3

Explicit Measures of Attitudes

That guy Arnold sure is hot!

Auto dealers are just out to make a quick buck, and they'll rip off their customers every time they get a chance.

School vouchers are a bad idea because they will take money away from public schools.

These are all expressions of attitudes. They describe a person's feelings toward another person, a group, a situation, or an idea. Attitudes can be expressed in many ways—with different words, different tonal inflections, and different degrees of intensity. Some of the color and richness of the ways in which attitudes and opinions are often expressed is captured in the quotations from actual public opinion interviews shown in Box 3–1.

How can statements like these be studied scientifically? To compare them in any systematic way, we have to classify them into two or more categories (e.g., pro or anti concerning some group or idea) or, preferably, measure them on a quantitative scale (e.g., indicating *degree* of favorability or unfavorability). Furthermore, the classification or measurement must be **reliable**, that is, consistent. Reliability means (a) that two different raters agree on their classification of the statements to a high degree, and also (b) that on two different occasions the respondents' statements are largely consistent. Reliability and validity of measurement are discussed later in this chapter.

In this chapter, we examine ways of measuring **explicit attitudes**—evaluations that a person is consciously aware of and can express. In the next chapter, we examine **implicit attitudes**—evaluations that are automatic and function without a person's awareness or ability to control them (Greenwald & Banaji, 1995; Dovidio, Kawakami, & Gaertner, 2002).

TYPES OF ATTITUDE QUESTIONS

All measures of explicit attitudes rely on self-report. There are two basic types of questions that are used to obtain statements of attitudes and opinions. Some of the interview questions quoted in Box 3–1 are **open-ended** questions—ones that give the respondent a free choice of how to answer and what to mention (e.g., "What do you think was the main cause of these disturbances?"). Other questions are **closed-ended**—that is, ones that present two or more alternative answers for the respondent to choose between (e.g., "Have the disturbances helped or hurt the cause of Negro rights?"). Often an interview will use both types of questions because they have complementary advantages and disadvantages.

Box 3-1 on page 45 has been removed. The text describing open-ended questions continues below on page 46.

Open-ended questions have the advantages of eliciting the full range, depth, and complexity of the respondent's own views, with minimal distortion, in his or her own words. They reduce the likelihood of overlooking important possible viewpoints which the investigator has not thought of or not included in the questionnaire. For these reasons they are often used as introductory questions to open up a topic that will subsequently be probed more deeply and intensively with closed-ended questions. (This is called a **funnel sequence** of questioning.) The primary disadvantages of open-ended questions are the difficulty and frequently the unreliability of scoring or **coding** them. That is, trying to decide how the response should be classified or what quantitative point on a scale it best represents can be difficult and time-consuming, and sometimes it cannot be done with adequate agreement between raters. For instance, how would you score the second respondent's answer in Box 3–1 that the disturbances "have helped.... They haven't helped yet but overall it will help..."?

For these reasons closed-ended questions are likely to make up a large majority of the items on most interviews and questionnaires. They have the advantages of being easy to score and relatively **objective**. That is, independent observers or scorers can reach a high percentage of agreement on which response was given or on what score should be assigned to the response. Of course, unlike open-ended questions, they have the possible disadvantage that they may force the respondent to use the concepts, terms, and alternative answers preferred by the investigator, rather than expressing his or her own ideas and preferences (Schuman & Presser, 1981b; Sudman & Bradburn, 1982; J. Converse, 1984; Schwarz, Groves, & Schuman, 1998; Robinson, Shaver, & Wrightsman, 1999).

Closed-ended questions must be written very carefully so as not to produce biased answers. Without such care in item construction, the results will be far less reliable, and sometimes they may be so slanted that they are seriously misleading. For instance, here are two biased items that were on questionnaires sent out by two political lobbying groups:

> Are you in favor of allowing construction union czars the power to shut down an entire construction site because of a dispute with a single contractor, thus forcing even more workers to knuckle under to union agencies? YES__ NO__ (Sudman & Bradburn, 1982, p. 2)

> The Soviets and other Communist countries have a record of breaking one treaty after another. They not only shoot down unarmed passenger planes but also lie about it afterwards. Do you agree with those Congressmen who want a so-called political solution based on signing agreements with Communist forces in Central America? YES__ NO__

Obviously, these questions are worded so as to encourage a "No" answer. Consequently, the response percentages reported by their sponsoring agencies will markedly exaggerate respondents' real attitudes about these issues. The lesson of these examples is that one should always look at the question's wording before making or accepting an interpretation of the meaning of survey response figures. For that reason, good practice requires that the exact question wording be stated whenever the quantitative results of survey questions are reported. For readers interested in constructing questionnaires and surveys, a multi-volume series by Fink (2003) is an excellent resource.

The most common way of measuring attitudes is to combine several items on the same topic to form a **scale** (e.g., a scale of political liberalism versus conservatism), and to compute a single score for each respondent for the group of items. In the following section we describe the major ways of constructing such attitude scales.

ATTITUDE-SCALING METHODS

During the late 1920s and early 1930s a number of attitude-scaling methods were developed that are still in common use today, and more recently a few additional methods have been developed. Each of the major attitude-scaling techniques are discussed here rather briefly, primarily to clarify their major characteristics and points of difference. This will not prepare you to use these methods yourself to build an attitude scale, but it will provide you with enough information to understand references to such methods later in this book or in the research literature.

In 1925 Bogardus was one of the first to use quantitative measurement methods in the field of social psychology. Thus, surprisingly, the quantitative study of attitudes is only about 80 years old, even though quantitative research in psychology goes back over 125 years to the founding of Wundt's laboratory in 1879, though the term attitude has been used in the psychological sense for well over a century, and though cognition, affect, and conation have been discussed by philosophers ever since the time of Plato. Given the relatively short history of quantitative research on attitudes and opinions, it is no wonder that many questions remain to be answered.

Bogardus' Social Distance Scale

Bogardus (1925) proposed a scale of **social distance** that could be used to determine attitudes toward various racial or nationality groups, many of which at that time were relatively recent immigrants to the United States. Respondents gave their judgments, following these instructions (p. 301):

> 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.
>
> 1. To close kinship by marriage
> 2. To my club as personal chums
> 3. To my street as neighbors
> 4. To employment in my occupation in my country
> 5. To citizenship in my country
> 6. As visitors only to my country
> 7. Would exclude from my country

By use of this scale, people's attitudes toward the English, Germans, Turks, and many other groups could be compared.

As can be seen in the example, the scale points progress systematically from acceptance of members of the racial or national group into the most intimate family relationships, down to complete exclusion of the group. The respondent's attitude score toward that group is taken as the closest degree of relationship that he or she is willing to accept. Some early findings showed that, to the average American, the English were the most accepted national group and Turks were one of the least accepted groups (Bogardus, 1928). A recent study compared the distance scores of 135 U.S. schoolteachers with those reported by Bogardus (Kleg & Yamamoto, 1998). In it, social distance scores were obtained for 39 ethnic and racial groups, and the scores of the teachers were strikingly similar to those found by Bogardus in 1928—the rankings of the

39 groups were highly correlated ($r = 0.86$). However, the distance ratings in the 1998 study were more homogenous, indicating less extreme positive and negative attitudes. As with the findings reported by Bogardus, the English were rated highest and the Turks lowest.

Other variations of this technique have allowed measurement of attitudes toward any social group, not just ethnic or nationality groups, and have also broadened the range of response options (e.g., Triandis, 1964, 1971).

Thurstone's Method of Equal-Appearing Intervals

A few years after Bogardus' work on social distance, Thurstone (1928) proposed the next attitude-scaling method. In contrast to the Bogardus scale, in which the scale points were not designed to be equidistant, Thurstone attempted to develop a method that would indicate rather precisely the *amount* of difference between two respondents' attitudes. The method that he developed is rather complex.

First, the investigator collects or constructs a large number (100 or so) of opinion statements representing favorable, neutral, and unfavorable views about the topic of interest (for instance, Thurstone studied attitudes toward the church, "Negroes," capital punishment, and birth control). Then the investigator obtains a large group of people to serve as judges and rate each statement's favorability or unfavorability toward the topic. Each judge sorts the statements into 11 equally spaced categories, disregarding his or her own attitude toward the topic, and considering only how *favorable* or *unfavorable* the statement is toward the attitude object. If there are statements on which different judges show substantial disagreement, they are discarded as ambiguous; other items may be discarded as irrelevant to the topic; and judges who make too few differentiations are omitted from later computations. The remaining statements are assigned **scale values** based on the median favorability rating of the judges. From these statements, a final scale of about 20 items (or sometimes more) is selected according to two criteria. The aim is to choose items having (a) scale values at approximately **equal intervals** along a 9-point or an 11-point scale of favorability, and (b) high agreement among the judges' ratings (that is, low spread or variability of their ratings).

After the items for the final scale have been chosen, they are randomly arranged on the questionnaire form without any indication of their scale values. Respondents check only the items they agree with and leave the others blank. A person's attitude toward the topic can then be defined as the mean (or the median—both methods have been used) of the scale values of the items that he or she has checked. An example of a Thurstone scale is shown in Box 3–2.

Thurstone's method makes the important assumption that the opinions of the judges do not affect the scale values of the items obtained from their judgments. This assumption has been shown to be reasonably correct when the judges do not have extreme views on the topic. However, if many of the judges have extreme views or are highly involved in the topic, the obtained scale values of the items will be affected (Hovland & Sherif, 1952). Specifically, judges who are highly favorable to a topic rate only a few of the most extreme statements as favorable, and they displace their ratings of most of the statements toward the unfavorable end of the judgment scale. The opposite is true for judges who are highly unfavorable toward a topic.

The other major drawback of Thurstone's method is that it is time-consuming and tedious to apply (Webb, 1955). For that reason it is used much less extensively than the method described next.

Box 3–2 A Thurstone Scale of Attitudes Toward Using Contraceptives

A selection of about half of the items from a Thurstone scale is shown below. Although the items are arranged here in the order of their scale values, on the actual questionnaire they would be arranged in a mixed-up order as indicated by their item numbers, and the scale values would not be shown. Respondents are to check or circle the numbers of the items with which they agree.

Scale value	*Item no.*	*Item*
1.28	*5*	*I detest the very word birth control.*
2.23	*13*	*I am afraid to use birth control.*
3.00	*3*	*My feelings would be hurt if someone advised me to practice birth control.*
4.17	*6*	*I am sorry for those who practice birth control.*
5.06	*11*	*It frightens me to think that the overcrowding is going to force birth control on us whether we want it or not.*
7.38	*10*	*It saddens me that so many persons are ignorant of the advantages of birth control.*
8.37	*9*	*I am happy about the positive effects of birth control.*
9.37	*12*	*I am so glad people are beginning to accept birth control.*
10.77	*1*	*It is a wonderful feeling to take advantage of birth control.*

Source: Kothandapani, V. (1971a). *A psychological approach to the prediction of contraceptive behavior* (pp. 26, 69–70). Chapel Hill: University of North Carolina, Carolina Population Center.

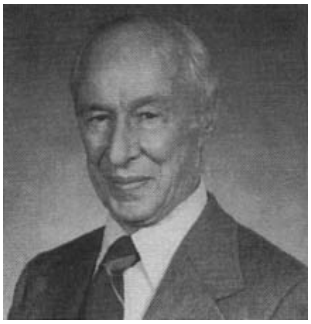
Likert's Method of Summated Ratings

Shortly after Thurstone's work, Likert (1932) proposed a simpler method of attitude-scale construction, which does not require the use of judges to rate the items' favorability. Better still, the reliability of Likert scales has been shown to be at least as high as that of the more difficult-to-construct Thurstone scales (Poppleton & Pilkington, 1964).

Likert's method was the first approach that measured the *extent* or *intensity* of the respondent's agreement with each item, rather than simply obtaining a "yes–no" response. In this method, again, a large number of opinion statements on a given topic are collected, but each one is phrased in such a way that it can be answered on a 5-point rating scale. For instance, here is an example from Likert's original scale of internationalism (Likert, 1932, p. 17)—it is interesting to note how many of these attitude items still have an up-to-date ring:

We should be willing to fight for our country whether it is in the right or in the wrong.
_____ Strongly approve
_____ Approve
_____ Undecided
_____ Disapprove
_____ Strongly disapprove

Respondents check one of the five choices, which are scored 1, 2, 3, 4, and 5 respectively. (Of course, items on the opposite end of the continuum—ones expressing a favorable attitude toward internationalism—would be scored in reverse: 5, 4, 3, 2, and 1, respectively.)



Photograph courtesy of Rensis Likert. Reprinted by permission.

Box 3–3 RENSIS LIKERT, *Attitude Measurement Pioneer*

Rensis Likert's distinguished career included pace-setting work in four major areas: attitude measurement, survey research methodology, research on organizational management, and applications of social science to important social problems. He earned his Ph.D. in psychology at Columbia with dissertation research, published in 1932, which developed the attitude measurement technique that bears his name. After teaching briefly at New York University, he moved to full-time research on organizational management. In 1939 he became the founding director of the Division of Program Surveys for the U.S. Department of Agriculture, where he made major contributions to methods of survey interviewing, probability sampling, and wartime public opinion research.

Following World War II, Likert founded the University of Michigan's Survey Research Center and later the Institute for Social Research, which under his leadership became the largest university-based social science research agency in the U.S. After retiring, he headed a consultation and research firm on organizational management until his death in 1981. Author of over 100 articles and six books, including New Patterns of Management *and* The Human Organization, *he was elected president of the American Statistical Association, and a director of the American Psychological Association, and he received the highest research award of the American Association for Public Opinion Research.*

This method uses only items that are clearly positive or negative toward the attitude object, whereas Thurstone's method also requires some relatively neutral items.

As the name "summated ratings" indicates, respondents' attitude scores are determined by adding their ratings for all of the items. This procedure is based on the assumption that all of the items are measuring the same underlying attitude. As a consequence of this assumption, it follows that all the items should be positively correlated, in contrast to the Thurstone method, which does not impose this requirement. Although the correlations among the items are not usually high, because each item is measuring its own unique content as well as the general underlying attitude, the assumption can be, and should be, checked. The usual way to do this is to correlate the score on each item with the total score for the whole pool of items combined (these are called item–total correlations). Any item with a correlation near zero is discarded because it is not measuring the common factor shared by other items.

A great strength of the Likert method is its use of **item analysis** techniques to "purify" the scale by keeping only the best items from the initial item pool. A common way of accomplishing this is to compare the group of respondents scoring highest on the total pool of items (say, the top 25%) with the group scoring lowest (the bottom 25%),

thus eliminating the middle group, whose attitudes may be less clear, less consistent, less strongly held, and less well-informed. If a particular item does not **discriminate** significantly between these groups—that is, does not have significantly different mean scores for the top and bottom groups—it is clear that it is measuring some other dimension than the general attitude involved in the scale. For example, in a scale of internationalist attitudes, a nondiscriminating item might be concerned with a hope for world peace, because high scorers (internationalists) and low scorers (isolationists) might both share this hope.

The Likert method of attitude scale construction quickly became and remains the most popular method, and a number of variations of it have also gained wide usage. One variation is to eliminate the "Undecided" or "Neutral" category, thus forcing respondents to choose between favorable and unfavorable stances. For instance, an item from the California F Scale, for measuring authoritarian or "fascist" attitudes, is scored as follows (Adorno et al., 1950, p. 68):

An insult to our honor should always be punished.

+ 1: slight support, agreement	− 1: slight opposition, disagreement
+ 2: moderate support, agreement	− 2: moderate opposition, disagreement
+ 3: strong support, agreement	− 3: strong opposition, disagreement

A more serious, and unfortunate, departure from Likert's procedure is the frequent omission of an item analysis. When this occurs, there is no empirical evidence that the items are all measuring the same underlying attitude, nor that they are useful, discriminating items. This situation is often signaled by use of the term "Likert-type" scale, which is apt to be an indication of hasty, slipshod research, quite out of keeping with Likert's own procedures.

Guttman's Cumulative Scaling Method

One of the limitations of both the Thurstone and the Likert techniques is that the respondent's attitude score does not have a unique meaning. That is, any given score can be obtained in many different ways. On a Likert scale, for instance, a respondent can obtain a midrange score by giving mostly "Undecided" responses, or by giving many "Strongly approve" responses offset by many "Strongly disapprove" responses, or by both "Approve" and "Disapprove" responses. Using the summated ratings (or more commonly, the average response to the items) does not tell us much about the pattern of responses or the responses to individual items.

Guttman (1944) proposed a method in which scores would have unique meanings. This was to be accomplished by ensuring that response patterns were **cumulative**. That is, in the Guttman method, a respondent who is moderately favorable to the attitude object should answer "yes" to all of the items accepted by a mildly favorable respondent *plus* one or more additional items. Similarly, a strongly favorable respondent should endorse all the items accepted by moderately favorable respondents *plus* additional one(s).

This reasoning can be clarified by some examples. Actually, the steps on the Bogardus Social Distance Scale, previously discussed in this chapter, apparently meet these requirements. A respondent who was very unfavorable toward Cubans, for instance, might be willing to accept them to citizenship in the country but not to the higher categories. Another person might agree to citizenship and also to equal employment. A favorable respondent might accept both of these items and also endorse accepting Cubans into his neighborhood and his social club; and so on, up to respondents who agreed with all the items.

Box 3–4 An Example of a Guttman Scale

Attitudes toward religious fundamentalism and its role in current American politics were measured in a study of the 1980 U.S. election conducted by the Center for Political Studies at the University of Michigan. Responses to interview questions were obtained from a representative national sample of over 1,200 white adults.

The six-item Guttman scale, which was constructed from the survey responses, is shown below. Items are listed here in rank order of the percentage of respondents agreeing with them, but in the actual interview they were arranged in a mixed-up order. The index of reproducibility of the scale was .925 (meaning only 7½% inconsistent responses). This is a Guttman scale because of the decreasing percentages of pro-fundamentalist answers on the successive questions (though it is unusual to have two items as close together in percentage of agreement as questions 3 and 4 here), and because most respondents who agreed with any given item also agreed with all *of the lower-numbered items (as shown by the index of reproducibility).*

Some evidence of the scale's validity is that, of 11 current political issues, its highest correlations were with opposition to abortion and support for school prayers.

	Items (in rank order, not in their order in the interview)	*% agreeing*
1	*Religion is an important part of one's life.*	*73*
2	*The Bible is God's word and all it says is true.*	*44*
3	*I feel favorable toward evangelical groups like the Moral Majority.*	*30*
4	*Religion provides a great deal of everyday guidance.*	*28*
5	*I am born again.*	*21*
6	*I feel close to evangelical groups active in politics such as the Moral Majority.*	*6*

Source: Miller, A. H., & Wattenberg, M. P. (1984). Politics from the pulpit: Religiosity and the 1980 elections. *Public Opinion Quarterly, 48,* 301–317. Copyright 1984 by The Trustees of Columbia University. Reprinted by permission of the University of Chicago Press.

Guttman suggested that, if a scale displays the cumulative pattern just described, we can be sure that it is **unidimensional**—that is, it is measuring just one underlying attitude. By contrast, Thurstone and Likert scales may be measuring two or more underlying dimensions. Guttman has proposed a quantitative index for determining the unidimensionality of a scale, and scales that meet Guttman's criteria are apt to be quite short (perhaps 4–10 items) and restricted to a narrow topic.

Box 3–4 presents an example of a Guttman scale constructed to measure attitudes toward politicized religious fundamentalism or the "religious right" (the attitude object). Notice that all six items are on a rather narrow topic, concerning various signs of religious fundamentalism, whereas many other aspects of religiosity are not represented. Of course, if desired, these could be measured by other Guttman scales on such topics as specific religious beliefs, frequency of religious activities, or degree of ethical behavior.

To develop a unidimensional scale by Guttman's procedure, an initial pool of items is given to a large group of respondents, each item being stated in a "yes–no" or "agree–disagree" format. Next, the items are arranged according to the number of respondents agreeing with them. In this procedure, by definition, the item agreed to by the *fewest* respondents is the item most favorable toward the attitude object (e.g., the "Moral Majority"

in the scale shown in Box 3–4); that is, it is the most-difficult-to-accept item. Each respondent's score is then determined very simply: It is merely the rank number of the most favorable item that he or she endorsed (answered in the scored direction). The answers of each respondent are examined separately (usually by computer). This is done to discover all instances of inconsistent response patterns: that is, cases in which a respondent endorses an item and fails to endorse one of the less-favorable items.

According to the theory of measurement underlying this scaling method, each such instance is considered a response error, and no more than 10% of inconsistent responses are allowed if a scale is to be considered unidimensional. (Guttman refers to this as an **index of reproducibility** of 0.90 or higher.) Items that have many inconsistent responses are probably measuring a different underlying dimension, and accordingly they are deleted from the pool of items. After a number of rounds of computation and discarding of items, a short scale may be developed that meets Guttman's criteria for unidimensionality. However, critical analyses have demonstrated that even more procedural safeguards than those recommended by Guttman are necessary to be sure that a truly unidimensional scale has been developed (Dawes & Smith, 1985).

Osgood's Semantic Differential

In contrast to the preceding methods of constructing attitude scales, Osgood's Semantic Differential is actually a scale in itself. However it is a scale of such a general sort that it can be applied to any attitude object. This has the great advantage that researchers do not have to construct and try out a new scale every time they want to study a new topic. No doubt this convenience is a major reason for the sustained popularity of the Semantic Differential since it was introduced (Osgood et al., 1957).

The reason for the name "Semantic Differential" is that the technique attempts to measure the **connotative meaning** of the concept or object being rated: that is, its implied meaning, or differential connotations to the respondent. In contrast to the other major attitude-scaling methods, the Semantic Differential does not consist of opinion statements about the attitude object. Instead it uses a series of 7-point scales with two opposing adjectives at the ends of each scale (e.g., "good" and "bad"). Respondents check the point on each scale that corresponds to their impressions of, or feelings about, the object or concept being rated. An abbreviated example of the instructions and the rating form is shown in Box 3–5.

Osgood and his colleagues (1957) reported a great deal of research on the application of this Semantic Differential approach to the measurement of a wide variety of concepts, including attitudes toward elderly people, gender groups, substance use, psychopathology, menopause, and work. Notably, the method has been successfully applied in many different cultures and subcultures.

Using the method of factor analysis, Osgood and his colleagues studied the underlying dimensions in connotative meaning, and time after time they came up with generally similar results. They concluded that there are three basic dimensions on which people make semantic judgments, and these are applicable quite universally to varied concepts, varied adjectival rating scales, and various cultures. The three dimensions are as follows: (a) the **evaluative** dimension, involving adjectives such as good–bad, beautiful–ugly, kind–cruel, pleasant–unpleasant, and fair–unfair; (b) the **potency** dimension, marked by adjectives such as strong–weak, large–small, and heavy–light; and (c) the **activity** dimension, identified by adjectives such as active–passive, hot–cold, and fast–slow.

Of these dimensions, the one most heavily weighted in people's judgments is evaluation. Osgood (1965) recommended using it as the prime indicator of attitude toward the object.

Box 3–5 An Example of a Semantic Differential Rating Task

Both the instructions and the rating form are substantially shortened in this demonstration example. Ordinarily many concepts to be rated would be presented to each respondent in a stapled booklet, one concept on each page; and more adjective scales might also be used for each concept. Note that the end of the scale representing the positive pole on the dimension is systematically varied between left and right.

INSTRUCTIONS: The purpose of this study is to measure meanings of certain things to various people by having them judge them against a series of descriptive scales. In taking this test, please make your judgments on the basis of what these things mean to you.

Here is how you are to use these scale: If you feel that the concept at the top of the page is very closely related to one end of the scale (for instance, very fair), you should place your check mark as follows:

fair X:__:__:__:__:__:__ unfair

If you feel that the concept is only slightly related to one or the other end of the scale (for instance, slightly strong), you should place your check mark as follows:

weak __:__:__:__:X:__:__ strong

The direction toward which you check, of course, depends on which of the two ends of the scale seems most characteristic of the thing you're judging.

If you consider the concept to be neutral on the scale, both sides of the scale equally associated with the concept, or if the scale is completely irrelevant, unrelated to the concept, then you should place your check mark in the middle space.

Rate the concept on each of these scales in order, and do not omit any. Please do not look back and forth through the items. Do not try to remember how you checked similar items earlier in the test. Make each item a separate and independent judgment. Work at fairly high speed throughout this test. Do not worry or puzzle over individual items. It is your first impression, the immediate "feelings" about the items, that we want. On the other hand, please do not be careless, because we want your true impressions.

SEPARATION OF CHURCH AND STATE	(Dimension)*
good __:__:__:__:__:__:__ bad	(evaluative)
weak __:__:__:__:__:__:__ strong	(potency)
active __:__:__:__:__:__:__ passive	(activity)
large __:__:__:__:__:__:__ small	(potency)
slow __:__:__:__:__:__:__ fast	(activity)
unfair __:__:__:__:__:__:__ fair	(evaluative)

*Of course, the dimensions are not shown on the respondents' forms.

Source: Adapted from Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning* (pp. 36–38, 82–84). Urbana: University of Illinois Press.

Clearly it is an affective dimension whereas the other two seem more cognitive in nature. Normally each dimension can be measured reliably by the use of only three or four adjective scales, so use of the Semantic Differential is simple and convenient for the investigator and relatively easy for respondents as well.

Final Comments on Attitude Scales

The work by Osgood illustrates the possibility of multidimensional scaling of attitudes. Although most attitude scales have concentrated on measuring the **magnitude** of attitudes—that is, their degree of favorability or unfavorability (also sometimes called their **valence**)—several other dimensions of attitudes have been suggested as worthy of study. In particular, these dimensions include the **complexity** or elaboration of attitudes, their **centrality** or importance to the person who holds them, and their **accessibility** (closeness to awareness, or readiness for expression). The structure of attitudes is considered in more detail in Chapter 5.

It should also be emphasized here, as was mentioned in Chapter 1, that carefully constructed attitude scales have quite rarely been used by researchers and only occasionally utilized by attitude pollers for practical assessment. Instead, the major contribution of these elaborate measurement methods has been to provide theoretical understanding of specific domains of attitudes.

Over the years, a number of other attitude-scaling methods have been proposed (cf. Edwards & Kilpatrick, 1948; Coombs, 1950; Hambleton, 1989; Mitchell, 1990; Kenny & Judd, 1996). In Chapter 4, we examine *implicit measures* of attitudes, which have been developed in the past 10 years and provide a very different approach to measuring attitudes. In addition to implicit measures, there is one other approach that deserves comment. Item response theory (IRT) has become increasingly used by researchers as computer programs became available to perform its required complex computations.

The five measurement approaches previously summarized (Bogardus, Thurstone, Likert, Guttman, and Osgood) all produce scores that are useful in describing a specific sample. However, when scores are based on parameters established with a prior sample (as they are with the Guttman and the Thurstone scales), then the scores are *group-dependent*. That is, they are difficult to compare across dissimilar groups. In addition, all of the scales described above are *test-dependent*, in that the meaning of the scores depends on the specific items used in the scale. It would not be appropriate to replace some of the items in the scale for one sample and then to make comparisons of scores across samples.

Item Response Theory (IRT) has been used primarily in the development of achievement and aptitude tests, but it is also beginning to make its way into attitude measurement. The goal of IRT is to obtain a measure that is applicable to groups and individuals with widely varying ability levels. In attitude terms, this would mean groups with extremely positive or negative attitudes. Items are included in the scale based on extensive testing, and they are selected to range from very easy (i.e., almost everyone agrees with the item) to very difficult (almost no one agrees with it). Different items are given to different samples, but because each item's favorability to the attitude object has been premeasured, comparable scores can be derived for the various samples. In many ways, IRT is an extension of the scale-value aspect of Thurstone scaling, but with a different mathematical approach to obtaining the scale values for each item. Typically, in IRT models, researchers obtain scores for each item in the scale as well as for each respondent. For an overview of IRT, see Hambleton (1989) or Embretson & Reise (2000).

Item response theory has received considerable attention by researchers over the past 20 years, but its merits are still widely debated (Anastasi & Urbina, 1997). Although it

has been used sporadically in attitude studies, the most common method used in attitude research continues to be Likert measures. Fortunately, studies comparing the different methods of attitude measurement that we have described have found them to be positively correlated—Fishbein and Ajzen (1974) reported typical intercorrelations of around +.7, though Tittle and Hill (1967) found lower figures averaging around +.5. Both studies showed the Likert scale to be most highly correlated with the various other attitude measures.

In addition to the limitations noted by IRT researchers, another limitation shared by all explicit measures of attitudes is that the scales they produce are **ordinal** scales rather than equal-interval or ratio scales. This means that respondents can successfully be placed in their *rank order* on the attitude dimension, but we cannot be sure that the actual attitudinal distance between two values on the scale is equal to the distance between two other values. For instance, on a Likert scale, is the distance between "Undecided" and "Approve" (3 and 4) the same as the distance between "Approve" and "Strongly approve" (4 and 5)? The two distances are numerically equal, but they may not be psychologically equal. Even though Thurstone's method strives to achieve "equal-appearing intervals," it is nevertheless an ordinal scale rather than an interval scale.

Technically, ordinal scales should be treated with nonparametric, distribution-free statistical techniques involving measures such as the median. For this reason it is statistically improper to add or multiply scores together, compute mean scores, use *t* tests, analysis of variance, or any of the other widely used parametric statistics. However, these restrictions are almost universally disregarded, largely because statistical research has shown that in most instances violations of the assumptions underlying the use of parametric techniques do not lead to serious distortions of their results. Thus scores are customarily derived through summation or averaging, and *t* tests and *F* tests are used on attitude scale results. It is well to keep in mind, however, that occasionally, when distributions are markedly skewed or variances are grossly different, use of parametric techniques may produce misleading conclusions (Dawes & Smith, 1985).

RELIABILITY AND VALIDITY OF MEASUREMENT

There are two essential characteristics for attitude scales, as for all other types of measurement: reliability and validity. **Reliability** means consistency of measurement. A measurement that is unreliable is like an elastic tape measure, which stretches a different amount every time it is used. Two kinds of reliability are commonly reported: **internal consistency** measures, showing the amount of agreement between different items intended to assess the same concept; and **stability** measures, indicating the consistency of scores on the same scale at two different points in time. Both kinds are generally reported in terms of correlation coefficients. Internal consistency measures include *split-half* coefficients, *alternate-form* agreement, and the *alpha coefficient* of internal homogeneity of items (Cronbach, 1984). Stability is usually reported as *test–retest* correlations for the same group of subjects taking the same test or other measurement at two points in time. For verbal attitude or information measures, these two occasions need to be far enough apart that subjects are unlikely to remember their previous answers and simply repeat them on the second measurement occasion—usually a week or two at a minimum.

Unreliability of measurement in verbal scales can often be combated by several means. Sometimes it results from very coarse measurement (e.g., simply "Agree" or "Disagree"), in which case it can usually be reduced by increasing the number of response alternatives (e.g., several degrees of agreement or disagreement). Thus, even if a person gives a slightly

different response on another occasion, he or she will not have shifted from one end of the dimension to the other. Another common source of unreliability in multi-item attitude scales is that items are not "pure" measures of the characteristic that one is attempting to measure, and thus they are often only weakly or moderately correlated with each other. The customary way to solve this problem is to add more items of the same sort to the scale, because statistical principles of measurement guarantee that, for any given level of item intercorrelation, a longer scale will be more reliable than a shorter one. However, limits to this approach are the availability of appropriate items and the feasible length of the scale. Other ways of reaching sounder statistical conclusions by improving measurement reliability are discussed by Cook and Campbell (1979), Cronbach (1984), and Thompson (2002).

Validity means accuracy or correctness of measurement. Measuring instruments can be reliable without being valid—for example, a bathroom scale that consistently gives too heavy readings. However, they cannot be valid if they are not reliable—for instance, the many different readings given by an elastic tape measure would almost all (or all) be wrong, and thus the tape measure would not be a valid instrument.

The validity of a measuring instrument is often determined by comparing its results with a **criterion**—an accepted, standardized measure of the same characteristic. For example, butchers' scales are calibrated and tested against a very accurate master instrument. In psychological measurement, a criterion may be a well-established instrument, as in using the Stanford-Binet intelligence test as a standard of comparison for the results of a newly devised IQ test. However, in many cases there may be no well-established criterion instrument for the characteristic being measured, as when research begins on a new topic that has not been measured before. This is frequently true in the area of attitudes, and it necessitates an approach similar to pulling oneself up by one's bootstraps. The typical approach here is termed **construct validation**, which involves computing a network of relationships between the new measure and other relevant characteristics and comparing the obtained correlations with those expected on a theoretical basis. If there is generally good correspondence, that constitutes support for the instrument's validity.

Other aspects of validity are discussed in Chapter 5, and extensive elaborations of threats to validity in reaching conclusions from psychological data and ways of counteracting these threats may be found in Cook and Campbell (1979), Cronbach (1984), and Bickman (2000).

PROBLEMS AFFECTING THE VALIDITY OF ATTITUDE SCALES

Pause for a moment, and think in detail about what respondents have to do in the process of answering an attitude question:

> Respondents first interpret the attitude question, determining what attitude the question is about. They then retrieve relevant beliefs and feelings [from their memory]. Next they apply these beliefs and feelings in rendering the appropriate judgment. Finally, they use this judgment to select a response. (Tourangeau & Rasinski, 1988, p. 299)

Problems can occur at each of these stages, which may reduce the validity of respondents' answers. Also, as mentioned in Chapter 1, the fact that people sometimes *construct* attitude responses on the spot without any prior consideration of the issue, rather than retrieving a previously formed attitude from their memory, would sharply decrease both the reliability and validity of such attitude statements.

The *wording* of attitude questions is one of the main factors affecting the validity of attitude scales. However, because principles regarding the wording of attitude questions are also applicable to the wording of public opinion interviews, they are discussed in detail in Chapter 6.

The major problem to be discussed here is the ways in which response sets can invalidate attitude questionnaire answers. **Response sets** are systematic ways of answering that are not directly related to the question content, but which represent typical behavioral characteristics of the respondents. Several types of response sets are mentioned in the next four subsections, and some possible solutions to them are discussed.

Carelessness

When respondents are unmotivated or careless, their answers will be variable and inconsistent from moment to moment or from one testing session to another. Such a situation will reduce the questionnaire's reliability, and unreliable questionnaires are necessarily low in validity.

Some carelessness and low motivation can be minimized by the researcher building good rapport with the respondent, stressing the importance of the task, and engaging the respondent's interest in it. However, despite such precautions, some respondents may still answer carelessly or fail to follow directions through misunderstanding or poor comprehension. Therefore the response sheets are usually scanned visually, and the data are either discarded or analyzed separately for respondents who (a) omit answers to many items, (b) answer almost all items in the same way, or (c) show systematic patterns of responding (for example, a, b, c, d, a, b, c, d).

Social Desirability

The social desirability response set is the tendency to give the most socially acceptable answer to a question, or to "fake good." It operates both in attitude scales and in public opinion interviews. For example, people will rarely describe themselves as dishonest, even though almost everyone occasionally fudges on the truth or cheats a little bit (by glancing at an opponent's cards, etc.). In extensive studies on this topic, Edwards (1964) showed that personality characteristics that are considered as desirable in our culture are also ones that are claimed by most respondents as applying to themselves, and vice versa. In one study of 140 characteristics, the correlation was +0.87, an almost perfect relationship. Edwards (1964) developed a personality scale that indicates the degree of an individual's tendency to give socially desirable answers, and other authors have constructed similar scales (Crowne & Marlowe, 1964; Schuessler, Hittle, & Cardascia, 1978).

These scales are useful for identifying respondents with a high tendency to provide socially desirable answers, but removing social desirability from the results of a study is more difficult. To control for social desirability responding, Edwards advocated the use of **forced-choice** items. In this technique of scale construction two items of approximately equal social desirability, but indicating, for instance, two different social needs, are paired together. The respondent is asked to choose the one that is most true of himself or herself. This was a creative proposal, but unfortunately the evidence of its success in solving the problem of social desirability responding is disappointing (Barron, 1959; Scott, 1968). Consequently, only a few scales have been built in this way, the best known of which is Rotter's (1966) scale of internal versus external locus of control.

Unfortunately, none of the available methods for combating social desirability responding is entirely satisfactory. The techniques that are most often used are as follows:

(a) selecting innocuous items, for which social desirability does not appear to be an issue; (b) providing anonymity for the respondents; (c) stating that there are no right or wrong answers, because the items cover matters of opinion rather than fact; (d) urging respondents to answer honestly and stressing that it is their own opinions that are desired; (e) use of the forced-choice technique of item construction, previously discussed; and (f) auxiliary use of personality scales to identify respondents who are particularly high or low in social desirability responding, and either excluding these participants from the analyses or statistically removing the variance stemming from their individual differences in social desirability. In his review of techniques for controlling social desirability response bias, Krosnick (1999a) suggested that researchers test for the presence of social desirability responding by having some respondents answer questions in an ordinary self-report fashion, and having others answer in a way that attempts to reduce social desirability through one or more of the approaches discussed above.

Extremity of Response

An extremity response set can occur only on items that have more than two alternative answers. For example, on a Likert-type scale having responses scored from +3 to −3, an extremity response set would be demonstrated by a respondent who picked mostly +3 and/or −3 answers. Its opposite, a midrange response set, would be shown by a large number of +1 and/or −1 answers. In one nationwide study of high school students, black students were found to give many more extreme responses than whites (Bachman & O'Malley, 1984). Other studies have found that Hispanic Americans give more extreme responses than European Americans (Marin, Gamba, & Marin, 1992), and older, less educated, and lower-income respondents give more extreme responses (Greenleaf, 1992).

There has been little study of the effects of extremity response sets or midrange response sets on questionnaire validity. Their effects can be reduced if equal numbers of items on a scale are keyed in the positive and negative directions, for then the +3 answers of an extreme responder will tend to counterbalance his or her −3 answers (and similarly for the +1 and −1 answers of a midrange responder). Another possible remedy is to eliminate the extremity response set altogether by use of items with only two alternatives (Yes–No or Agree–Disagree).

Acquiescence (Yea-Saying)

The most thoroughly studied aspect of acquiescence is the **agreement** response set, or yea-saying, defined as a tendency to agree with any questionnaire item regardless of its content. It has been studied extensively in the California F Scale measure of authoritarianism (Adorno et al., 1950), but it also is an issue in many other attitude and personality scales, particularly in the Minnesota Multiphasic Personality Inventory (Bradburn & Sudman, 1979). An example of agreement responding is answering "Yes" to both of the following items: "Jews are more willing than others to use shady practices to get ahead" and "Jews are just as honest as other businessmen" (Jackman, 1973). Such patterns of response have been found to be more common among people with lower education and income (Ware, 1978), women, children (Poole & Lindsay, 2001), individuals diagnosed as mentally retarded (Finlay & Lyons, 2002), and in more collectivistic cultures (Cheung & Rensvold, 2000; see also Narayan & Krosnick, 1996). Acquiescence bias occurs most often with difficult questions, or when respondents are fatigued from answering a large number of questions, and during phone interviews more than during face-to-face interviews (Krosnick, Narayan, & Smith, 1996). In a major research review on the topic,

Krosnick (1999a) concluded that "acquiescence occurs when people lack the skills and motivation to answer thoughtfully and when a question demands difficult cognitive tasks be executed in order for a person to answer precisely" (p. 41).

An example of a scale on which acquiescence is a problem is the California F Scale. As the result of an unfortunate decision during the construction of the scale, all 28 items were worded in such a way that agreement indicated authoritarianism and disagreement indicated lack of authoritarianism—that is, all items were keyed in the positive direction. Before 1950, when the authoritarianism studies were being formulated, this was not recognized as a major issue in scale construction, but it has since become so.

One approach for reducing agreement response bias effects during the construction of a scale is to *reverse the wording and the keying* of half of the items from that of the other half. For example, in addition to the item "I like to eat sushi," we might also present the item "I do not enjoy eating sushi." Responses to the latter item would then be *reverse-coded*, so that scores on the two items would be positively correlated. The result is called a **balanced scale**—that is, one having half of the items on the scale scored if the answer is "true," and half scored if the answer is "false." If the two groups of items are equally good, are positively intercorrelated, and have an equal spread of responses, this procedure will cause any agreement response effect to cancel out across the two groups of items.

However, this was not done on the California F Scale, and debates raged for years about the resulting problems. One group of authors (e.g., Bass, 1955; Campbell et al., 1960) claimed that the scale was more a measure of acquiescence than of authoritarianism. Another group, using different statistical methods, concluded that there was little relationship between authoritarianism and acquiescence (Couch & Keniston, 1960). A third group (e.g., Christie, Havel, & Seidenberg, 1958) found that there was some mixture of acquiescence in F Scale scores, but argued that there *should* be, because agreeing with an authoritatively worded statement is really one aspect of being an authoritarian.

The use of reverse-coded items and balanced scales has become common practice in attitude measurement, and as a result it might appear that the problem of acquiescence bias has been solved. However, it is not a simple matter to construct attitude measurement items that are reverse-coded. It is often difficult to devise questions that avoid using the word "not" or another similar negation; and questions containing "not" are apt to be cumbersome and can increase a respondent's fatigue (and thereby inadvertently *increase* acquiescence). In addition, the increased cognitive resources needed to interpret and respond to these reverse-coded items may lead to differential rates of acquiescence across respondents. For example, when confronted with a longer or more confusing item, respondents who are less motivated to think about the item might simply say "yes"—an acquiescent response. In the final analysis, reverse-coded questions are only a partial solution to acquiescence bias, and it is more important to ask a few direct, carefully worded items, encourage respondents to answer honestly, and implement steps to ensure confidentiality or anonymity (Krosnick, 1999a).

A nay-saying or disagreement response set—that is, a tendency to disagree with any item regardless of its content—is the other end of the agreement dimension (cf. Knowles & Condon, 1999). It is relatively rare and has been little investigated. One study found the disagreement response set more common among Republicans than among Democrats (Milbrath, 1962).

THE BOGUS PIPELINE

Because of the response sets discussed in the previous section, the validity of self-report measures is always open to question. However, using them in conjunction with a

related objective measure can sometimes increase their validity. For instance, in a study of adolescents' self-reports of smoking, the amount of smoking reported was significantly higher when reports were taken after a demonstration that recent smoking could be detected from the presence of carbon monoxide in their breath (Bauman & Dent, 1982).

This carbon monoxide measure was a true indicator of smoking, but the same effect of increased validity should occur if respondents merely *believe* that there is a true measure of their behavior or feelings available. This principle is the basis of the so-called **bogus pipeline**, in which participants are falsely convinced that some elaborate electronic apparatus can detect their true feelings. This technique typically results in their reporting higher levels of various socially undesirable attitudes or behaviors such as racial prejudice, eating disorders such as bulimia, smoking marijuana, and drunk driving (Jones & Sigall, 1971; Quigley-Fernandez & Tedeschi, 1978; Roese & Jamieson, 1993; Tourangeau, Smith, & Rasinski, 1997).

We should emphasize that using techniques such as the bogus pipeline raises several ethical questions (Aguinis & Henle, 2001). Issues of privacy, informed consent, deception, and debriefing must all be carefully considered. However, these are not clear-cut issues, and researchers differ on how they should be resolved. On the one hand, minor deception is often socially acceptable (as in conventional politeness and "little white lies"), and on the other hand, full debriefing of participants about the research they took part in may sometimes do more harm than good. Dawes and Smith's (1985) often-cited review recommended following social norms about what is considered ethical outside of the laboratory, and using deception only in cases where it seems so innocuous that no debriefing should be needed. A fuller discussion of ethical issues in research is presented in Chapter 12.

Given the ethical issues associated with the bogus pipeline, an important question is whether there are other ways to increase the validity of self-report measures that do not require deception. Several studies have found that using techniques that ensure anonymity can yield equally valid responses (Hill, Dill, & Davenport, 1988).

OTHER WAYS OF MEASURING EXPLICIT ATTITUDES

In the preceding sections, all of the methods that we have described for measuring explicit attitudes have relied on language, either written or oral. But there are other ways of assessing explicit attitudes that do not require linguistic skills.

Graphical Scales

Feeling Thermometer. One technique for measuring attitudes is through graphical or pictorial rating scales. One example of such a scale is the feeling thermometer. It asks respondents to indicate their attitude on a scale of degrees, typically ranging from 0° (Very cold) to 100° (Very warm), with 50° representing "No feeling at all." The feeling thermometer has been used quite often in public opinion research (Berman & Stookey, 1980), and particularly to measure attitudes toward political candidates (e.g., Granberg & Brent, 1980; Beasley & Joslyn, 2001).

For example, Fox and Smith (1998) used a feeling thermometer to examine how attitudes toward political candidates were affected by the candidate's gender. Students enrolled in American government classes at two universities, in California and Wyoming, rated four hypothetical candidates for the U.S. House of Representatives. The name of each candidate (indicating gender) was followed by his or her positions on a number of issues. On half of

the forms the candidate was female (e.g., Lisa Jennings), and on the other half the candidate was male (e.g., Bill Jennings). Results for the California sample (which was selected to be a liberal sample) showed no evidence of gender bias—ratings of the candidates did not differ by their gender. For example, the average rating for Lisa Jennings was 68.3, and for the similarly described Bill Jennings it was 69.4—a nonsignificant difference. However, the more conservative Wyoming sample did show evidence of gender bias—Lisa Jennings' mean was 59.4, whereas Bill Jennings' mean was 65.1.

Body-Shape Preference. Another example of a graphical scale used to measure attitudes comes from research on female body size and shape. Singh (1993) developed a pictorial measure that assesses preference for female body size (thin, normal, overweight) and shape (waist-to-hip ratio). The waist-to-hip ratio is determined by the smallest width of the waist divided by the largest width of the hips. As shown in Figure 3–1, a waist-to-hip ratio of 1.0 means equal waist and hip sizes, whereas a ratio of 0.7 indicates a waist considerably smaller than the hips. Research with this scale has generally found a preference for "normal" weight combined with a smaller waist-to-hip ratio of around 0.7

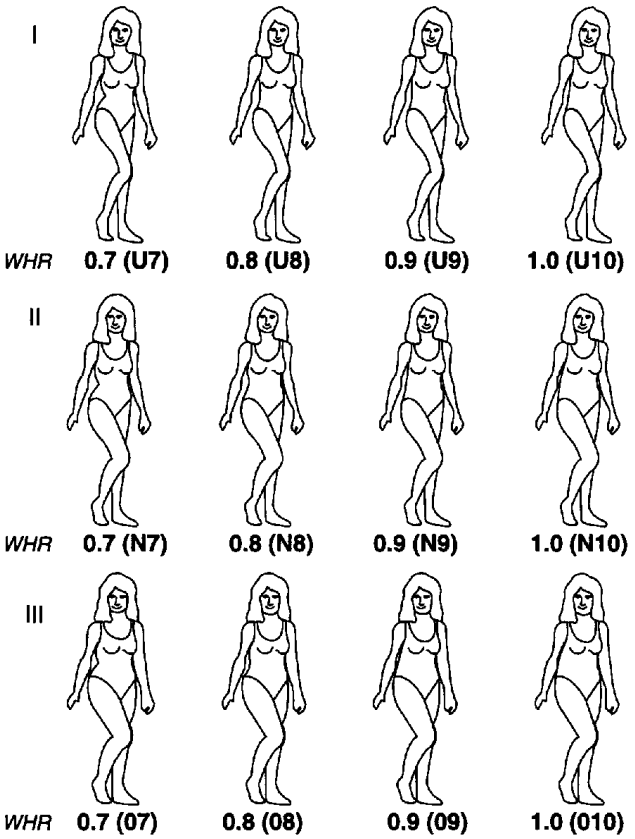


FIGURE 3–1 Stimulus figures representing three body-weight categories: underweight (I), normal weight (II), and overweight (III). Waist-to-hip ratios (WHR) are shown under each figure in each weight category, along with a letter and a number in parentheses that identify the body-weight category and WHR.

Source: Singh, D. (1993). Adaptive significance of female physical attractiveness: Role of waist-to-hip ratio (p. 298). *Journal of Personality and Social Psychology, 65,* 293–307. Copyright © 1993 by the American Psychological Association. Reprinted with permission.

(Markey et al., 2002), although several studies have indicated a tendency for European American respondents to prefer thinner weight (Gordon, 2000). The preference for smaller waist-to-hip ratios has been found among both men and women, across a wide range of ages (sample ages ranged from 25 to 85), and across various ethnic groups (African American, Asian American, Mexican American, and Caucasian).

Observations of Behavior

Compared with verbal self-report measures, behavioral measures of attitudes have been seldom used and consequently are poorly developed and crude in their methodology. In large part this is because they are difficult, time-consuming, and expensive to utilize.

The most straightforward type of behavioral observation is one made in a natural setting, such as watching for aggressive episodes in a schoolyard. However, the time-consuming, tedious nature of such observation has led to use of more standardized situations, which are structured so as to elicit the behavior of interest more easily. Cook and Selltiz (1964) described three different types of such standardized approaches: (a) apparently unstaged standardized situations in which a person's behavior can be observed, (b) staged role-playing situations in which the person is asked either to respond as he or she would in real life, or to take the part of a particular other person, and (c) use of sociometric choices which the participant believes will have real-life consequences (e.g., choice of which members of a group to work with on a joint task). In all three of these approaches, of course, the situation chosen is one in which the attitude objects (e.g., children of a different racial group) are presented in some way.

A key advantage of this approach is that *participants can be convinced that there will be real-life consequences* flowing from their responses (e.g., they will actually get to work with the classmates they choose). Alternatively, it is also possible to represent the attitude objects only *symbolically* (i.e., in words or pictures) rather than having them physically present. An example is Harter and Pikes' (1984) pictorial scale that measures perceived competence and social acceptance in young children (cf. Rainey & Rust, 1999). However, this procedure tends to measure the respondents' behavioral intentions (what they *say* they would do) rather than their actual behavior toward the attitude object—thus it is a return to a self-report form of measurement.

Because Fishbein and Ajzen (1972) have reported very high correlations between behavioral intentions and behavior, the use of behavioral-intention measures may be justifiable here. Cook and Selltiz (1964) defended it on the grounds that it is less sensitive than real-life behavior to a variety of extraneous influences (e.g., previous acquaintance or lack of it with individuals being responded to). However, we should emphasize that it is sometimes possible to observe actual behavior in situations where extraneous influences are relatively inoperative. For instance, slight vertical or horizontal head movements are good indicators of a person's attitude toward a persuasive message (Wells & Petty, 1980). Similarly, in a small-group discussion situation, choice of a seat next to someone in a wheelchair, rather than one farther away, could indicate a person's attitude toward people with disabilities. Because Wicker (1969) and others have shown that there is often only a low relationship between verbal self-report attitude measures and behavior, the use of actual behavior measures may be preferable to behavioral-intention measures.

One famous example illustrates the types of behavioral-intention measures that have been used. DeFleur and Westie (1958) developed a method in which white subjects, after seeing some relevant interracial slides, were asked whether they would be willing to be photographed with a black person of the opposite sex. The subjects were also requested to sign a "standard photograph release agreement" indicating which of a variety of purposes

they would be willing to have such a photograph used for—ranging from showings solely to professional sociologists for research purposes to a nationwide publicity campaign in favor of racial integration. The number of uses that they checked was taken as an indicator of favorableness toward blacks.

Unobtrusive Measures

One of the most promising ways of supplementing attitude scale scores is the use of **unobtrusive measures** of behavior (observations made without attracting the attention of the people being studied), as suggested in a fascinating paperback book by Webb et al. (1981). Such measures may be direct observations of behavior, such as standing in a high lookout and counting the number of students taking different paths across campus, or watching children's aggressive behavior in a schoolyard. However, several types of unobtrusive measurement can also substitute for tedious long-term observation as indicators of attitudes: (1) Direct measures of **preference** can be counted, such as candidate bumper stickers in a parking lot (Wrightsman, 1969). (2) **By-products** or waste products can show people's attitudes; for instance, counts of beer cans and liquor bottles in trash can gauge the amount of drinking and the preferred beverages in an area (Rathje & Ritenbaugh, 1984). (3) Measures of **erosion**; for instance, paths worn in the grass across campus, or the rate of emptying of ice cream tubs, can indicate preferred routes or flavors. (4) Measures of **accretion**; people's interests can be estimated from the amount of dirt on pages of library books or the number of fingerprints and nose smudges on glass cases in museums (Webb et al., 1981).

Another good example of unobtrusive measures used in research studies is the percentage size of tips left for a restaurant server (Lynn & Simons, 2000). Similarly, the forwarding of letters in the **lost letter technique**, in which stamped and addressed letters are dropped in shopping areas, can gauge community sentiment toward local organizations or election issues (Simmons & Zumpf, 1983) or attitudes toward specific social groups such as gay men or lesbians (Bridges & Rodriguez, 2000). Recently a newer version of the lost letter technique has been developed for electronic mail—the "lost e-mail" (Stern & Faber, 1997). Further use of such imaginative approaches could help to solve the problems inherent in interpreting the results of attitude-scale and opinion-interview research.

Performance on Objective Tasks

This measurement approach has been used somewhat more widely than the previous ones. Cook and Selltiz (1964) described it as follows:

> Approaches in this category present the respondent with specific tasks to be performed; they are presented as tests of information or ability, or simply as jobs that need to be done. The assumption common to all of them is that performance may be influenced by attitude, and that a systematic bias in performance reflects the influence of attitude. (p. 50)

Thus, in a sense, this approach is similar to observations of behavior. It differs in that the task is structured for the subjects, and that the relevance of their performance to measurement of their attitudes is usually quite thoroughly disguised.

Some examples may clarify how this can be done. Hammond (1948) devised an "information" test with alternative answers that were equally far on either side of the correct response (which was not provided as an alternative). He showed that the respondents' choices of erroneous responses were generally consistent with their own attitudes. For instance, a pro-union person would generally choose an answer that overestimated labor

Photograph courtesy of Lehigh University. Reprinted by permission.

Box 3–6 DONALD CAMPBELL, *Methodologist and Attitude Researcher*

Donald Campbell received nearly every major honor that psychology had to offer—notably, election to the National Academy of Sciences and the American Academy of Arts and Sciences, the presidency of the American Psychological Association, and its Distinguished Scientific Contribution Award. He was honored as a methodologist and a philosopher, a field researcher and a laboratory experimenter, and for work in anthropology, political science, and sociology as well as psychology.

Born in 1916, Campbell worked on a turkey ranch before taking his B.A. at the University of California at Berkeley. Following wartime service in the Navy he returned to Berkeley and completed a noteworthy dissertation on the consistency of racial attitudes. After teaching briefly at Ohio State and the University of Chicago, he settled in 1953 at Northwestern, remaining there until 1979, when he moved to Syracuse for three years and then to Lehigh University, where he continued to write until his death in 1996.

Campbell was widely known as a coauthor of books on unobtrusive measures and on experimental and quasi-experimental research methods. Among his 200-plus articles, one on indirect methods of measurement is particularly relevant to the topic of this chapter. Chapter 5 cites his research on attitude consistency; and in Chapter 12 his critique of attitude–behavior pseudo-inconsistency is described, and his call for planned experimentation on social and governmental programs is applauded.

unions' membership size, rather than an answer that underestimated it, whereas the opposite would usually be true for an anti-union individual. Similarly, Brigham and Cook (1970) had respondents judge the plausibility of pro-integration and anti-integration arguments, and the judgments were treated as indicators of the person's own attitudes toward racial integration.

Two problems are present in interpreting measures of this sort. If a person shows a consistent bias in performance, it seems safe to infer that the individual's attitudes are responsible. However, if a consistent bias is not shown, it may not be safe to infer that the person's attitude is a weak one, for we do not know how sensitive such measures are. Second, a particular bias in response might reflect either wishes or fears—"a member of the Communist party may overestimate the number of Communists in the United States, but so may a member of [an anti-Communist group]" (Cook & Selltiz, 1964, p. 51). Thus, additional information may be needed to determine the direction of the person's attitude from a biased performance.

In using any of the kinds of measures described in this chapter, it should be emphasized that the researcher's conclusions about people's attitudes is an inference from the particular

measures taken. This is true even when the measures used are individuals' self-reports of their own attitudes, for the researcher still has to decide whether the respondents truly are aware of their own attitudes and are reporting them accurately.

SUMMARY

Attitudes and opinions may be expressed in many colorful ways, but for purposes of scientific study, they must be classified into categories or measured on a quantitative scale. The development of attitude-scaling methods in the 1920s and 1930s was the first major application of quantitative measurement in the field of social psychology. In terms of the frequency of their use, the five most widely used scaling methods are Bogardus' scale of social distance toward various ethnic groups, Thurstone's method of equal-appearing intervals, Likert's method of summated ratings (the most popular of all), Guttman's cumulative scaling method of constructing a unidimensional scale, and Osgood's scale of connotative meaning, the Semantic Differential. All of these methods produce scales that are ordinal in nature, and therefore some caution must be exercised if parametric statistics are used in analyzing their results.

It is essential for attitude scales, like all measurement methods, to be both reliable (consistent) and valid (accurate) in their results. Problems that affect the validity of attitude scales include the response sets of carelessness, social desirability, extremity, and acquiescence (yea-saying). With due care in constructing and interpreting attitude scales, all of these problems can be at least partially overcome.

In conjunction with attitude scales, it is recommended that other less-common methods of studying attitudes also be more widely used in research, in order to provide a broader multidimensional measurement approach. These supplementary techniques include methods of increasing the validity of self-report measures, graphical scales, observations of behavior (particularly unobtrusive observations), and measures of performance on objective tasks in attitude-relevant situations. In addition, the following chapter discusses ways of measuring implicit attitudes.