In these segmented scales the respondent must make a decision that reflects either positively or negatively on the person being rated, because there is no neutral category in any graphic scale divided into an even number of segments. In segmented scales in which there is an uneven number of categories, the implication of the middle category would be "neutrality" (e.g., neither unpopular nor popular).

The implication in the use of segmented scales, as in the case of numerical ratings, is that adjacent segments or categories are psychologically equidistant. That is a problematic assumption in some cases (cf. Surber, 1984), even though it is also a basic one in certain standardized scaling procedures (discussed in the next chapter). Interestingly, one study that compared a number of different procedures (e.g., a numerical rating scale, a checklist, ranking of preferences, a forced choice between paired items) in terms of test-retest reliability found no appreciable differences among them—all the methods, at least in the context of marketing research, generated satisfactory reliability values (Kassarjian and Nakanishi, 1967). The results suggest that the selection of a rating format might best be based on considerations other than concerns about stability. For example, one might decide to use a multipoint scale in order to develop a more precisely defined pattern of what was being rated (King, King, and Klockars, 1983).

How many points or rating segments are optimal in a multipoint scale? Is it better when making a numerical rating scale, for example, to construct a 3-point scale, a 7-point scale, an 11-point scale, etc.? Another study examined the extent to which interrater reliability of a clinical rating scale was affected by the number of scale points from 2 all the way up to 100 (Cicchetti, Showalter, and Tyrer, 1985). The results were that internal-consistency reliability increased steadily up to a 7-point rating scale, beyond which no substantial increases occurred.

In making up *anchor* words, such as "extremely unpopular" and "extremely popular," researchers should try to select terms or short statements that are simple, unidimensional, and unambiguous. The anchors also need to be clearly relevant to the behavior or variables being rated, and consistent with other cues. Figure 8-2 shows a seven-point "mood questionnaire" based on the segmented type of graphic rating scale, in which the same anchor words are relevant to all eight variables (Rosnow, 1968). It is also important that the anchor words be as precise as possible, and that they stay clear of expressions with ethical or moral overtones. In the next chapter we also present several standardized rating scales, most of which incorporate some aspect of one or more of the three basic formats just discussed.

SYSTEMATIC RATING ERRORS AND THEIR CONTROL

The use of rating scales proceeds on the assumption that the person doing the rating is capable of an acceptable degree of precision and objectivity. However, there are a number of potential rating errors that need to be considered by researchers employing these instruments. One type of bias, the *halo effect*, refers to the fact either (1) that a judge who forms a favorable impression of someone with regard to some central trait will then tend to paint a rosier picture of the

Indicate the way you feel *now* by placing a check mark in the appropriate place.

	Not at all		A little		Quite a bit		Extremely
	1	2	3	4	5	6	7
Jittery							
Depressed							
Troubled							
Shaky							
Unhappy							
Excitable							
Downhearted							
Anxious							
	1	2	3	4	5	6	7
	Not at all		A little		Quite a bit		Extremely

FIGURE 8-2
An illustration of the segmented format of a graphic rating scale.

person on other characteristics or (2) that, if a rater has knowledge of some previous outstanding performance (positive or negative) or trait, this knowledge influences the person's current rating of the ratee for the better or the worse. That is, because we tend to judge a person in terms of a general mental attitude toward him or her, this "halo" will influence our opinions about specific qualities of the individual. For example, a student who is athletic or good-looking might be judged as more popular than is really the case. The halo effect is most prevalent when the trait or characteristic that is being rated (1) is not easily observable, (2) is not clearly defined, (3) involves relations with other people, and (4) is of some moral importance (Symonds, 1925).

As noted, forced-choice rating methods are designed to reduce this tendency of arriving at particular judgments based on a general mental attitude. Suppose the judge is dominated by a desire to make the person being rated "look good" and to avoid making him or her "look bad." In this case the judge would simply pile up favorable ratings using a numerical scale. In using the forced-choice procedure, however, there is no inkling as to which favorable and unfavorable traits receive weight toward the score.

Another type of bias is an *error of leniency*, named from the fact that some judges (i.e., lenient judges) will tend to rate someone who is very familiar, or someone with whom they are ego-involved, more positively than they should. Judges who are made aware of this failing may, however, "lean over backwards" and rate the person more negatively than they should, thus making the opposite type of error. When we anticipate the leniency error, a way to help counteract it is to arrange the rating scale differently. We might give only one unfavorable cue word ("poor") and have most of the range given to degrees of favorable responses ("fairly good," "good," etc.):

Poor	Fairly good	Good	Very good	Excellent

However, we treat or analyze the anchors numerically such that "good" is only a 3 on a five-point scale.

A third type of bias, an *error of central tendency*, occurs when the observer hesitates to give extreme ratings and instead tends to rate in the direction of the mean of the total group. This can usually be controlled in the way that the positive range was expanded in the sample shown above. Thus, in a numerical or a segmented graphic scale, it is usually a good idea to allow for one or two more points than are absolutely essential. If it is essential, for instance, that we have at least five alternative responses or segments, then it would be better to use a seven-point than a five-point rating scale, i.e., if observers are reluctant to use the extremes.

A fourth type of error, a *logical error in rating*, refers to the fact that many judges are likely to give similar ratings for variables or traits that seem logically related in their own minds but may not be related in any given target person. This type of error is similar to the halo effect, in that both increase the intercorrelation of the variables or traits being rated. The difference is that the halo effect results from the observer's favorable attitude about one personality as a whole, whereas the logical error results from the observer's perceptions as to the relatedness of certain variables or traits irrespective of individuals. The way to deal with this problem is to constuct very precise definitions and to make the instructions as explicit as possible.

In general, the most effective strategy for improving ratings is to use multiple judges who have been very carefully trained and to pool their ratings over many different trials employing similar and different rating methods. Observers who have been lectured on rating errors should be more sensitive to the different kinds of biases and also more mindful about how they respond. Training that includes practice sessions followed by discussions of each possible error will teach them what to watch for and how to be more precise and objective in their evaluations (see Cooper, 1981; Guilford, 1954).