Likert Attitude Scaling

Measuring someone's attitude is an attempt to locate his position on an affective continuum ranging from "very positive" to "very negative" toward an attitudinal object. In the Likert scaling technique this quantification is accomplished by tallying respondents' affirmation of positive and negative belief statements about the attitudinal object.

**DELIMITING THE ATTITUDBINAL OBJECT**

In the discussion of the nature and definition of attitude in Chapter 1, frequent reference was made to the "psychological object" or "attitudinal object." Attitude, it was indicated, is the extent of liking or disliking something. That "something" is the attitudinal object. The first consideration in any attempt at attitude measurement is to identify this object. "Attitude toward what?" is the question that must be answered before proceeding farther. The more precisely the attitudinal object is delimited, the more successful will be the measurement effort. In order for attitude-scale scores to be meaningful, it is essential for the researcher who uses the scale scores to have precisely the same object in mind as do the respondents to the scale. Similarly, the scale-construction effort will not be successful if the attitudinal object is so nebulous or ill defined that different respondents have slightly different objects in mind.

In general, simple objects are more easily delimited than are complex objects, tangible objects are more easily delimited than are abstract objects, and individual objects are more easily delimited than are classes of objects. Proper nouns are among the easiest objects to work with: Martin Luther King, Honda Accord, Statue of Liberty. But attitude measurement would be of limited practical value if we were restricted to measuring attitude toward simple, tangible objects.

In measuring attitudes toward complex and/or abstract objects, it is vitally important to define the object thoroughly and completely. If you want to measure attitude toward education, for instance, you must be clear in your own mind, and in communication with others (scale respondents, consumers of your research findings, and so forth) which of the following components are included in your definition of education:

1. The general level of societal literacy (e.g., "A highly educated society makes a strong country.")
2. Education as a profession (e.g., "Education is a fulfilling profession.")
3. The public school system (e.g., "Formal education in this country is the best in the world.")
4. One's own educational level (e.g., "More education will make me a better person.")

If you included all these items in a single scale and summed item scores for each respondent, the total scores would have little or no clear meaning. The solution is either to construct multiple scales, one for each attitudinal object, or to define education narrowly and specifically enough so that some of these multiple dimensions are precluded. Rather than bothering scale respondents with a definition, such delimitation of the attitudinal object is generally accomplished through careful selection of scale items.

**GENERATING AN ITEM POOL**

Once the attitudinal object has been determined, a pool of items stating beliefs or opinions about that object is constructed or collected. It is important, in generating this item pool, to tap a broad diversity of opinions about the attitudinal object. Since completed Likert scales typically contain only about 20 of the dozens or even hundreds and thousands of possible beliefs about a particular object, they will be most valid if they constitute a fairly comprehensive and representative sampling of this almost infinite theoretical domain. One way to achieve this comprehensiveness in generating potential scale items is to ask a diverse group of people (those with substantial knowledge about, and with both positive and negative attitudes toward, the attitudinal object) to write down several of their own beliefs and feelings about the attitudinal object. With a little editing, belief statements generated in this manner can become the item pool that serves as the basis for your attitude scale. Interviews and other conversations about the attitudinal object are similarly productive in generating useful pool items, as are editorial-type writings about the object.

While in a broad sense all attitudinal items can be called belief or opinion statements, attitude measurers sometimes distinguish among belief or cognitive items, feeling or affective items, and behavioral-tendency or conative items. Cognitive items express beliefs about the attitudinal object.
Affective items constitute a very direct expression of feeling toward the attitudinal object. Conative items express behavioral intention or behavioral preference with regard to the object. There are actually two kinds of behavioral-tendency items: would items and should items. Would items express a personal behavioral intention toward the attitudinal object. Should items express a behavioral preference for social action. Following are examples of these various item types, with diesel cars the attitudinal object:

Belief: Diesel cars are economical to run.
Feeling: I like diesel cars.
Would: I would buy a diesel car if I had my choice.
Should: The government should give a tax credit to people who buy diesel cars.

Since beliefs about, feelings toward, and behavioral tendencies with regard to objects tend to be highly related (people who like X tend to have positive beliefs about X and to express tendencies or desires to behave positively toward X), any or all of these types of items can be used in attitude scales, without distinction.

It is important, however, to distinguish between behavioral-tendency items and actual-behavior items. Actual-behavior items can be somewhat of a problem in attitude scales. Someone living in Alaska might think diesel cars are “the greatest” but might have to respond “no” or “disagree” to the statement “I own a diesel car,” because diesel cars are not generally available in Alaska. If you think about it, you’ll realize that actual behaviors are influenced by many things besides attitudes and therefore are not always accurate indices of attitude. As a general rule, items inquiring about respondents’ actual behaviors are best left out of attitude scales. A more thorough discussion of the relation of attitude to behavior is found in Chapter 7.

Each Likert item should be clearly positive or negative with regard to the attitudinal object. Neutral items won’t work in Likert scales. Positive items are items that state favorable beliefs about or feelings toward the attitudinal object. Following are examples of positive, neutral, and negative items:

Positive: I would prefer having a diesel-powered car. Diesel cars use less fuel.
Neutral: Diesel cars are a kind of automobile. Diesel cars are produced in the U.S.
Negative: Diesel cars don’t start easily in cold weather. Diesel cars cost more than they are worth.

Some items may connote a positive attitude for some respondents but a negative attitude for others. For example, the item “I like diesel cars about as well as I like tutti-frutti ice cream” is as positive as or negative as tutti-frutti ice cream is to each respondent. Care should be taken to select only items for which there is a very high degree of consensus regarding the direction (positive or negative) of the affective component. An objective method for making this determination is described later in this chapter, in the section entitled “Item Analysis.”

Another type of item that won’t benefit your scale is an item to which all respondents make the same response. When your scale is complete, it will be used to discriminate among respondents, that is, to distinguish people with very positive attitudes from those with moderately positive attitudes, and these, in turn, from those with negative attitudes, and so forth. Every item must contribute a little to this discrimination. In order to discriminate, items must “spread people out” in terms of score points. If all respondents make the same response to an item, the item isn’t spreading them out. In the extreme case, if all respondents make the same response to every item in the scale, they will all obtain the same total score. There will be no discrimination whatsoever, and the scale will be useless.

Some factual items don’t spread respondents out. If a fact is so well established that all respondents “agree,” it won’t make a good attitudinal item. For example, probably all or nearly all respondents would “strongly agree” that “diesel cars burn diesel fuel.” Some factual items, however, can discriminate just fine. It is objectively established that “diesel cars get better mileage per gallon than do gasoline cars,” but not all respondents know and/or believe this fact. Those who know this positive “fact” will agree and thus have legitimate points added to their attitude score. Those who have misinformation will disagree and thus not gain attitudinal points. Those who don’t know may answer “uncertain” or may guess based upon their overall attitude toward diesel cars. A factual item such as this can spread respondents out and thus will contribute to the discrimination of the scale.

Strongly or extremely worded items can also suffer from the problem of not spreading respondents out. The item “I would never purchase a diesel car” won’t spread people out as well as this less extremely worded rewrite: “I am unlikely to purchase a diesel car.” In general, the use of absolutes such as always and never should be avoided in writing Likert scale items. Of course, you’ll know better how well your items are spreading people out after you administer them, but even before administration you can be on the lookout for items that state well-established facts and items that are very strongly worded.

It should be clear, by now, that good item writing is in large part simply a
matter of good writing. The wording of each item must convey the same meaning to the respondent that it conveys to the item writer. Furthermore, it must convey the same meaning to all respondents. Items should be worded as briefly, as clearly, and as concisely as possible.

A particular problem for respondents in taking attitude scales is compound items (Likert called these “double-barreled” items). Compound items really contain two beliefs or opinions in one sentence, for example, “Diezel cars are noisy and smelly.” Respondents may agree with one part of the statement (one belief) and disagree with the other part. Two potentially good items on a scale measuring attitude toward diesel cars could be constructed from this double-barreled statement.

About half positive and half negative items is the proportion usually recommended in order to preclude the possible effect of “acquiescence response set.” (A thorough discussion of response sets appears in Chapter 7.)

A final consideration in constructing the item pool is how many items to include. The answer depends partly on how well your items are written, partly on how specific your attitudinal object is, and partly on how reliable you need your final scale to be. If you are a novice at writing and editing items for attitude scales, you may have to start out with a pool of more than 50 items. If you have been systematically editing items and eliminating poorly worded and redundant items during the item-pool construction process, and if you are a good writer, you may need only 25 or 30 items in your initial pool. Attitudes toward specific and “tightly” conceptualized objects can be measured accurately with fewer items than attitudes toward loosely defined and amorphous objects. The question of reliability will be discussed at length in Chapter 7. Suffice it to say here that, all else being equal, the magnitude of the reliability coefficient is directly related to scale length.

RESPONSE FORMAT AND SCORING

Likert items use response categories ranging from “strongly agree” through “strongly disagree.” Five categories are fairly standard (“strongly agree,” “agree,” “uncertain” or “undecided,” “disagree,” and “strongly disagree”). Some scale constructors use seven categories, and some prefer four or six response categories (without middle category). All of these options seem to work satisfactorily. It should be noted in this regard that reducing the number of response categories reduces the spreading out of scores (reduces variance) and thus tends to reduce reliability. Increasing the number of response categories adds variance. As the number of response categories is increased, a point is reached at which respondents can no longer reliably distinguish psychologically between adjacent categories (e.g., “Do you agree very strongly, or very, very strongly?”). Increasing the number of categories beyond this point simply adds random (error) variance to the score distribution.

In scoring positively stated Likert items “strongly agree” receives 5 points, “agree” 4 points, and so on. For negatively worded items the scoring is reversed (“strongly disagree” equals 1, “disagree” equals 2, and so on). Thus, responses indicating a positive attitude toward the attitudinal object (agree responses to positive items; disagree responses to negative items) result in high scale scores. Responses indicating a negative attitude toward the attitude object result in low scale scores. In calculating the total scale score for each respondent, item scores are summed. (Thus the Likert scaling procedure is commonly known as the method of “summed” ratings.)

The highest possible scale score is $5 \times N$ (number of items). This would certainly be interpreted as a strongly positive attitude. The lowest possible score, indicating a strongly negative attitude, is $1 \times N$. A neutral attitude would result in a score of approximately $3 \times N$. More precise score interpretation should be normative (i.e., relative to the scores of other respondents) rather than “absolute.” It is not necessarily true, for instance, that scores above $3 \times N$ are “high” scores. With reference to some attitudinal objects (e.g., democracy, motherhood, apple pie) most or all respondents may score higher than $3 \times N$.

ITEM ANALYSIS

In order to refine the item pool into a finished scale, item analysis is necessary. Following administration of the item pool to a group of respondents, three sets of statistics are typically computed for each item: (1) percentage of respondents making each response, (2) item mean and standard deviation, and (3) item discrimination index. In addition, item analysis usually entails computing a reliability coefficient for the total scale. No fixed number of respondents is required for item analysis, but item and scale statistics will be much more stable when computed on 100 respondents than on 10.

Percentage of respondents making each response to each item and item mean and standard deviation tell about item response distribution, spread, and skew. In general, items on which respondents are spread out across response categories are better than items on which respondents are clustered primarily in two or three response categories. (The problem with well-established factual items, you will remember, is that respondents are not spread out.)
Item discrimination index shows the extent to which each item discriminates among respondents in the same manner as the total scale score. If high scorers on an item have high scale scores and low scorers on the item have low scale scores, the item is discriminating among respondents in the same manner as the total score and thus will have a high discrimination index. In a scale measuring a psychological construct, such as attitude, each item must contribute to the measurement of this construct. Items that don't discriminate among respondents in the same manner as the total score aren't measuring the same thing as the other items. Such items are typically rejected from the scale. When the discrimination index is used as a criterion for rejecting items, the result will be a more homogeneous scale—all remaining items will be contributing to the measurement of the same underlying construct.

If a computer or a somewhat sophisticated calculator is available, correlating item scores with scale scores is the most efficient method of computing item discrimination index. Correlating item scores with total test scores has the same meaning as does intercorrelating any two variables. If the scores on the two variables "go together"—if high scorers on one variable (e.g., test item) tend to be high scorers on the other variable (e.g., total scale) and if low scorers on the one variable tend to be low scorers on the other variable—the variables are positively correlated. Just as in other applications of the correlation statistic, item discrimination indices computed by this method can range from 1.00 to -1.00. Table 2.1 shows the performance of eight students on four test items and on the total test. Try to estimate the correlation of each item with total scale scores.

Item 11 has a very high, positive correlation with total scale score. This item discriminates among respondents in the same manner that total scale score discriminates among respondents. Item 16 has a moderate, positive correlation with total score. Item 17 correlates about zero, and item 27 has a moderate to high, negative correlation.

A zero or near zero correlation indicates that the item is discriminating among respondents in a manner unlike the discrimination of the total score (and thus of most of the other scale items). Such an item is not measuring the same thing that the other items are measuring and is not contributing to the measurement being accomplished in the other items. It is usually wise to eliminate such items from the scale. A negatively correlating item is actually working against the discrimination being accomplished in the rest of the scale. A moderate or high negative correlation sometimes signals a miskeyed item (i.e., scored in reverse). If the item has not been miskeyed, it should definitely be removed from the scale.

A more crude, albeit acceptable, method for calculating item discrimination indices is to compare, for each item, the responses made by the high-scoring and low-scoring respondents. Two groups of test takers are selected on the basis of total scale score, a high-scoring group and an equal-sized low-scoring group. For each item the mean is calculated for each group, and then the mean of the low-scoring group (\(X_n\)) is subtracted from the mean of the high-scoring group (\(X_h\)). Table 2.2 illustrates this procedure for nine items from an attitude scale that was administered to 30 respondents (the middle-scoring 10 respondents are not used in these calculations).

Since the highest possible item score is 5 and the lowest possible item score is 1, the possible range for the mean difference discrimination index statistic is from 4.00 to -4.00. A 4.00 can occur only if all high-scoring group members make the most extreme positive item response and all low-scoring group members make the most extreme negative response (as for item 8 in Table 2.2). A -4.00 can occur only with the opposite response pattern. Both are highly unlikely conditions. Items with positive mean differences (i.e., items 1, 2, 3, 5, 6, and 8) are discriminating in a manner similar to the total score. Items with mean differences around zero (items 4 and 9) are not discriminating between high- and low-scoring respondents and are thus not contributing to the measurement of attitude. Items with negative discrimination indices (item 7) are working in opposition to the discrimination being accomplished by the rest of the scale.

There are several possible causes for items failing to discriminate in the same manner as the total scale score. They may be ambiguous in wording, they may be keyed incorrectly (reversing the keying will solve this problem), respondents may not agree about their positiveness/negativeness (reversing the keying for some respondents would solve this problem, but...
that of course is not practical), or they may not connote affect about the attitudinal object at all. Whatever the reason, items with zero and negative discrimination indices should generally be removed from the scale. (Miskey items can be kept with rekeying.) The number of items you throw out will depend largely on the size of your initial pool. If you started with a pool of 40 items you can—and probably should—throw out about half. Even many items with good item statistics will be unnecessary. If you began with but a few items to spare in your original pool, it may be necessary to rewrite some items in order to retain a long enough scale for adequate reliability. In reducing scale length on the basis of item statistics, care should be taken to maintain both a diversity of opinion content and a balance of positively and negatively worded items.

A COMPLETED LIKERT SCALE

Figure 2.1 is a standard Likert-type scale designed to measure attitude toward marijuana. Nine items are positive, 11 negative. When the scale was administered in several graduate classes in social psychology, alpha internal consistency reliability coefficients ranged from .83 to .88. (This statistic will be explained in detail in Chapter 7.)

**SUMMARY OF LIKERT SCALING PROCEDURES**

1. Identify the attitudinal object; delimit it quite specifically.
2. Collect a pool of opinion items (30 or more) about the attitudinal object. All items must state or imply something positive or negative about the attitudinal object. Neutral items cannot be used.
3. Administer the item pool to a group of respondents. Each respondent indicates his degree of agreement to each item.
4. Score each item for each respondent. On a five-point continuum, "strongly agree" receives five points for positive items, "agree" four points, and so forth. For negative items this scoring procedure is reversed.
5. Sum each respondent's item scores. The highest possible score (indicating most positive attitude) is five times the number of items. The lowest possible score is one times the number of items.
6. Correlate total scale scores for all respondents with item scores for all respondents (one item at a time).
7. Eliminate negatively correlating and zero-correlating items. Keep enough positively correlating items to maintain desired level of reliability. Maintain a balance of positive and negative items.
FIGURE 2.1
Likert attitude scale measuring attitude toward marijuana.

Opinions About Marijuana

Indicate on the line to the left of each statement how much you agree or disagree with it. Please mark every item.
Use the following response categories:

A = strongly agree
B = agree
C = uncertain
D = disagree
E = strongly disagree

1. No right-thinking person would use marijuana. (N)
2. Marijuana use leads to heroin use. (N)
3. Only hippies and weirdos use marijuana. (N)
4. Marijuana should be legalized. (P)
5. Marijuana use causes birth defects. (N)
6. Since there is no hangover, marijuana is a good substitute for alcohol. (P)
7. Marijuana is a narcotic drug. (N)
8. Most people who criticize marijuana use don’t know anything about the drug. (P)
9. Habitual marijuana users are neurotic. (N)
10. As a symbol of the youth culture, epitomizing disobedience and disregard for authority, marijuana usage should be put down. (N)
11. In our highly impersonal society, marijuana helps one express feelings and relate to others, and should therefore definitely be used by those who feel the need. (P)
12. Marijuana is a good social stimulator and should be allowed, especially at parties, where mixing is important. (P)
13. Marijuana is not a "hard" drug. (P)
14. If a son or daughter uses marijuana, Mom and Dad should be willing to try it before they condemn it. (P)
15. Since we aren’t sure if it can harm us, we should avoid marijuana. (N)
16. Marijuana use is illegal and therefore wrong. (N)
17. Marijuana has psychological therapy potential. (P)
18. Marijuana causes dehumanization. (N)
19. Criminals have a higher rate of marijuana use than does the general public. (N)
20. Intelligence test scores of marijuana users are higher on the average than scores of nonusers. (P)

Note: P = positively keyed item; N = negatively keyed item.
The Semantic Differential

As was explained in Chapter 1, the semantic differential, developed by Charles Osgood and his associates, was not originally designed for the purpose of attitude measurement. Osgood was studying the nature of meaning. He believed that the thousands of adjectives that we use to describe the world around us have considerable connotative overlap. Using a statistical technique called factor analysis, he attempted to identify the underlying dimensions of meaning. He found that a large proportion of all meaning can be accounted for with three cognitive dimensions: evaluation, potency, and activity.

The semantic differential is Osgood's instrument for measuring the extent to which respondents attribute each of the several meaning dimensions to particular objects. Pairs of opposite adjectives that are highly representative of the dimension(s) to be measured serve as "items." Respondents indicate the extent to which each adjective (or its paired opposite) describes the object.

Figure 6.1 is an example of a semantic differential designed to measure the three major dimensions of meaning found by Osgood. This figure illustrates the commonly used semantic differential administrative format. (The meaning dimension represented by each adjective pair is indicated for instructional purposes; these are not so identified for respondents.) Additional dimensions of meaning can be measured by including adjective pairs representative of those dimensions in the instrument. Adjective pairs are normally separated by seven response categories representing equal units along the adjective-opposites continua. A common practice is to randomly alter the "direction" of the adjective continua (e.g., in Figure 6.1 valuable and clean, both connoting positive evaluation, appear on the left ends of their respective continua; good and fair, also connoting positive evaluation, appear on the right ends of their continua). This convention is designed to overcome response sets that could adversely affect the validity of the instrument. Response sets are discussed in greater detail in Chapter 7.

FIGURE 6.1
Sample semantic differential designed to measure three dimensions of meaning.

<table>
<thead>
<tr>
<th>My Spouse</th>
<th>worthwhile</th>
<th>worthless</th>
</tr>
</thead>
<tbody>
<tr>
<td>valuable</td>
<td>clean</td>
<td>dirty</td>
</tr>
<tr>
<td>bad</td>
<td>small</td>
<td></td>
</tr>
<tr>
<td>unfair</td>
<td>weak</td>
<td></td>
</tr>
<tr>
<td>large</td>
<td>deep</td>
<td>shallow</td>
</tr>
<tr>
<td>fast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>active</td>
<td>passsive</td>
<td>cold</td>
</tr>
</tbody>
</table>

Note: The first four adjective pairs measure the evaluation dimension; the next three measure potency; and the last three measure activity.

USE IN ATTITUDE MEASUREMENT

For purposes of attitude measurement a special form of the semantic differential, consisting entirely of adjective pairs representing the evaluation dimension, is constructed. Figure 6.2 lists adjective pairs that Osgood found to "load" highly on the evaluation dimension in his factor analyses.

Usually a few adjective pairs are sufficient to produce a scale with high internal consistency. Selected adjective pairs will not work equally well, though, across attitudinal objects. The adjective pair high-low, for example, normally contains a large proportion of evaluative meaning. But when this adjective pair is used with reference to the attitudinal object "marijuana," it may not connote evaluation. Likewise, a Volkswagen "Beetle" owner and devotee may rate VW Beetle toward the ugly end of a beautiful-ugly continuum, with no negative evaluation intended.

The safe course is to select several (perhaps six or seven) adjective pairs
initially, perform an item analysis1 (just as in Likert scaling), and eliminate the weaker items. Figure 6.3 is an example of a semantic differential consisting of nine potentially evaluative adjective pairs designed to measure attitudes toward marijuana. Administered to a group of 24 college students, the scale had an internal consistency (alpha) coefficient of .86. When the four adjective pairs with the lowest correlations with total scale score (ugly-beautiful, clean-dirty, useless-valuable, and high-low) were eliminated, leaving a five-item scale, the alpha increased to .95. (The high-low adjective pair, incidentally, correlated -.28 with the total score from the nine-item scale.)

In scoring the semantic differential, each item (adjective pair) can contribute from one to seven points to the total score, the most positive response receiving a 7 and the most negative a 1. Thus, a semantic differential consisting of five pairs of evaluative adjectives would have a score range from 5 to 35.

STRENGTHS AND WEAKNESSES OF THE SEMANTIC DIFFERENTIAL

The semantic differential has much to recommend it as a self-report attitude-measurement technique. It is relatively easy to construct, it is

---

1 See the descriptions of the SPSS RELIABILITY and ESTAT TESTAT computer programs in Appendix B.

short and thus quick to administer, and it is usually highly reliable. (Test-retest reliabilities and internal-consistency coefficients around .90 are common for well-constructed semantic differential scales used to measure the evaluative dimension.) Furthermore, semantic differential attitude-scale scores typically correlate very highly with scores from Likert and Thurstone attitude scales. The semantic differential does have several drawbacks, though. Administration rapport is sometimes a problem. Respondents who insist on a literal interpretation will sometimes balk at applying some adjective pairs to some attitudinal objects ("What do you mean, how dirty or clean is my mother? I'm not going to mark anything on that item"). As you can see, it's critical to impress on respondents the importance, for meaningful measurement, of answering all items—even if they experience a small amount of psychological discomfort in the process. It might help to tell them that some items will be discarded after statistical analysis.

A second problem is in the realm of validity. Respondents can slant their
attitude-scale scores at will. Semantic differential attitude scales are very blatant or "transparent" in their purpose. It is relatively easy for respondents to figure out what is being measured. This is true, to some extent, with Likert, Thurstone, and Guttman scales, but of the four, the semantic differential comes closest to simply asking respondents, "What is your attitude toward object x?" or "How positively (negatively) do you feel about object x?" Methods of combating such (conscious or unconscious) biasing of responses are discussed in Chapter 7.

Another possible problem in using the semantic differential for attitude measurement is failing to isolate the evaluative dimension. It should be clear from the above discussion that the measurement of meaning and the measurement of attitude are not synonymous. Attitude, or evaluation, is only one dimension of meaning. But sometimes researchers or evaluators in a hurry to select an attitude-measurement instrument find and use an instrument such as is shown in Figure 6.1 when they should be using one such as in Figure 6.3. When items measuring two or more dimensions of meaning are combined in a single scale score, the meaning of that score is unclear. It should most certainly not be characterized as a measure of attitude.

Even worse, some researchers bring together a conglomeration of adjective-opposite pairs in a semantic differential format, with no particular psychological construct in mind at all, and call it attitude measurement. Their mistake is to associate the attitude label with the semantic differential instrumentation format. If the items are not working together to measure a particular psychological variable, not only is this not attitude measurement, it is not the measurement of any psychological trait.

Reliability and validity are the benchmark criteria for assessing the quality of all measurement devices and procedures. If a measurement instrument is valid, it is measuring the right thing—what it is supposed to be measuring. If it is reliable, its measurement is consistent and accurate, rather than random. In the study of social and psychological phenomena, the elements of measurement are abstract constructs (concepts) such as attitudes, values, and personality traits. Unlike the measurement of physical traits or conditions such as height, weight, and temperature, psychological traits cannot be seen or felt. Nor can they be measured directly. They must be inferred from people's beliefs and behaviors. This measurement process is extremely prone to error. For this reason psychometric and sociometric researchers must take particular care to maximize the quality (the reliability and validity) of their measurement instruments and procedures.

This chapter will discuss the commonly used methods for assessing the reliability and validity of psychological measurement instruments. Since many of these methods use the statistical technique of correlation, it will be useful to review the important characteristics of the correlation statistic before proceeding farther.

Correlation was introduced in Chapter 2 as a statistical device for studying the relationship of each item in an attitude scale to scores on the total scale (item discrimination). You will remember that a correlation coefficient of 1.00 indicates a perfect, positive relationship between two variables, and a zero correlation (0.00) indicates no relationship at all. Thus a reliability (correlation) coefficient of 1.00 indicates perfect reliability, whereas a reliability coefficient of .00 indicates no reliability whatsoever (essentially random scoring). Likewise, validity coefficients of 1.00 and .00 indicate, respectively, total or perfect validity and no validity.

Never do we see either reliability or validity coefficients of 1.00, how-