., votes for a candidate or party; employee absenteeism; purchases of a particular product). In these applied situations, validity is determined by whether the measure predicts what it was designed to predict. If it does not, then it is an invalid predictor in this practical sense. Nonetheless, the instrument itself could have reasonable validity as a measure of attitude because a failure to predict a specific behavior may arise from a variety of

causes (see Chapter 4). Moreover a particular measure could be valid for predicting one criterion and not another.

When attitude measures are created primarily for the scientific purpose of understanding the processes underlying attitude change, assessment of the validity of the attitude measure may hinge on theory-relevant predictions (for example, that attitudes polarize when people are given an opportunity to think about the attitude object; see Chapter 12). As relevant theory evolves and a measure enters into more and more empirical relationships, the construct validity of the measure increases, as does the validity of the theory. Thus, the establishment of construct validity, of which criterion validity is a part, is an ongoing process.

Response Distortions

As this chapter has shown, most attitude measures rely on self-reports of beliefs, feelings, or behavior. This practice is potentially problematic because people may evade answering questions or distort their reports to protect their privacy, to avoid legal prosecution, to gain economic advantage, to obtain social approval and avoid social disapproval, and to project or protect particular identities. *Response distortions* of these and other types could produce systematic errors in attitude measurement.

Attitude researchers typically adopt several strategies to reduce response distortions. One strategy is to embed the measure of interest among items that are of little or no interest to the researcher. The use of such *filler items* is intended to disguise the researcher's interest from the respondents in order to decrease their efforts to provide answers in accordance with their perception of the researcher's or interviewer's expectations. Another strategy is to enlist respondents' cooperation by assuring them of the acceptability of all responses. Generally, they are told that there are "no right or wrong answers. The correct answer is an honest and truthful answer." Furthermore, the respondents are assured of confidentiality by the promise that no one but the research staff will ever know how each individual reacted and that all reports of the research will present the data aggregated across the respondents. In addition, respondents usually complete attitude scales under conditions of anonymity, that is, without giving their names or providing other identifying information.

Research on the efficacy of some of these strategies for reducing response distortion suggests that these strategies are, at best, only partially successful in reducing response distortion (see Bradburn, 1983; Nederhof, 1985; Schuman & Kalton, 1985; Sudman & Bradburn, 1974). Consequently, several other techniques have been developed to reduce motivated distortions.

Bogus Pipeline. The efficacy of the bogus pipeline in reducing response distortions stems from respondents' beliefs that their self-reports are subject to validation. Specifically, the bogus pipeline attempts to control response distortions by leading respondents to believe that the investigator has a foolproof procedure for detecting their true attitudes. Adapted by E.E. Jones and Sigall (1971) from a technique introduced by Gerard (1964), the bogus pipeline typically uses fake electronic

apparatus and a set of electromyographic (EMG) electrodes that are attached to respondents' arms. Respondents are led to believe that this apparatus records minute muscular contractions that yield a precise assessment of their true beliefs and feelings. A meter, controlled by a confederate in an adjacent room, ostensibly provides output from the EMG machine to be viewed by the respondent. In addition, subjects are given a steering wheel connected to a pointer on a meter that allows them to rate the attitude object by turning the wheel (see Sigall & Page, 1971). The electrodes, they are told, can predict how far and in what direction they will turn the wheel. To demonstrate to the respondents that the electrodes work, the EMG output meter then predicts their responses to a number of questions. The meter actually reproduces responses that the respondents gave on another occasion that was apparently unrelated to the pipeline study. On several questions respondents may even be invited to try to "fake out" the EMG machine. After the respondents are convinced that the pipeline works, they are asked to see "how in touch they are with their true feelings" by predicting the EMG results by responding to attitude items using the steering wheel. In essence, the bogus pipeline is a fake "truth detector," but respondents are made to believe it is real. Its main premise, of course, is that respondents' motivations to distort their responses will be reduced if they are subjected to this procedure.

Tests of whether the bogus pipeline reduces response distortion have yielded controversial findings. Sigall and Page (1971) found that white subjects' stereotypes about blacks were more unfavorable and their stereotypes about Americans more favorable under bogus pipeline conditions than under standard rating scale conditions (although Schlenker, Bonoma, Hutchinson, & Burns, 1976, were unable to replicate these findings fully). Other studies have found differences in responses obtained in bogus pipeline and standard rating conditions that are in keeping with the idea that the bogus pipeline reduces social desirability or impression management concerns (Arkin, Appelman, & Burger, 1980; Gaes, Kalle, & Tedeschi, 1978; R.A. Page & Moss, 1975; Riess, Kalle, & Tedeschi, 1981; see also Ostrom, 1973). In contrast, Cherry, Byrne, and Mitchell (1976) obtained no differences between conditions and raised the possibility that the bogus pipeline may heighten conformity to the demand characteristics of the experiment (Orne, 1962) among subjects high in social desirability. However, the evidence on this point is inconclusive (Arkin & Lake, 1983; Byrne & Cherry, 1978; Gaes, Quigley-Fernandez, & Tedeschi, 1978). Also, E.E. Jones and Sigall (1971) noted that the pipeline may cause respondents to focus more on feelings and affect toward the attitude object than they ordinarily would in responding to standard assessment techniques.

The best evidence in support of the bogus pipeline as a means of reducing response distortion comes from two studies reported by Quigley-Fernandez and Tedeschi (1978). In these experiments, subjects who were waiting to participate in an experiment overheard information about the correct answers to a test given in the experiment from someone who presumably had just participated in that experiment. Later in the experiment, these subjects were questioned under either standard or bogus pipeline conditions about whether they had previously heard anything about the test. Both studies showed higher "confession" rates with bogus pipeline assessment.

Related to the bogus pipeline method are a number of studies that have obtained differences in self-reports under standard conditions compared with conditions in which the respondents are led to believe that their reports will be validated in some way (Arkin & Lake, 1983; Bauman & Dent, 1982; Evans, Hansen, & Mittelmark, 1977; P. C. Hill, Henderson, Bray, & Evans, 1981). For example, Bauman and Dent (1982) compared self-reports of smoking behavior by respondents who did or did not have prior knowledge that they would be asked subsequently to provide breath specimens that would reveal their smoking behavior. As an objective measure of smoking behavior, breath specimens from all respondents were analyzed for carbon dioxide levels. According to this test, among adolescents who had smoked recently, only 64 percent indicated under standard self-report conditions that they had smoked within the last four hours. In contrast, 86 percent of those who were aware that the objective measure would be used reported smoking within the last four hours. In another variant of the standard bogus pipeline method, Jamieson and Zanna (1983) found that subjects who expected that their attitudinal responses would be validated by a "lie detector" failed to show attitude shifts that normally occur for self-presentational purposes in anticipation of a counterattitudinal message.

Randomized Response Technique. Warner's (1965) randomized response technique (RRT), like the bogus pipeline, was designed to reduce response distortion in answering questions that are sensitive or potentially embarrassing. Unlike the bogus pipeline, it does not require the use of deception or elaborate apparatus and therefore has wider applicability. Essentially the randomized response technique attempts to eliminate refusals to answer and response distortions by guaranteeing respondents that no one can know for certain whether the answers they gave were in response to the sensitive question.

In Warner's (1965) original model of the RRT, the respondent is confronted with two questions, the sensitive question (e.g., Are you in favor of quarantining people with AIDS?) and its logical complement (e.g., Are you *not* in favor of quarantining people with AIDS?). Through the aid of a randomizing device (e.g., a die), respondents are directed to answer either Question 1, the sensitive question, or Question 2, its logical complement. For example, respondents may be instructed to roll a die and conceal its outcome from the interviewer. They may be told that, if the roll of the die results in a 1, 2, 3, or 4, they are to answer Question 1, but if the roll of the die results in a 5 or 6, they are to answer Question 2. They are told not to disclose which question they answered but merely to report the answer. Thus, a response of "Yes" can mean that the person either is or is not in favor of quarantining people with AIDS. Because only the respondents know which question was answered, complete confidentiality has been guaranteed to the respondents. Warner reasoned that this guarantee would be sufficient to eliminate refusals to answer and response distortions.

Despite lack of knowledge of which question the respondent answered, the application of elementary probability theory yields an estimate of the proportion of people in the population who favor the quarantine of people with AIDS. For example, if P is the probability that a respondent is directed to answer Question 1 and π is the proportion

of people in the population who favor quarantine, then $P\pi$ is the proportion of people who would answer "Yes" to Question 1. Similarly, $(1 - P)$ is the probability that a respondent is directed to answer Question 2, and $(1 - \pi)$ is the proportion of them who would answer "Yes" because they do not favor quarantine. Assuming that all respondents answer as instructed and truthfully, the probability of a "Yes" response, τ, is given by:

$$\tau = P\pi + (1 - P)(1 - \pi) \tag{2.8}$$

Because P is known and τ can be obtained from the data, Warner showed that Equation 2.8 can be solved for π to obtain an estimate of the proportion of people in the population who are in favor of the quarantine of people with AIDS. The sample estimate, $\hat{\pi}$, is given by:

$$\hat{\pi} = [\hat{\tau} + P - 1]/(2P - 1) \tag{2.9}$$

when $P \neq \frac{1}{2}$ and where $\hat{\tau}$ is the obtained proportion of "Yes" responses in the sample. Thus, the randomized response technique does permit inferences about the population parameters even though the question any given respondent answered is unknown.

Despite its simplicity, the randomized response technique does have a major drawback: The randomization process introduces an additional source of random error that makes estimation of population parameters less efficient than it is for direct questioning. Consequently, much of the research on the randomized response technique since Warner's initial presentation has been concerned with the development of alternative RRT models that would make the technique more statistically efficient and, thereby, more practical. One widely used development is the unrelated question RRT (Greenberg, Abul-Ela, Simmons, & Horvitz, 1969). In this variant, the respondent is directed with probability P to answer the sensitive question and with probability $(1 - P)$ to answer a totally innocuous question (e.g., Were you born in the month of April?). Other major developments include extensions to questions involving more than two response categories (Abul-Ela, Greenberg, & Horvitz, 1967; Liu & Chow, 1976; Liu, Chow, & Mosley, 1975) and to questions requiring a quantitative or numerical response (Greenberg, Kuebler, Abernathy, & Horvitz, 1971; Himmelfarb & Edgell, 1980). This literature has been reviewed and summarized by Horvitz, Greenberg, and Abernathy (1975) and by Fox and Tracy (1986), and comprehensive bibliographies are available in Nathan (1988) and Himmelfarb and Edgell (1988).

Individuals' responses to questions obtained by various RRT models can be correlated with each other or with other variables to investigate their relationships. That is, even though respondents' answers in an RRT procedure are not always in response to the sensitive question, their respective answers can be scored and treated as individual values in standard formulas for various correlation coefficients. The obtained correlations will be attenuated relative to the true correlation between the variables, but the correlations can be corrected for this attenuation (see Fox & Tracy, 1984; Himmelfarb & Edgell, 1982; Kraemer, 1980). Statistical tests of the corrected

correlation coefficients must be adjusted for the additional error introduced by an RRT procedure (Edgell, Himmelfarb, & Cira, 1986).

A number of experiments have compared responses obtained through direct questioning and an RRT procedure. Many of these studies involved questions about drug and other abuses or illegal activities. In general, higher rates of drug and alcohol use and fewer refusals to respond were reported under RRT conditions than under direct questioning (G. H. Brown & Harding, 1973; Goodstadt & Gruson, 1975; Reaser, Hartsock, & Hoehn, 1975; Zdep, Rhodes, Schwarz, & Kilkenny, 1979). Also, RRT compared with standard interview conditions produced higher rates of reported child abuse (Zdep & Rhodes, 1976) and reported abortions (I-Cheng, Chow, & Rider, 1972; Krotki & Fox, 1974). Yet, the strongest evidence for the superior validity of the RRT was obtained in a laboratory study by Shotland and Yankowski (1982), which resembled Quigley-Fernandez and Tedeschi's (1978) test of the bogus pipeline. Subjects, while waiting to participate in an experiment, overheard information about the correct answers to a test given in the experiment. When questioned in a face-to-face interview, 27 percent reported receiving this information and only 10 percent said they had used it. In the RRT condition, 64 percent confessed hearing the information, and 80 percent reported using it. Finally, attesting to the fact that social desirability does bias conventional self-reports and that the RRT reduces this distortion, Himmelfarb and Lickteig (1982) found a significant relation between the social desirability and undesirability of behaviors and attitudes and the extent to which these behaviors and attitudes are overreported and underreported on an anonymous self-administered questionnaire compared with a questionnaire completed through the use of an RRT procedure. Although more evidence is needed on the validity of the RRT, the research has supported the ability of the technique to reduce distortion in self-reports of socially undesirable behaviors.

Response Sets. In addition to motivated response distortions, a variety of more subtle response sets have been implicated as sources of invalidity in psychological measurement (Cronbach, 1946, 1950; Guilford, 1954). *Response sets* are tendencies to respond in particular ways that are not tied to the particular content of items or scales. These response tendencies have traditionally been viewed as reflecting consistent habits within individuals that vary across persons (Guilford, 1954, p. 453). It is also usually assumed that respondents are unaware of their response sets.

Response sets that have been considered particularly important in attitude measurement include tendencies to answer *yes* or to *agree* with items (acquiescence; Couch & Keniston, 1960), tendencies to give or avoid giving extreme responses, and tendencies to be cautious and noncommittal by choosing the neutral category. Although the pervasiveness of these response sets has been questioned (Rorer, 1965), strategies for minimizing them are frequently implemented by researchers. For example, to correct for acquiescence sets, researchers often include equal numbers of positively worded and negatively worded items on their scales. And to avoid noncommitment tendencies, investigators sometimes omit neutral response options (see Table 2.10), thereby forcing respondents to choose among nonneutral alternatives.

Because techniques for eliminating the effects of response sets and motivated response distortions are not always successful, attitude researchers have attempted to develop disguised, unobtrusive, or indirect measures of attitudes that do not rely on self-reports or even on verbal measures (see D.T. Campbell, 1950; Kidder & Campbell, 1970; Sechrest & Belew, 1983; Webb, Campbell, Schwartz, & Sechrest, 1966; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Some of these instruments were considered in the earlier discussion of specific indicators of attitudes (e.g., physiological measures, Hammond's error-choice method).

Context and Other Response Effects

Survey researchers have identified a number of other factors that may bias questionnaire responses. A detailed examination of these *response effects* is beyond the scope of this chapter (see Bradburn, 1983; Schuman & Presser, 1981; Schuman & Kalton, 1985).[24] Here, we consider a subset of response effects that are particularly relevant to social psychological studies of attitudes. Because such studies are likely to use closed-ended, self-administered questionnaire measures of attitudes, interviewer effects and effects due to method of administration (e.g., face-to-face interviews versus telephone interviews) are of marginal concern. Also of marginal concern are response alternative order effects, which are minimized when attitudes are measured by questionnaires that allow respondents to review all the available alternatives. Schwarz, Strack, Hippler, and Bishop (1991) have provided an excellent discussion of how differences in administration method produce various types of response effects.

The question of whether to include a noncommittal or neutral response category in attitude assessment has been of considerable interest to survey researchers (e.g., Kalton, Roberts, & Holt, 1980; Schuman & Presser, 1981). Schuman and Presser's (1981) analysis of research on the *don't know* response category concluded that including or excluding an explicit don't know or neutral response alternative has little overall impact on estimates of the relative proportion of people who favor, versus oppose, an attitude issue. Nonetheless, Krosnick and Schuman's (1988) meta-analysis of the impact of attitude strength on response effects found that people with weaker attitudes tended to use neutral response alternatives more than did people whose attitudes were stronger (see discussions of attitude strength in Chapters 3, 4, and 12). Similar findings were obtained by Bishop (1990) in a meta-analysis of 18 telephone interview experiments.

Survey researchers have also examined a variety of response effects that reflect the order in which questions are asked. When questioned about either fictitious or unfamiliar issues, respondents frequently use the context created by earlier questions to interpret the question (see Schwarz & Strack, 1991). For example, Strack, Schwarz, and Wänke (cited in Schwarz & Strack, 1991) asked German college students about their attitudes toward an "educational contribution." For half of the students, this question was preceded by a question about the amount of tuition fees paid by students at U.S. universities. For the other half, the target question was preceded by a question about the amount the Swedish government pays college students for financial support. Attitudes

were more favorable toward the "educational contribution" item when the question preceding it concerned the fees paid *to students*, rather than *by students*.

Question order effects are not limited to ambiguous questions and unknown issues. For example, Hyman and Sheatsley (1950) varied the order in which respondents were asked two questions about allowing newspaper reporters to enter foreign countries in order to report back to their own countries. One question asked whether U.S. reporters should be allowed into communist countries, and the other asked whether communist reporters should be allowed into the United States. The results indicated that respondents were *more* willing to allow communist reporters into the United States if they had first answered the question about U.S. reporters. Similarly, respondents were *less* willing to allow U.S. reporters into communist countries if they had first answered the question about communist reporters. According to Schuman and Ludwig (1983), who reported several replications of this finding, respondents' answers to the first question that was posed reflected their attitudes toward the United States and communism. Responses to the second question, however, reflected respondents' tendencies to be "even-handed." For example, if they had just endorsed the rights of U.S. reporters to enter communist countries, they presumably felt compelled to extend the same rights to communist reporters.

Feldman and Lynch (1988) have suggested a somewhat different and more general interpretation of this and other question order effects. In their view, people's responses to attitude queries reflect both constructive and memorial processes. Thus, earlier items often make salient to respondents information that they might not ordinarily consider in responding to subsequent, thematically related questions. In the Hyman and Sheatsley (1950) study, then, respondents exposed to the U.S. reporter–communist reporter order may have based their responses to the latter item on information that came to mind in responding to the U.S. reporter item, information that may not have been accessible to respondents exposed to the communist reporter–U.S. reporter order.

A more recent example of how earlier questions can bias responses to subsequent, related questions is contained in the results of the National Crime Survey (cited in Schuman & Presser, 1981). Respondents were asked to report victimization experiences. Half of the respondents answered 16 attitude items about crime prior to reporting their actual experiences, whereas these questions were omitted for the remaining respondents. Reports of victimization, especially of less serious crimes, were higher for the sample who responded to the earlier attitude items. Apparently, the attitude questions activated memories or images of victimization experiences. Another item order effect was reported by Turner and Kraus (1978). Some of their respondents were asked whether spending should be increased in 11 federal spending areas (e.g., defense, environment) and were then asked a general question about taxes (whether they paid overly high federal income taxes). Other respondents were asked about the spending areas *after* responding to this general question. Results showed that when the spending area questions came first, 14 percent fewer respondents thought income taxes were too high. The specific questions about federal spending areas presumably made respondents more aware of the many services funded by income tax revenue. As Schuman and Presser (1981) suggested, respondents may have derived their general conclusions

about necessary income tax levels from the implications of the specific spending items. Such *general-specific* versus *specific-general* order effects have also been documented in other topic domains (e.g., Kalton, Collins, & Brook, 1978; Krosnick & Schuman, 1988; Schuman & Presser, 1981; Schwarz, Strack, & Mai, 1991; T. W. Smith, 1982).

Although various response effects have been demonstrated by survey researchers, the conditions under which they occur and their relation to seemingly relevant moderator variables such as attitude strength are not well understood at present (see Schuman & Kalton, 1985). Attempts to broaden our knowledge of context and other response effects are now focusing on ideas from basic research and theory in social cognition and information processing. Many of these ideas are discussed in this book (see Chapters 3 through 8). Significant conceptual contributions to this area have been made by Tourangeau and Rasinski (1988), Strack and Martin (1987), Schuman and his colleagues (e.g., Krosnick & Schuman, 1988; Schuman & Presser, 1981), and Schwarz and his colleagues (e.g., Hippler, Schwarz, & Sudman, 1987; Schwarz, Strack, Hippler, & Bishop, 1991; Schwarz & Sudman, 1992). This collaboration between survey researchers and cognitive social psychologists holds considerable promise for enriching our understanding of both basic processes in social cognition and attitudes as well as applied issues in attitude measurement (for an overview of this collaboration, see Jobe & Mingay, 1991).

Single- Versus Multiple-Item Measures

Surveys frequently measure attitudes by a single evaluative item (e.g., Do you approve of the way the President is doing his job?). The justification for using single-item measures is primarily economic: Surveys are costly, and the costs increase with the number of questions asked. As we have explained, however, there are good methodological reasons to prefer a composite score based upon a multiple-item index.

Multiple-item measures can compensate for the limitations inherent in most individual items. Any single item typically contains nuances of meaning and tone that may exert unintended influences on subjects' responses. Indeed, survey research has provided abundant evidence that slight differences in item wording can exert pronounced effects on responses. For example, in a classic study by Rugg (1941), one national sample was asked: "Do you think the United States should allow public speeches against democracy?" Another comparable sample was asked: "Do you think the United States should forbid public speeches against democracy?" Approximately 20 percent more of the respondents were willing to *not allow* such speeches than were willing to *forbid* them. This *forbid-versus-allow* effect has been replicated in several subsequent studies (Schuman & Presser, 1981). Similarly, T. W. Smith (1987) reported the results of a number of surveys that yielded systematic differences in the percentages of people indicating favorable attitudes toward the government doing more for "the poor" or "the unemployed" than for people "on welfare."

Clauses or phrases that are ostensibly irrelevant to the main issue posed in questions have also been shown to have a substantial impact on survey responses. For example,

Cantril (1940) asked one group of respondents: "Do you think the U.S. should do more than it is now doing to help England and France?" When the phrase "in their fight against Hitler" was added to the end of this question, the percentage saying yes increased from 13 percent to 20 percent. Similarly, support for sending U.S. troops to intervene in regional wars such as Vietnam was increased by adding the phrase "to stop a Communist takeover" (Mueller, 1973; Schuman & Presser, 1981).

Although research has documented that wording effects can occur, the extent to which minor wording differences influence survey responses is presently unknown, and a systematic understanding of when and why wording effects occur has yet to be achieved (Schuman & Kalton, 1985). Even in the absence of systematic theory about question wording, a number of principles of good question writing have been articulated (e.g., A.L. Edwards, 1957b; Payne, 1951; Sheatsley, 1983; Sudman & Bradburn, 1982).

In our discussion of various scaling techniques, we saw how poor items can be eliminated through item analysis procedures. By examining an item's operating characteristic, we can frequently tell whether the item is appropriate or not. The operating characteristic of an item relates scores on an item to attitude scores obtained over many other items (i.e., to total scores). With a single item, however, we cannot determine the item's operating characteristic nor can we correlate scores on that item with scores on other items. In essence, we lack an internal way of distinguishing between good and bad items when we only have a single item. Moreover, we have no way of estimating the reliability of the item or of estimating the magnitude of the relationships that would exist, in the absence of errors of measurement, between the variable that the item assesses and other variables.

In discussing the split-half method of assessing reliability, we noted that the correlation between the two halves is less than the reliability of the total scale. The Spearman-Brown formula was introduced to correct for this reduction. What may not have been apparent in the formula, however, is the general relationship that exists between the average reliability of items and the number of items in the scale. This relationship is shown in Figure 2.8 for several different possible item reliabilities. Although the relationship between the reliability of the total score and the number of items is one of diminishing returns, Figure 2.8 shows that the reliability of the total score always increases as the number of items increases if the average inter-item correlation remains constant. Indeed, given a particular average inter-item reliability estimate (correlation), one could determine from the Spearman-Brown formula how many items of that same reliability would be needed to boost the reliability of the total scale to any particular value. Thus, multiple-item measures have the added advantage of greater reliability.

In our discussion of validity, we noted that the validity of any measure is in part determined by the reliability of the measure. Unreliable measures of variables attenuate the relationships between the variables and, therefore, make it more difficult to observe the true relationships between variables. A reliable measure not only yields consistent scores from one observation to another, but also has greater potential for correlating highly with other variables. The value of aggregated measures for establishing strong

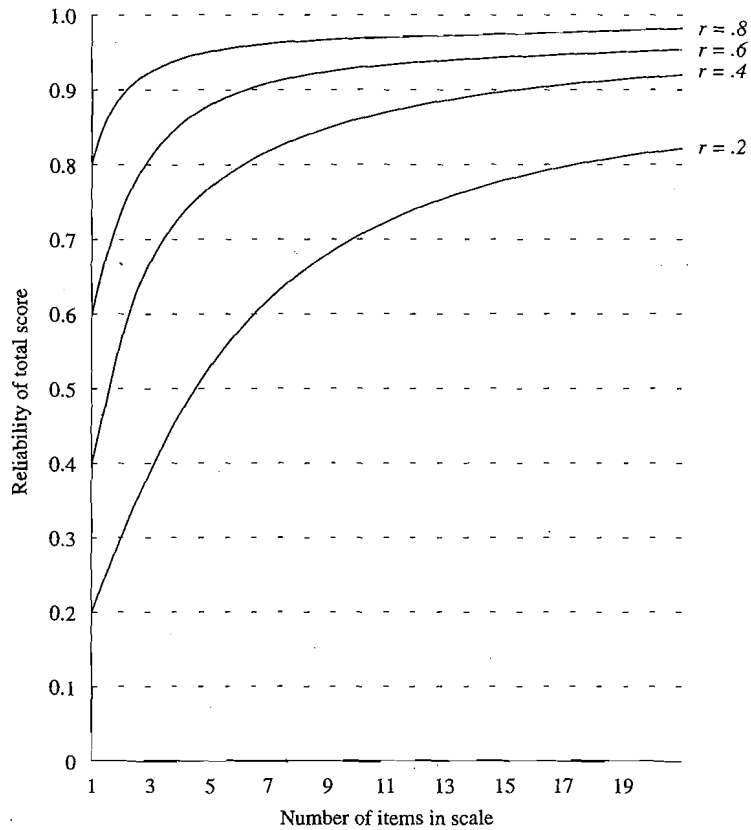

FIGURE 2.8. Total scale reliability as a function of the number of items in the scale and inter-item reliability values of .2, .4, .6, and .8.

relations between variables is illustrated in Chapter 4 by Fishbein and Ajzen's (1974, 1975) research on the attitude-behavior relationship.

This chapter has emphasized attitude measurement methods based on multiple-item scales because of these well-known psychometric principles. Yet, as will become apparent in the following chapters, many successful studies of attitudes have assessed attitudes informally by one or two rating scales. These successes indicate that these single-item measures are reliable enough to detect mean differences between groups of reasonable size when variables are powerfully manipulated in carefully controlled settings. Yet, they may not be sufficiently reliable to correlate strongly with other measures, particularly hypothesized mediating variables. The reliability of our measures could be improved and, therefore, the relationships between variables enhanced with multi-item measures. Many of the other techniques discussed in this chapter should also prove useful for the development of more reliable and valid measures of attitudes.

Conclusion

Sixty years have passed since Thurstone initiated the development of formal scaling techniques of attitude measurement. Since then, a variety of both simple and mathematically complex models have been developed and applied to attitude measurement. The more popular unidimensional techniques have been summarized here.

Investigators of the 1930s, 1940s, and early 1950s were particularly interested in developing attitude measures and assessing their validity and susceptibility to bias. Such concerns continue today mainly among applied researchers. Interest in attitude measurement and related methodological issues declined in the late 1950s among most attitude researchers. By then, existing research suggested that attitude measures based on the Thurstone, Likert, Guttman, and semantic differential scaling techniques usually correlated quite well with one another, with none having any particular advantage over the others except for differences in their ease of construction. From the 1960s onward researchers became more involved with the testing of theoretical propositions related to issues of attitude formation and change.

Interest in attitude measurement may be increasing once more. This renewal of interest may be traced to several factors. First, there is increased concern about the applied relevance of attitude research. Applied researchers desire to develop scales for measuring attitudes toward a variety of contemporary issues. Because these scales are intended for general use beyond a single investigation, they are ordinarily constructed in accord with modern psychometric principles. These principles encourage multiple investigations to demonstrate that scales have sufficiently high reliability and validity to be useful for predicting and understanding the correlates of attitudes. Second, there is growing awareness that the study of relationships between theoretical variables cannot be divorced from the issue of how these variables were measured. This point was made salient in research on the attitude-behavior relation, where single-item measures of behavior contributed to the erroneous conclusion that attitudes were not related to behavior (see Chapter 4). Also, the increased use of sophisticated statistical techniques (e.g., structural equation models with latent variables; see Chapter 4) for testing mediational models of attitude change focuses concern on measurement issues.

Many of the measurement techniques considered in this chapter were imported from other areas of psychology at a time when there was little or no attitude theory. In contrast, as the chapters that follow attest, there is now an abundance of theories. A closer integration of theory and research on attitudes with the methodology of attitude measurement might prove to be quite fruitful. For example, current interest in attitude structure (see Chapter 3) might well lead to a greater emphasis on multidimensional aspects of attitude measurement and to the development of new techniques for measuring attitudes. The traditional scaling techniques described in this chapter could be adapted more fully for the assessment of attitude structure.

Particularly relevant to issues of attitude measurement is current theory on stages of information processing. Indeed, Tourangeau and Rasinski (1988) provided an excellent analysis of context effects in such terms. More generally, not all methods of assessing

attitudes require that respondents engage in the same underlying processes, even when the attitude scores that are produced by various methods correlate highly with one another. Some methods require more elaborate and deeper cognitive processing, particularly when respondents must consider the implications of a large number of belief statements. Other methods, especially single-item measures, may encourage respondents to answer on the basis of a stored general evaluation of an attitude object. Moreover, certain methods may heighten respondents' self-presentational concerns or provide superficial cues that may guide their answers to attitudinal items. Obviously, these issues (e.g., depth of processing, self-presentational pressures, use of attitude-relevant cues) are of considerable interest in attitude theory and are considered throughout this book. Greater recognition of the interdependence of theory and measurement has considerable potential for the future of attitude research.

Notes

1. A linear transformation is of the form: $Y = bX + a$, where X is the original set of values, Y is the new set of values, b is the ratio of the unit of measurement of Y to the unit of measurement of X, and a is the origin or zero point—the value of Y when $X = 0$. For example, the transformation of Celsius to Fahrenheit is given by the linear equation $F = (9/5)C + 32$, where 9/5 is the ratio of the units of measurement of F and C and the 32 is the value of F when $C = 0$.

2. Measurement theorists often define different types of scales or levels of measurement by classes of *admissible scale transformations* (e.g., Krantz, Luce, Suppes, & Tversky, 1971; Suppes & Zinnes, 1963). Statements are regarded as meaningful only if they are invariant under all admissible scale transformations. In the case of an interval scale, only positive linear transformations are admissible. By that criterion, the statement that D's attitude is 2.5 times more favorable than B's is not a meaningful one because it is not invariant under all positive linear transformations (see Michell, 1986).

3. In the measurement and scaling literature these checks for consistency are often referred to as *internal consistency tests*. We have omitted the word *internal* and substituted *checks* for *tests* to avoid confusion with certain methods of estimating reliability that are frequently labeled measures of internal consistency (e.g., alpha; see subsequent discussion).

4. These statements about the Davison and Sharma (1988, 1990) findings simplify certain highly technical and important conditions that must be met before conclusions concerning measured variables apply to the underlying latent variable. Moreover, their proofs concern tests of the null hypothesis about differences between means or associations between variables. Conclusions about the size of a difference between means or the strength of an association do not apply to the underlying latent variable unless the observed variable was measured on an interval scale. Also, the results Davison and Sharma obtained for t-tests, one-way analyses of variance, and correlations do *not* generalize to the analysis of variance of factorial designs (see Davison & Sharma, 1990).

5. Coombs' (1950) unfolding technique also locates both stimuli and persons simultaneously on the attribute being scaled. Because this technique has received only quite limited attention within the attitudes domain, it is omitted from this chapter (see Coombs, 1950, 1964; Dawes, 1972; Dawes & Smith, 1985; McIver & Carmines, 1981).

6. It is unclear whether the instruction to sort the items into equally spaced intervals is an integral part of the method. Thurstone and Chave's (1929, p. 31) description of their instructions to their judges does not mention this instruction. The end intervals and the middle interval were labeled, but Thurstone and Chave thought any further descriptions would prevent the subjects from sorting the items into what appeared to the subjects to be equal shifts of opinion between successive piles (p. 30). However, in a subsequent paper by Thurstone (1930) that described a scale for measuring attitude toward the movies, the judges were instructed to treat the intervals as equal steps. A number of general descriptions of scaling techniques describe the method of equal-appearing intervals as including the instruction (e.g., B. F. Green, 1954; Torgerson, 1958), but others do not (e.g., A. L. Edwards, 1957b). Regardless of whether the instruction is included, it is clear that the method assumes that the interval widths are equal.

7. The method of successive intervals was apparently independently derived by Guilford (1938), who called it the "method of absolute scaling" and by Attneave (1949), who called it the "method of graded dichotomies."

8. Setting the standard deviations equal to some number is equivalent to choosing a unit of measurement. The unit of measurement is arbitrary for an interval scale.

9. Because the normal distribution extends from $-\infty$ to $+\infty$, proportions of 0.00 and 1.00 have indeterminant z values. Also, when a proportion is quite extreme (e.g., $\leq .02$ or $\geq .98$), its z-score value can vary considerably depending on the value in the third decimal place of the proportion. Reliable determination of the value in the third decimal place would require an impractically large number of judges. Therefore, B. F. Green (1954) recommended that z-score cell entries with values greater than ± 2 should also be eliminated.

10. The mathematical details for deriving the interval widths and scale values are quite simple. The z-score in any cell of Table 2.6 is the location in normal curve units of the upper boundary of that interval (column) in the distribution for that item (row). Thus, t_c, the upper boundary of interval c in the distribution of item i, has a z-score value of $z_{ci} = (t_c - s_i) / \sigma_i$. The difference between the upper boundaries in any two adjacent intervals, c and c', in the same distribution is then given by:

$$z_{ci} - z_{c'i} = [(t_c - s_i) / \sigma_i] - [(t_{c'} - s_i) / \sigma_i] \quad (2.1a)$$

Because the s_i values in the first and second terms on the right side of the equation cancel, $z_{ci} - z_{c'i} = (t_c - t_{c'}) / \sigma_i$. We see, then, that the difference between z-scores in adjacent columns of the same row is proportional to the width of the interval between their boundaries. The difference between the upper boundary of interval c in any two different distributions i and j is given by a similar expression:

$$z_{ci} - z_{cj} = [(t_c - s_i) / \sigma_i] - [(t_c - s_j) / \sigma_j] \quad (2.1b)$$

If we make the usual assumption that the standard deviations of all the item distributions are equal (i.e., $\sigma_i = \sigma_j$ for all i and j), then the σs in the above equations are equal to a constant which can be set equal to 1 with no loss of generality (see note 8). Equation 2.1a then simplifies to $z_{ci} - z_{c'i} = t_c - t_{c'}$ and Equation 2.1b simplifies to $z_{ci} - z_{cj} = -s_i - (-s_j) = s_j - s_i$. We see then that the difference between the z-scores in adjacent columns in the same row provides an estimate of the width of the interval between the two columns and that the difference between the z-scores in the same column but in different rows provides an estimate of the difference in the scale values of the items.

11. Under the assumption that the item distributions all have standard deviations equal to 1, the z-score in any cell ci is $z_{ci} = t_c - s_i$. When there are entries in each of the cells, the mean of any column c is $\bar{z}_c = t_c - \bar{s}$. By fixing the zero point of the scale at $\bar{s} = 0$, the mean of the z-scores in a column is just t_c. The mean of any row i is $\bar{z} = \bar{t} - s_i$, where $\bar{t}$ is the mean of the column means (grand mean). The scale value of the item in row i, s_i, then is just $\bar{t} - \bar{z}_i$.

12. Since Thurstone's early work, a number of procedures have been developed to reduce the labor involved in judging a large number of stimuli. For example, subsets of stimuli can be judged by subgroups of judges (see Torgerson, 1958, pp. 191-194). Nonetheless, these techniques have not been put to much use in attitude item scaling.

13. Guttman scaling is not limited to items that involve only two response categories, such as *agree* or *disagree*, but can handle multiple response categories that indicate the degree of agreement and disagreement (see Guttman, 1944, 1947a).

14. In theory no allowance is made for error, although Guttman's concept of quasi-scales and the acceptability of coefficients of reproducibility between .85 and .90 indicates some tolerance of random error in practice. *Latent structure analysis* (Lazarsfeld, 1950, 1954; Lazarsfeld & Henry, 1968), which incorporates aspects of Guttman scaling, more realistically allows for a probabilistic relationship between the latent, underlying attribute and the overt response (B. F. Green, 1954; Torgerson, 1958).

15. T. L. Kelley (1939) found that if the total scale scores are distributed normally, the selection of respondents from the upper and lower 27 percent of the distribution provides optimal discrimination. For flatter than normal distributions, the

percentage needed for optimal discrimination is larger (Cureton, 1957).

16. The item-total score correlation is calculated with the item score excluded from the total score.

17. A high negative correlation indicates that the item was scored incorrectly either because the investigator incorrectly judged its favorability or in some way mixed up the scoring of the alternatives. In either case, items that correlate negatively should have their scoring reversed.

18. It is common in the research literature to see various sorts of rating scales erroneously described as *Likert-type* scales, even scales that do not require respondents to indicate the extent of their agreement or disagreement with the content of an item about an attitude object. General rating scales should certainly not be called Likert scales. Furthermore, agree-disagree scales should not be labeled Likert scales because the term *scale* should be reserved for the *set of items* that has been chosen based on Likert scaling procedures (i.e., item analysis) and that can therefore be used as a scale to assess an attitude.

19. Factor analysis is a statistical technique that attempts to account for the intercorrelations among variables by a smaller number of underlying dimensions or factors. It examines the intercorrelations among items and attempts to find clusters of items that are highly correlated with one another but are not correlated highly with items in other clusters. These clusters are called "factors" or "dimensions."

    Confirmatory factor analysis tests hypotheses about which variables are related to or "load on" which underlying factors and how the factors themselves are related. Most confirmatory factor analysis programs incorporate a test of goodness-of-fit that assesses how well the observed relationships among the variables can be accounted for by the hypothesized model. To assess whether a set of items has a single factor structure, one would specify that all the items load on a single common factor. The fit of the single factor model could be contrasted with that provided by a multidimensional factor structure.

20. The notion of strictly parallel measures with same means and standard deviations should regarded as part of an idealized model that only be approximated in reality. Similarly, assumptions of independence of true scores errors in classic true score theory are also approximation to reality. Discussion of so alternative item response theories that do make these restrictive assumptions is beyond scope of this chapter (see Crocker & Algi 1986; Nunnally, 1978).

21. By definition, two measures are parallel if, each individual, they yield the same true sc and differ only in their errors, that is, $X = T + e$ and $X' = T + e'$. The correlation between paral measures, $\rho_{XX'}$, is given by the expression:

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}}$$

    where $\sigma_{XX'}$ is the covariance of measures X a X'. Because parallel measures have the sa means and standard deviations, the denomina of the equation above is σ^2_x. The numerator of t equation can be rewritten as $\sigma_{(T+e)(T+e')}$. When t expected value of the numerator is taken and t independence assumptions of classic true sc theory are invoked, the numerator becomes Then:

$$\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X}$$

    which is the proportion of the observed sc variance that is true score variance or the r liability of a measure.

22. An alternative to the theory of parallel measu ment is domain sampling theory, which assu that items are sampled randomly from the cont domain (see Nunnally, 1978). Both theories lead the same results and equations, given the assu tions of classic true score theory. We have cho to explicate reliability theory in terms of para measurement theory because the ideas of dom sampling, while fitting attitude measures based the psychometric model, seem less appropriate attitude measures based on psychophysical other models.

23. In classic true score theory, the reliability of a measure also can be shown to be equal to the square of the correlation between true and observed scores. Since a variable cannot correlate with the true score of another variable higher than it correlates with its own true score, the square root of the reliability coefficient of a measure is the upper limit for its validity coefficient.

24. Biases that may affect the proportion of people who favor an issue are of particular concern to survey researchers because they are interested in generalizing their results to some real population. Although this issue is less troublesome to laboratory researchers, whose main interests are in finding relative differences between experimental and control conditions, response effects can affect the precision of experimental outcomes and restrict generalizability of results. Moreover, even in the laboratory, an attitude measure is rarely administered by itself. Information on other variables (e.g., manipulation checks, measures of possible mediators and moderators) is often collected along with the attitude measure. Context effects can occur between different measuring instruments as well as within a particular instrument.