chapter 12
Measurement Scales

>learning objectives

After reading this chapter, you should understand . . .

1 The nature of attitudes and their relationship to behavior.
2 The critical decisions involved in selecting an appropriate measurement scale.
3 The characteristics and use of rating, ranking, sorting, and other preference scales.

Any measurement must take into account the position of the observer. There is no such thing as measurement absolute, there is only measurement relative.

Jeanette Winterson,
journalist and author
They board the sleek corporate jet in Palm Beach and are taken aft to meet with the general manager of MindWriter, who is seated at a conference table that austerely holds one sheaf of papers and a white telephone.

"I'm Jean-Claude Malraison," the general manager says. "Myra, please sit here . . . and you must be Jason Henry. On the flight up from Caracas I read your proposal for the CompleteCare project. I intend to sign your contract if you answer one question to my satisfaction about the schedule.

"I took marketing research in college and didn't like it, so you talk fast, straight, and plainly unless we both decide we need to get technical. If the phone rings, ignore it and keep talking. When you answer my one question, I'll put you off the plane in the first Florida city that has a commercial flight back to . . . to . . ."

"This is Palm Beach, Jean-Claude," says the steward.

"What I don't like is that you are going to hold everything up so you can develop a scale for the questionnaire. Scaling is what I didn't like in marketing research. It is complicated and it takes too much time. Why can't you use some of the scales our marketing people have been using? Why do you have to reinvent the wheel?" The manager is looking toward Myra.

"Our research staff agrees with us that it would be inappropriate to adapt surveys developed for use in our consumer products line," says Myra smoothly.

"OK. Computers are not the same as toaster ovens and VCRs. Gotcha. Jason, what is going to be different about the scales you intend to develop?"

"When we held focus groups with your customers, they continually referred to the need for your product service to 'meet expectations' or 'exceed expectations.' The hundredth time we heard this we realized . . ."
This chapter covers procedures that will help you understand measurement scales so that you might select or design measures that are appropriate for your research. We concentrate here on the problems of measuring more complex constructs, like attitudes. Conceptually, we start this process by revisiting the research process (see Exhibit 12-1) to understand where the act of scaling fits in the process.

Scales in business research are generally constructed to measure behavior, knowledge, and attitudes. Attitude scales are among the most difficult to construct, so we will use attitudes to develop your understanding of scaling.

> The Nature of Attitudes

Jason is properly concerned about attitude measurement for the MindWriter study. But what is an attitude? There are numerous definitions, but one seems to capture the essence: An attitude is a learned, stable predisposition to respond to oneself, other persons, objects, or issues in a consistently favorable or unfavorable way. Important aspects of this definition include the learned nature of attitudes, their relative permanence, and their association with socially significant events and objects. Because an attitude is a predisposition, it would seem that the more favorable one's attitude is toward a product or service, the more likely that the product or service will be purchased. But, as we will see, that is not always the case.

Let's use Myra as an example to illustrate the nature of attitudes:

1. She is convinced that MindWriter has great talent, terrific products, and superior opportunities for growth.
2. She loves working at MindWriter.
3. She expects to stay with the firm and work hard to achieve rapid promotions for greater visibility and influence.

The first statement is an example of a cognitively based attitude. It represents Myra's memories, evaluations, and beliefs about the properties of the object. A belief is an estimate (probability) about the truth of something. In this case, it is the likelihood that the characteristics she attributes to her work environment are true. The statement "I think the cellular market will expand rapidly to incorporate radio and video" is also derived from cognition and belief. The second statement above is an affectively based attitude. It represents Myra's feelings, intuition, values, and emotions toward the object. "I love the Yankees" and "I hate corn flakes" are other examples of emotionally oriented attitudes. Finally, researchers recognize a third component, conative or behaviorally based attitudes. The concluding statement reflects Myra's expectations and behavioral intentions toward her firm and the instrumental behaviors necessary to achieve her future goals.

The Relationship between Attitudes and Behavior

The attitude-behavior relationship is not straightforward, although there may be close linkages. Attitudes and behavioral intentions do not always lead to actual behaviors; and although attitudes and behaviors are expected to be consistent with each other, that is not always the case. Moreover, behaviors can influence attitudes. For example, marketers know that a positive experience with a product or service reinforces a positive attitude or makes a customer question a negative attitude. This is one reason that restaurants where you have a bad dining experience may give you a coupon for a free meal on your next visit. They know a bad experience contributes mightily to formation of negative attitudes.

Business researchers treat attitudes as hypothetical constructs because of their complexity and the fact that they are inferred from the measurement data, not actually observed. These qualifications
cause researchers to be cautious about the ways certain aspects of measured attitudes predict behavior. Several factors have an effect on the applicability of attitudinal research:

- Specific attitudes are better predictors of behavior than general ones.
- Strong attitudes (strength is affected by accessibility or how well the object is remembered and brought to consciousness, how extreme the attitude is, or the degree of confidence in it) are better predictors of behavior than weak attitudes composed of little intensity or topical interest.
- Direct experiences with the attitude object (when the attitude is formed, during repeated exposure, or through reminders) produce behavior more reliably.
- Cognitive-based attitudes influence behaviors better than affective-based attitudes.
- Affective-based attitudes are often better predictors of consumption behaviors.
Using multiple measurements of attitude or several behavioral assessments across time and environments improves prediction.

The influence of reference groups (interpersonal support, urges of compliance, peer pressure) and the individual’s inclination to conform to these influences improves the attitude-behavior linkage.

Researchers measure and analyze attitudes because attitudes offer insights about behavior. Many of the attitude measurement scales used have been tested for reliability and validity, but often we craft unique scales that don’t share those standards. An example is an instrument that measures attitudes about a particular tourist attraction, product, or candidate, as well as the person’s intention to visit, buy, or vote. Neither the attitude nor the behavioral intent instrument, alone or together, is effective in predicting the person’s actual behavior if it has not been designed carefully. Nevertheless, managers know that the measurement of attitudes is important because attitudes reflect past experience and shape future behavior.

**Attitude Scaling**

Attitude scaling is the process of assessing an attitudinal disposition using a number that represents a person’s score on an attitudinal continuum ranging from an extremely favorable disposition to an extremely unfavorable one. Scaling is the “procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some of the characteristics of numbers to the properties in question.” Procedurally, we assign numbers to indicants of the properties of objects. Thus, one assigns a number scale to the various levels of heat and cold and calls it a thermometer. To measure the temperature of the air, you know that a property of temperature is that its variation leads to an expansion or contraction of mercury. A glass tube with mercury provides an indicant of temperature change by the rise or fall of the mercury in the tube. Similarly, your attitude toward your university could be measured on numerous scales that capture indicators of the different dimensions of your awareness, feelings, or behavioral intentions toward the school.

The percent of workers who are considered truly loyal.

34

> Selecting a Measurement Scale

Selecting and constructing a measurement scale requires the consideration of several factors that influence the reliability, validity, and practicality of the scale:

- Research objectives.
- Response types.
- Data properties.
- Number of dimensions.
- Balanced or unbalanced.
- Forced or unforced choices.
- Number of scale points.
- Rater errors.

**Research Objectives**

Researchers’ objectives are too numerous to list (including, but not limited to, studies of attitude, attitude change, persuasion, awareness, purchase intention, cognition and action, actual and repeat purchase). Researchers, however, face two general types of scaling objectives:

- To measure characteristics of the participants who participate in the study.
- To use participants as judges of the objects or indicants presented to them.
Assume you are conducting a study of customers concerning their attitudes toward a change in corporate identity (a company logo and peripherals). With the first study objective, your scale would measure the customers' orientation as favorable or unfavorable. You might combine each person's answers to form an indicator of overall orientation. The emphasis in this first study is on measuring attitudinal differences among people. With the second objective, you might use the same data, but you are now interested in how satisfied people are with different design options. Each participant is asked to choose the object he or she favors or the preferred solution. Participants judge which object has more of some characteristic or which design solution is closest to the company's stated objectives.

Response Types

Measurement scales fall into one of four general types: rating, ranking, categorization, and sorting. A rating scale is used when participants score an object or indicant without making a direct comparison to another object or attitude. For example, they may be asked to evaluate the styling of a new automobile on a 7-point rating scale. Ranking scales constrain the study participant to making comparisons and determining order among two or more properties (or their indicants) or objects. Participants may be asked to choose which one of a pair of cars has more attractive styling. A choice scale requires that participants choose one alternative over another. They could also be asked to rank-order the importance of comfort, ergonomics, performance, and price for the target vehicle. Categorization asks participants to put themselves or property indicants in groups or categories. Asking auto show attendees to identify their gender or ethnic background or to indicate whether a particular prototype design would appeal to a youthful or mature driver would require a category response strategy. Sorting requires that participants sort cards (representing concepts or constructs) into piles using criteria established by the researcher. The cards might contain photos or images or verbal statements of product features such as various descriptors of the car's performance.

Data Properties

Decisions about the choice of measurement scales are often made with regard to the data properties generated by each scale. In Chapter 11, we said that we classify scales in increasing order of power; scales are nominal, ordinal, interval, or ratio. Nominal scales classify data into categories without indicating order, distance, or unique origin. Ordinal data show relationships of more than and less than but have no distance or unique origin. Interval scales have both order and distance but no unique origin. Ratio scales possess all four properties' features. The assumptions underlying each level of scale determine how a particular measurement scale's data will be analyzed statistically.

Number of Dimensions

Measurement scales are either unidimensional or multidimensional. With a unidimensional scale, one seeks to measure only one attribute of the participant or object. One measure of an actor's star power is his or her ability to "carry" a movie. It is a single dimension. Several items may be used to measure this dimension and by combining them into a single measure, an agent may place clients along a linear continuum of star power. A multidimensional scale recognizes that an object might be better described with several dimensions than on a unidimensional continuum. The actor's star power variable might be better expressed by three distinct dimensions—ticket sales for last three movies, speed of attracting financial resources, and column-inch/amount-of-TV coverage of the last three films.

Balanced or Unbalanced

A balanced rating scale has an equal number of categories above and below the midpoint. Generally, rating scales should be balanced, with an equal number of favorable and unfavorable response choices. However, scales may be balanced with or without an indifference or midpoint option. A balanced scale
might take the form of “very good—good—average—poor—very poor.” An unbalanced rating scale has an unequal number of favorable and unfavorable response choices. An example of an unbalanced scale that has only one unfavorable descriptive term and four favorable terms is “poor—fair—good—very good—excellent.” The scale designer expects that the mean ratings will be near “good” and that there will be a symmetrical distribution of answers around that point, but the scale does not allow participants who are unfavorable to express the intensity of their attitude.

The use of an unbalanced rating scale can be justified in studies in which researchers know in advance that nearly all participants’ scores will lean in one direction or the other. Raters are inclined to score attitude objects higher if the objects are very familiar and if they are ego-involved. Brand-loyal customers are also expected to respond favorably. When researchers know that one side of the scale is not likely to be used, they try to achieve precision on the side that will most often receive the participant’s attention. Unbalanced scales are also considered when participants are known to be either “easy raters” or “hard raters.” An unbalanced scale can help compensate for the error of leniency created by such raters.

**Forced or Unforced Choices**

An unforced-choice rating scale provides participants with an opportunity to express no opinion when they are unable to make a choice among the alternatives offered. A forced-choice rating scale requires that participants select one of the offered alternatives. Researchers often exclude the response choice “no opinion,” “undecided,” “don’t know,” “uncertain,” or “neutral” when they know that most participants have an attitude on the topic. It is reasonable in this circumstance to constrain participants so that they focus on alternatives carefully and do not idly choose the middle position. However, when many participants are clearly undecided and the scale does not allow them to express their uncertainty, the forced-choice scale biases results. Researchers discover such bias when a larger percentage of participants express an attitude than did so in previous studies on the same issue. Some of this bias is attributable to participants providing meaningless responses or reacting to questions about which they have no attitudes (see Chapter 13). This affects the statistical measures of the mean and median, which shift toward the scale’s midpoint, making it difficult to discern attitudinal differences throughout the instrument. Understanding neutral answers is a challenge for researchers. In a customer satisfaction study that focused on the overall satisfaction question with a company in the electronics industry, an unforced scale was used. Study results, however, revealed that 75 percent of those in the “neutral” participant group could be converted to brand loyalists if the company excelled (received highly favorable ratings) on only 2 of the 26 other scaled questions in the study. Thus, the participants in the neutral group weren’t truly neutral, and a forced-choice scale would have revealed the desired information.

**Number of Scale Points**

What is the ideal number of points for a rating scale? Academics and practitioners often have dogmatic reactions to this question, but the answer is more practical: A scale should be appropriate for its purpose. For a scale to be useful, it should match the stimulus presented and extract information proportionate to the complexity of the attitude object, concept, or construct. A product that requires little effort or thought to purchase, is habitually bought, or has a benefit that fades quickly (low-involvement products) can be measured generally with a simple scale. A 3-point scale (better than average-average-worse than average) is probably sufficient for a deodorant, a fast-food burger, gift-wrapping, or a snack. There is little support for choosing a scale with 5 or more points in this instance. But when the product is complex, plays an important role in the consumer’s life, and is costly (e.g., financial services, luxury goods, automobiles, and other high-involvement products), a scale with 5 to 11 points should be considered.

As we noted in Chapter 11, the characteristics of reliability and validity are important factors affecting measurement decisions. First, as the number of scale points increases, the reliability of the measure increases. Second, in some studies, scales with 11 points may produce more valid results than 3-, 5-, or 7-point scales. Third, some constructs require greater measurement sensitivity and the opportunity to extract more variance, which additional scale points provide. Fourth, a larger number of
Online surveys are increasingly common due in large part to their speed in data collection. They also offer versatility for use with various types of measurement scales; flexibility in containing not only verbal but graphical, photographic, video, and digital elements; access to difficult-to-contact or inaccessible participants; and lower cost of large-sample completion. The visual appearance of the measurement scale is very important in getting the participant to click through to completion. This invitation from Nortel Networks and the opening screen of the questionnaire are designed to encourage participation. Informative, Inc., fielded this survey for Nortel Networks (designed to evaluate Nortel's website). The first screen of the questionnaire indicates two strategies: a multiple-choice, single-response strategy incorporating forced choice, and a multi-item rating grid that does not force choice (notice the NA column). If you look closely, you can also see a scroll bar on the first screen. Some designers will put only one question to a screen in Web questionnaires believing that participants who have to scroll may not fully complete the survey. This survey was designed for a technical audience, so that was not as much a concern. www.nortelnetworks.com

A 10-point scale is needed to produce accuracy when using single-dimension versus multiple-dimension scales. Finally, in cross-cultural measurement, the cultural practices may condition participants to a standard metric—a 10-point scale in Italy, for example. Exhibit 12-2 previews the scales discussed in this chapter along with the characteristics of scale types described in Chapter 11.

Rater Errors

The value of rating scales depends on the assumption that a person can and will make good judgments. Before accepting participants' ratings, we should consider their tendencies to make errors of central tendency and halo effect. Some raters are reluctant to give extreme judgments, and this fact accounts for the error of central tendency. Participants may also be "easy raters" or "hard raters," making what is called an error of leniency. These errors most often occur when the rater does not know the object or property being rated. To address these tendencies, researchers can:

- Adjust the strength of descriptive adjectives.
- Space the intermediate descriptive phrases farther apart.
- Provide smaller differences in meaning between the steps near the ends of the scale than between the steps near the center.
- Use more points in the scale.
The *halo effect* is the systematic bias that the rater introduces by carrying over a generalized impression of the subject from one rating to another. An instructor expects the student who does well on the first question of an examination to do well on the second. You conclude a report is good because you like its form, or you believe someone is intelligent because you agree with him or her. Halo is especially difficult to avoid when the property being studied is not clearly defined, is not easily observed, is not frequently discussed, involves reactions with others, or is a trait of high moral importance. Ways of counteracting the halo effect include having the participant rate one trait at a time, revealing one trait per page (as in an Internet survey, where the participant cannot return to change his or her answer), or periodically reversing the terms that anchor the endpoints of the scale, so positive attributes are not always on the same end of each scale.

**Rating Scales**

In Chapter 11, we said that questioning is a widely used stimulus for measuring concepts and constructs. For example, a researcher asks questions about participants' attitudes toward the taste of a soft drink. The responses are "thirst quenching," "sour," "strong bubbly," "orange taste," and "syrupy." These answers alone do not provide a means of discerning the degree of favorability and thus would be of limited value to the researcher. However, with a properly constructed scale, the researcher could develop a taste profile for the target brand. We use rating scales to judge properties of objects without reference to other similar objects. These ratings may be in such forms as "like—dislike," "approve—disapprove," or other classifications using even more categories.

Examples of rating scales we discuss in this section are shown in Exhibit 12-2. Since this exhibit amplifies the overview presented in this section, we will refer you to the exhibit frequently.

**Simple Attitude Scales**

The *simple category scale* (also called a *dichotomous scale*) offers two mutually exclusive response choices. In Exhibit 12-3 they are "yes" and "no," but they could just as easily be "important" and "unimportant," "agree" and "disagree," or another set of discrete categories if the question were different. This response strategy is particularly useful for demographic questions or where a dichotomous response is adequate.

When there are multiple options for the rater but only one answer is sought, the *multiple-choice, single-response scale* is appropriate. Our example has five options. The primary alternatives should encompass 90 percent of the range, with the "other" category completing the participant's list. When
there is no possibility for an "other" response or exhaustiveness of categories is not critical, the "other" response may be omitted. Both the multiple-choice, single-response scale and the simple category scale produce nominal data.

A variation, the multiple-choice, multiple-response scale (also called a checklist), allows the rater to select one or several alternatives. In the example in Exhibit 12-3, we are measuring seven items with one question, and it is possible that all seven sources for home design were consulted. The cumulative feature of this scale can be beneficial when a complete picture of the participant's choice is desired, but it may also present a problem for reporting when research sponsors expect the responses to sum to 100 percent. This scale generates nominal data.

Simple attitude scales are easy to develop, are inexpensive, and can be designed to be highly specific. They provide useful information and are adequate if developed skillfully. There are also weaknesses. The design approach is subjective. The researcher's insight and ability offer the only assurance that the items chosen are a representative sample of the universe of attitudes about the attitude object. We have no evidence that each person will "view all items with the same frame of reference as will other people. Although such scales are frequently used, there has been a great effort to develop construction techniques that overcome some of their deficiencies.

**Likert Scales**

The **Likert scale**, developed by Rensis Likert (pronounced Lick-ert), is the most frequently used variation of the summated rating scale. **Summated rating scales** consist of statements that express either a favorable or an unfavorable attitude toward the object of interest. The participant is asked to agree or disagree with each statement. Each response is given a numerical score to reflect its degree of attitudinal favorableness, and the scores may be summed to measure the participant's overall attitude. Summation is not necessary and in some instances may actually be misleading, as our caution below clearly shows.

---

> **Exhibit 12-3** Sample Rating Scales

<table>
<thead>
<tr>
<th>Simple Category Scale</th>
<th>&quot;I plan to purchase a MindWriter laptop in the next 12 months.&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>data: nominal</td>
<td>[ ] Yes</td>
</tr>
<tr>
<td></td>
<td>[ ] No</td>
</tr>
</tbody>
</table>

| Multiple-Choice,     | "What newspaper do you read most often for financial news?"   |
| Single-Response Scale|                                                               |
| data: nominal        | [ ] East City Gazette                                         |
|                       | [ ] West City Tribune                                         |
|                       | [ ] Regional newspaper                                        |
|                       | [ ] National newspaper                                        |
|                       | [ ] Other (specify:_________________________________________)|

| Multiple-Choice,     | "Check any of the sources you consulted when designing your new home:" |
| Multiple-Response    |                                                               |
| Scale (checklist)    |                                                               |
| data: nominal        | [ ] Online planning services                                 |
|                       | [ ] Magazines                                                 |
|                       | [ ] Independent contractor/builder                             |
|                       | [ ] Developer's models/plans                                  |
|                       | [ ] Designer                                                  |
|                       | [ ] Architect                                                 |
|                       | [ ] Other (specify:_________________________________________)|

| Likert Scale,        | "The Internet is superior to traditional libraries for comprehensive searches." |
| Summated Rating Scale|                                                               |
| data: interval       | STRONGLY AGREE (5) AGREE (4) NEITHER AGREE NOR DISAGREE (3) DISAGREE (2) STRONGLY DISAGREE (1) |
In Chapter 11 we noted that researchers differ in the ways they treat data from certain scales. If you are unable to establish the linearity of the measured variables or you cannot be confident that you have equal intervals, it is proper to treat data from these scales as ordinal.
Campbell-Ewald, an award-winning integrated communications agency headquartered in Detroit, believes it is good business to treat customers with respect—and the agency can prove it. As part of a major research initiative to discover why customer relationship management (CRM) solutions were falling short of expectations, Campbell-Ewald mailed more than 5,000 surveys to adults 18 or older who were customers in each of three business sectors: insurance, automotive, and retail. The goal? To answer the question “Does respect influence customer loyalty and, thereby, purchasing?” With partner research company Synovate and Campbell-Ewald clients, three surveys were developed. Each included 27 to 29 attitudinal statements that queried the adults on how they defined respect and the importance of respect to purchase behavior in each sector. Customers responded to the statements using a 5-point scale (strongly agree to strongly disagree). Using analysis of the results, Campbell-Ewald validated the relevance of its five “People Principles,” which, in turn, have helped clients like General Motors, Continental Airlines, and Farmers Insurance incorporate respectful behavior into their business practices. The five “People Principles” of respect are:

- Appreciate me.
- Intentions don’t matter; actions do.
- Listen; then you’ll know what I said.
- It’s about me, not about you.
- Admit it—you goofed!

How would you operationalize the construct of respect?

To learn more about this research, read the case: “Campbell-Ewald: R-E-S-P-E-C-T Spells Loyalty.”

www.campbell-ewald.com; www.synovate.com

In Exhibit 12-3, the participant chooses one of five levels of agreement. This is the traditional Likert scale because it meets Likert’s rules for construction and testing. The numbers indicate the value to be assigned to each possible answer, with 1 the least favorable impression of Internet superiority and 5 the most favorable. Likert scales also use 7 and 9 scale points. Technically, this is known as a Likert-type scale since its construction is often less rigorous. However, the advantages of the 7- and 9-point scales are a better approximation of a normal response curve and extraction of more variability among respondents. The values for each choice are normally not printed on the instrument, but they are shown in Exhibit 12-4 to illustrate the scoring system.

The Likert scale has many advantages that account for its popularity. It is easy and quick to construct. Conscientious researchers are careful that each item meets an empirical test for discriminating ability between favorable and unfavorable attitudes. Likert scales are probably more reliable and provide a greater volume of data than many other scales. The scale produces interval data.

Originally, creating a Likert scale involved a procedure known as item analysis. In the first step, a large number of statements were collected that met two criteria: (1) Each statement was relevant to the attitude being studied; (2) each was believed to reflect a favorable or unfavorable position on that attitude. People similar to those who are going to be studied were asked to read each statement and to state the level of their agreement with it, using a 5-point scale. A scale value of 1 indicated a strongly unfavorable attitude (strongly disagree). The other intensities were 2 (disagree), 3 (neither agree nor disagree), 4 (agree), and 5 (strongly agree), a strongly favorable attitude (see Exhibit 12-3). To ensure consistent results, the assigned numerical values are reversed if the statement is worded negatively (1 is always strongly unfavorable and 5 is always strongly favorable). Each person’s responses are then added to secure a total score. The next step is to array these total scores and select some portion representing the highest and lowest total scores (generally defined as the top and bottom 10 to 25 percent of the distribution). The middle group (50 to 80 percent of participants) are excluded from the subsequent analysis.

The two extreme groups represent people with the most favorable and least favorable attitudes toward the attitude being studied. These extremes are the two criterion groups by which individual items are evaluated. Item analysis assesses each item based on how well it discriminates between those persons whose total score is high and those whose total score is low. It involves calculating the mean scores for each scale item among the low scorers and high scorers. The mean scores for the high-score
Exhibit 12-4 Evaluating a Scale Statement by Item Analysis

For the statement "My digital camera's features are exciting," we select the data from the bottom 25 percent of the distribution (low total score group) and the top 25 percent (high total score group). There are 73 people in each group. The remaining 50 percent of the middle of the distribution is not considered for this analysis.

<table>
<thead>
<tr>
<th>Response Categories</th>
<th>Low Total Score Group</th>
<th>High Total Score Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>f</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Agree</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Undecided</td>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>Disagree</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>177</td>
</tr>
</tbody>
</table>

\[
\bar{X}_L = \frac{177}{73} = 2.42
\]

\[
\Sigma(X_L - \bar{X}_L)^2 = 503 - \frac{(177)^2}{73}
\]

\[
= 73.84
\]

\[
t = \frac{\bar{X}_H - \bar{X}_L}{\sqrt{\frac{\Sigma(X_L - \bar{X}_L)^2 + \Sigma(X_H - \bar{X}_H)^2}{n(n-1)}}}
\]

\[
= \frac{3.90 - 2.42}{\sqrt{70.33 + 73.84}}
\]

\[
= 6.92
\]

Legend:
1. For each of the response categories, the scale's value (X) is multiplied by the frequency or number of participants (f) who chose that value. These values produce the product (fX). This number is then multiplied by X. For example, there are 3 participants in the low-score group who scored a 5 (strongly agree with the statement): (fX) = (3)(5) = 15; (X( )fX) = (5)(15) = 75.
2. The frequencies and products are summed.
3. A mean score for each group is computed.
4. Deviation scores are computed, squared, and summed as required by the formula.
5. The data are tested in a modified t-test that compares high- and low-scoring groups for the item. Notice the mean scores in the numerator of the formula.
6. The calculated value is compared with a criterion, 1.75. If the calculated value (in this case, 6.92) is equal to or exceeds the criterion, the statement is said to be a good discriminator of the measured attitude. (If it is less than the criterion, we would consider it a poor discriminator of the target attitude and delete it from the measuring instrument.) We then select the next measurement item and repeat the process.

and low-score groups are then tested for statistical significance by computing t values. (In evaluating response patterns of the high and low groups to the statement "My digital camera's features are exciting," we secure the results shown in Exhibit 12-4.) After finding the t values for each statement, they are rank-ordered, and those statements with the highest t values are selected. The 20 to 25 items that have the highest t values (statistically significant differences between mean scores) are selected for inclusion in the final scale. Researchers have found that a larger number of items for each attitude object improve the reliability of the scale. As an approximate indicator of a statement's discrimination power,
one authority also suggests using only those statements whose t value is 1.75 or greater, provided there are 25 or more subjects in each group.15

Although item analysis is helpful in weeding out attitudinal statements that do not discriminate well, the summation procedure causes problems for researchers. The following example on website banner ads shows that the same summated score can mean different things:

1. This banner ad provides the relevant information I expect.
2. I would bookmark this site to use in the future.
3. This banner ad is annoying.
4. I would click for deeper links to discover more details.

If a 5-point scale is used, the maximum favorable score would be 20 (assuming 5 is assigned to the strongly agree response and question 3, a negation, is reverse-scored). Approximately one-half of the statements are worded favorably and the other half unfavorably to safeguard against halo effects. The problem of summation arises because different patterns are concealed by the same total score. One participant could find the website's ad relevant, worth returning to, and somewhat pleasing but not desire deeper information, whereas another could find the ad annoying but have favorable attitudes on the other three questions, thereby producing the same total score.

**Semantic Differential Scales**

The semantic differential (SD) scale measures the psychological meanings of an attitude object using bipolar adjectives. Researchers use this scale for studies such as brand and institutional image. The method consists of a set of bipolar rating scales, usually with 7 points, by which one or more participants rate one or more concepts on each scale item. The SD scale is based on the proposition that an object can have several dimensions of connotative meaning. The meanings are located in multidimensional property space, called semantic space. Connotative meanings are suggested or implied meanings, in addition to the explicit meaning of an object. For example, a roaring fire in a fireplace may connote romantic as well as its more explicit meaning of burning flammable material within a brick kiln. One restaurant trying to attract patrons on slow Tuesday evenings offered a special Tuesday menu and called it “down home cooking.” Yankee pot roast, stew, and chicken pot pie, although not its usual cuisine, carried the connotative meaning of comfort foods and brought patrons into the restaurant, making Tuesday one of the busiest nights of the week. Advertisers, salespeople, and product and package designers have long known that they must use words, shapes, associations, and images to activate a person's connotative meanings.

Osgood and his associates developed the semantic differential method to measure the psychological meanings of an object to an individual.16 They produced a list of 289 bipolar adjective pairs, which were reduced to 76 pairs and formed into rating scales for attitude research. Their analysis allowed them to conclude that semantic space is multidimensional rather than unidimensional. Three factors contributed most to meaningful judgments by participants: (1) evaluation, (2) potency, and (3) activity. These concepts from the historical thesaurus study (Exhibit 12-5) illustrate the wide applicability of the technique to persons, abstract concepts, events, institutions, and physical objects.17

Researchers have followed a somewhat different approach to SD scales than did the original study advocates. They have developed their own adjectives or phrases and have focused on the evaluative dimension more often (which might help explain the popularity of the Likert scale). The positive benefit is that the scales created have been adapted to specific management questions. One study explored a retail store image using 35 pairs of words or phrases classified into eight groups. These word pairs were especially created for the study. Excerpts from this scale are presented in Exhibit 12-6. Other categories of scale items were “general characteristics of the company,” “physical characteristics of the store,” “prices charged by the store,” “store personnel,” “advertising by the store,” and “your friends and the store.” Since the scale items are closely associated with the characteristics of the store and its use, one could develop image profiles of various stores.

The semantic differential has several advantages. It is an efficient and easy way to secure attitudes from a large sample. These attitudes may be measured in both direction and intensity. The total set of responses provides a comprehensive picture of the meaning of an object and a measure of the person's connotative meanings.

**Results of the thesaurus study are shown in Exhibit 12-5.**
Exhibit 12-5 Results of the Thesaurus Study

<table>
<thead>
<tr>
<th>Evaluation (E)</th>
<th>Potency (P)</th>
<th>Activity (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good-bad</td>
<td>Hard-soft</td>
<td>Active-passive</td>
</tr>
<tr>
<td>Positive-negative</td>
<td>Strong-weak</td>
<td>Fast-slow</td>
</tr>
<tr>
<td>Optimistic-pessimistic</td>
<td>Heavy-light</td>
<td>Hot-cold</td>
</tr>
<tr>
<td>Complete-incomplete</td>
<td>Masculine-feminine</td>
<td>Excitable-calm</td>
</tr>
<tr>
<td>Timely-untimely</td>
<td>Severe-lenient</td>
<td></td>
</tr>
<tr>
<td>Tenacious-yielding</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exhibit 12-6 Adapting SD Scales for Retail Store Image Study

<table>
<thead>
<tr>
<th>Convenience of Reaching the Store from Your Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearby</td>
</tr>
<tr>
<td>Short time required to reach store</td>
</tr>
<tr>
<td>Difficult drive</td>
</tr>
<tr>
<td>Difficult to find parking place</td>
</tr>
<tr>
<td>Convenient to other stores I shop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Products Offered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wide selection of different kinds of products</td>
</tr>
<tr>
<td>Fully stocked</td>
</tr>
<tr>
<td>Undependable products</td>
</tr>
<tr>
<td>High quality</td>
</tr>
<tr>
<td>Numerous brands</td>
</tr>
<tr>
<td>Unknown brands</td>
</tr>
</tbody>
</table>

Exhibit 12-7 Steps in Constructing an SD Scale

1. Select the concepts: nouns, noun phrases, or nonverbal stimuli such as visual sketches. Concepts are chosen by judgment and reflect the nature of the investigative question. In the MindWriter study, one concept might be “Call Center accessibility.”

2. Select bipolar word pairs or phrase pairs appropriate to your needs. If the traditional Osgood adjectives are used, several criteria guide your selection:
   - Three bipolar pairs are required when using evaluation, potency, and activity. Scores on these individual items can be averaged, by factor, to improve their reliability.
   - The scale should be relevant to the concepts being judged. Choose adjectives that allow connotative perceptions to be expressed. Irrelevant concept-scale pairings yield neutral midpoint values that convey little information.
   - Scales should be stable across raters and concepts. A pair such as “large-small” may be interpreted as denotative when judging a physical object such as “automobile” but may be used connotatively with abstract concepts such as “product quality.”
   - Scales should be linear between polar opposites and pass through the origin. A pair that fails this test is “rugged–delicate,” which is nonlinear on the evaluation dimension. When used separately, both adjectives have favorable meanings.

3. Create the scoring system and assign a weight to each point on the scale. The negative signs in the original scoring procedure (−3, −2, −1, 0, +1, +2, +3) were found to produce coding errors, and the 0 point is arbitrary. Most SD scales have 7 points: 7, 6, 5, 4, 3, 2, and 1.

4. As with Likert scales, about half of the adjective pairs are randomly reversed to minimize the halo effect.


Exhibit 12-8 SD Scale for Analyzing Industry Association Candidates

<table>
<thead>
<tr>
<th>Analyze (candidate) for current position:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E) Sociable (7): __ __ __ __ __ __ __ (1) Unsociable</td>
</tr>
<tr>
<td>(F) Weak (1): __ __ __ __ __ __ __ (7) Strong</td>
</tr>
<tr>
<td>(A) Active (7): __ __ __ __ __ __ __ (1) Passive</td>
</tr>
<tr>
<td>(E) Progressive (7): __ __ __ __ __ __ __ (1) Regressive</td>
</tr>
<tr>
<td>(F) Yielding (1): __ __ __ __ __ __ __ (7) Tenacious</td>
</tr>
<tr>
<td>(A) Slow (1): __ __ __ __ __ __ __ (7) Fast</td>
</tr>
<tr>
<td>(E) True (7): __ __ __ __ __ __ __ (1) False</td>
</tr>
<tr>
<td>(F) Heavy (7): __ __ __ __ __ __ __ (1) Light</td>
</tr>
<tr>
<td>(A) Hot (7): __ __ __ __ __ __ __ (1) Cold</td>
</tr>
<tr>
<td>(E) Unsuccessful (1): __ __ __ __ __ __ __ (7) Successful</td>
</tr>
</tbody>
</table>

Doing the rating. It is a standardized technique that is easily repeated but escapes many problems of response distortion found with more direct methods. It produces interval data. Basic instructions for constructing an SD scale are found in Exhibit 12-7.

In Exhibit 12-8 we see a scale being used by a panel of corporate leaders evaluating candidates for a high-level position in their industry’s lobbying association. The selection of the concepts is driven by the characteristics they believe the candidate must possess to be successful in advancing their agenda. There are three candidates.

Based on the panel’s requirements, we choose 10 scales to score the candidates. The letters along the left side, which show the relevant attitude dimension, would be omitted from the actual scale, as would the numerical values shown. Note that the evaluation, potency, and activity scales are mixed.
To analyze the results, the set of evaluation (E) values is averaged, as are those for the potency (P) and activity (A) dimensions.

The data are plotted in a “snake diagram” in Exhibit 12-9. Here the adjective pairs are reordered so that evaluation, potency, and activity descriptors are grouped together, with the ideal factor reflected by the left side of the scale.

### Numerical/Multiple Rating List Scales

**Numerical scales** have equal intervals that separate their numeric scale points, as shown in Exhibit 12-3. The verbal anchors serve as the labels for the extreme points. Numerical scales are often 5-point scales but may have 7 or 10 points. The participants write a number from the scale next to each item. If numerous questions about a product’s performance were included in the example, the scale would provide both an absolute measure of importance and a relative measure (ranking) of the various items rated. The scale’s linearity, simplicity, and production of ordinal or interval data make it popular for managers and researchers. When evaluating a new product concept, purchase intent is frequently measured with a 5- to 7-point numerical scale, with the anchors being “definitely would buy” and “definitely would not buy.”

A **multiple rating list scale** (Exhibit 12-3) is similar to the numerical scale but differs in two ways: (1) It accepts a circled response from the rater, and (2) the layout facilitates visualization of the results. The advantage is that a mental map of the participant’s evaluations is evident to both the rater and the researcher. This scale produces interval data.

### Stapel Scales

The **Stapel scale** is used as an alternative to the semantic differential, especially when it is difficult to find bipolar adjectives that match the investigative question. In the example in Exhibit 12-3 there are three attributes of corporate image. The scale is composed of the word (or phrase) identifying the image dimension and a set of 10 response categories for each of the three attributes.

Fewer response categories are sometimes used. Participants select a plus number for the characteristic that describes the attitude object. The more accurate the description, the larger is the positive number. Similarly, the less accurate the description, the larger is the negative number chosen.
Ratings range from +5 to −5, with participants selecting a number that describes the store very accurately to very inaccurately. Like the Likert, SD, and numerical scales, Stapel scales usually produce interval data.

### Constant-Sum Scales

A scale that helps the researcher discover proportions is the **constant-sum scale**. With a constant-sum scale, the participant allocates points to more than one attribute or property indicant, such that they total a constant sum, usually 100 or 10. In the Exhibit 12-3 example, two categories are presented that must sum to 100. In the restaurant example, the participant distributes 100 points among four categories:

You have 100 points to distribute among the following characteristics of the Dallas Steakhouse.

Indicate the relative importance of each attribute:

- Food Quality
- Atmosphere
- Service
- Price

100 TOTAL

Up to 10 categories may be used, but both participant precision and patience suffer when too many stimuli are proportioned and summed. A participant’s ability to add is also taxed in some situations; thus, this is not a response strategy that can be effectively used with children or the uneducated. The advantage of the scale is its compatibility with percent (100 percent) and the fact that alternatives that are perceived to be equal can be so scored—unlike the case with most ranking scales. The scale is used to record attitudes, behavior, and behavioral intent. The constant-sum scale produces interval data.

### Graphic Rating Scales

The **graphic rating scale** was originally created to enable researchers to discern fine differences. Theoretically, an infinite number of ratings are possible if participants are sophisticated enough to differentiate and record them. They are instructed to mark their response at any point along a continuum. Usually, the score is a measure of length (millimeters) from either endpoint. The results are treated as interval data. The difficulty is in coding and analysis. This scale requires more time than scales with predetermined categories.

Never ______ X ______ Always

Other graphic rating scales (see Exhibit 12-3) use pictures, icons, or other visuals to communicate with the rater and represent a variety of data types. Graphic scales are often used with children, whose more limited vocabulary prevents the use of scales anchored with words.

### Ranking Scales

In ranking scales, the participant directly compares two or more objects and makes choices among them. Frequently, the participant is asked to select one as the “best” or the “most preferred.” When there are only two choices, this approach is satisfactory, but it often results in ties when more than two choices are found. For example, assume participants are asked to select the most preferred among three or more models of a product. In response, 40 percent choose model A, 30 percent choose model B, and 30 percent choose model C. Which is the preferred model? The analyst would be taking a risk to suggest that A is most preferred. Perhaps that interpretation is correct, but 60 percent of the participants chose some model other than A. Perhaps all B and C voters would place A last, preferring either B or C to A. This ambiguity can be avoided by using some of the techniques described in this section.
Assume you are asked by Galaxy Department Stores to study the shopping habits and preferences of teen girls. Galaxy is seeking a way to compete with specialty stores that are far more successful in serving this market segment. Galaxy is considering the construction of an instore boutique catering to these teens. What measurement issues would determine your construction of measurement scales?

Using the paired-comparison scale, the participant can express attitudes unambiguously by choosing between two objects. Typical of paired comparisons would be the sports car preference example in Exhibit 12-10. The number of judgments required in a paired comparison is \((n)(n - 1)/2\), where \(n\) is the number of stimuli or objects to be judged. When four cars are evaluated, the participant evaluates six paired comparisons \(((4)(3)/2 = 6)\).

In another example we might compare packaging design proposals considered by a brand manager (see Exhibit 12-11). Generally, there are more than two stimuli to judge, resulting in a potentially tedious task for participants. If 15 suggestions for design proposals are available, 105 paired comparisons would be made.

Reducing the number of comparisons per participant without reducing the number of objects can lighten this burden. You can present each participant with only a sample of the stimuli. In this way, each pair of objects must be compared an equal number of times. Another procedure is to choose a few objects that are believed to cover the range of attractiveness at
Exhibit 12-11  Response Patterns of 200 Heavy Users' Paired Comparisons on Five Alternative Package Designs

Paired-comparison data may be treated in several ways. If there is substantial consistency, we will find that if A is preferred to B, and B to C, then A will be consistently preferred to C. This condition of transitivity need not always be true but should occur most of the time. When it does, take the total number of preferences among the comparisons as the score for that stimulus. Assume a manager is considering five distinct packaging designs. She would like to know how heavy users would rank these designs. One option would be to ask a sample of the heavy-users segment to pair-compare the packaging designs. With a rough comparison of the total preferences for each option, it is apparent that B is the most popular.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Total</th>
<th>Rank order</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>-164*</td>
<td>-</td>
<td>136</td>
<td>50</td>
<td>70</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>36</td>
<td></td>
<td>-54</td>
<td>14</td>
<td>30</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>62</td>
<td>146</td>
<td></td>
<td>-32</td>
<td>-</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>150</td>
<td>186</td>
<td>108</td>
<td></td>
<td>-</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>130</td>
<td>170</td>
<td>150</td>
<td>82</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>378</td>
<td>686</td>
<td>510</td>
<td>178</td>
<td>268</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank order</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Interpret this cell as 164 of 200 customers preferred suggested design B (column) to design A (row).

Pair comparisons run the risk that participants will tire to the point that they give ill-considered answers or refuse to continue. Opinions differ about the upper limit, but five or six stimuli are not unreasonable when the participant has other questions to answer. If the data collection consists only of paired comparisons, as many as 10 stimuli are reasonable. A paired comparison provides ordinal data.

The forced ranking scale, shown in Exhibit 12-10, lists attributes that are ranked relative to each other. This method is faster than paired comparisons and is usually easier and more motivating to the participant. With five items, it takes 10 paired comparisons to complete the task, and the simple forced ranking of five is easier. Also, ranking has no transitivity problem where A is preferred to B, and B to C, but C is preferred to A—although it also forces a false unidimensionality.

A drawback to forced ranking is the number of stimuli that can be handled by this method. Five objects can be ranked easily, but participants may grow careless in ranking 10 or more items. In addition, rank ordering produces ordinal data since the distance between preferences is unknown.

Often the manager is interested in benchmarking. This calls for a standard by which other programs, processes, brands, point-of-sale promotions, or people can be compared. The comparative scale is ideal for such comparisons if the participants are familiar with the standard. In the Exhibit 12-10 example, the standard is the participant's previous hair dryer. The new dryer is being assessed relative to it. The provision to compare yet other dryers to the standard is not shown in the example but is nonetheless available to the researcher.

Some researchers treat the data produced by comparative scales as interval data since the scoring reflects an interval between the standard and what is being compared. We would treat the rank or position of the item as ordinal data unless the linearity of the variables in question could be supported.
Should Northwest Airlines, Marriott, or Alaskan Airlines attempt to attract the business of Americans with disabilities? If so, what would it take to capture the segment? Eric Lipp, executive director of the Open Doors Organization (ODO), an advocacy organization for those with disabilities, sponsored a study to find out. High on his agenda was providing an incentive to the travel industry to make accommodations to attract the 22 million adults with disabilities who have traveled in the last two years on 63 million trips—and who may want to travel more. "We now estimate that Americans with disabilities currently spent $13.2 billion in travel expenditures and that amount would at least double [to $27.2 billion] if travel businesses were more attuned to the needs of those with disabilities."

ODO hired Harris Interactive, a global market research and consulting firm best known for The Harris Poll and for pioneering the Internet method to conduct scientifically accurate market research. Harris Interactive conducted a hybrid study via both online and phone surveys to determine the magnitude of the disability travel segment, its purchasing power, and the accommodations the segment needed to increase travel. "Those with disabilities can't all be reached with one method," explained Laura Light, project director with Harris Interactive. "The nature of their physical limitation might preclude one method or the other."

And how did the firm evaluate all the possible accommodations—from Braille safety cards on airplanes to a designated person to handle problems in a hotel? Harris Interactive used its proprietary COMPASS™ methodology, which uses paired comparisons as a measurement tool. "COMPASS™ saves the participant time and energy," explained Light. "Even with a long list, COMPASS™ can be done quickly." In the ODO study, COMPASS™ was used twice: once to measure 17 possible airline accommodations and once to measure 23 possible hotel accommodations. By having each participant evaluate only a portion of the large number of accommodation pairs rather than the full list (138 for airline accommodations and 253 for hotel accommodations), each question was answered in under four minutes. By using this process with all members of the sample, Harris Interactive is able to rank-order the items and measure the magnitude of difference between items. This makes it easier for Delta, Marriott, or Alaskan Airlines to make the right choices about accommodations for those with disabilities.

www.opendoorsnfp.org; www.harrisinteractive.com

To learn more about this research, read the case "Open Doors: Extending Hospitality to Travelers with Disabilities."

---

**Paired Comparison Increases Hospitality**

Q-sorts require sorting of a deck of cards into piles that represent points along a continuum. The participant—or judge—groups the cards based on his or her response to the concept written on the card. Researchers using Q-sort resolve three special problems: item selection, structured or unstructured choices in sorting, and data analysis. The basic Q-sort procedure involves the selection of a set of verbal statements, phrases, single words, or photos related to the concept being studied. For statistical stability, the number of cards should not be less than 60; and for convenience, not be more than 120. After the cards are created, they are shuffled, and the participant is instructed to sort the cards into a set of piles (usually 7 to 11), each pile representing a point on the judgment continuum. The left-most pile represents the concept statements, which are "most valuable," "favorable," "agreeable," and so forth. The right-most pile contains the least favorable cards. The researcher asks the participant to fill the center, or neutral, pile with the cards about which the participant is indecisive. In the case of a structured sort, the distribution of cards allowed in each pile is predetermined. With an unstructured sort, only the number of piles will be determined. Although the distribution of cards in most structured sorts resembles a normal distribution, there is some controversy about analyzing the data as ranking (ordinal data) versus interval data.

The purpose of sorting is to get a conceptual representation of the sorter's attitude toward the attitude object and to compare the relationships between people. The relative ranking of concepts allows researchers to derive clusters of individuals possessing similar preferences. By researchers varying
Jason has been working on scaling for MindWriter's CompleteCare project for a week when the request comes to Myra Wires, to report her progress to MindWriter's general manager. He has narrowed the choice to the three scales in Exhibit 12-12: a Likert scale, a numerical rating scale with two verbal anchors, and a hybrid expectation scale. All are 5-point scales that are presumed to measure at the interval level.

He needs a statement that can accompany the scale for preliminary evaluation. Returning to their list of investigative questions, he finds a question that seems to capture the essence of the repair process: Are customers' problems resolved? Translated into an assertion for the scale, the statement becomes "Resolution of problem that prompted service/repair." He continues to labor over the wording of the verbal anchors. Appropriate versions of the investigative question are constructed, and then the scales are added.

After consulting with MindWriter's research staff, Myra and Jason discuss the advantages of the scales. Myra suggests it is unlikely that CompleteCare would meet none of the customers' expectations. And, with errors of leniency, "none" should be replaced by the term "few" so that the low end of the scale would be more relevant. Jason has read a Marketing News article that said Likert scales and scales similar to MindWriter's numerical scale frequently produced a heavy concentration of 4s and 5s—a common problem in customer satisfaction research.

Others suggest a 7-point scale to remedy this, but Jason thinks the term "exceeded" on the expectation scale could compensate for scores that cluster on the positive end.

Now ready for a pilot test, Jason decides to compare the expectation scale with MindWriter's numerical rating scale. (The Likert scale requires that they create more potential items than they had room for on the postcard.) Using the CompleteCare database, names, addresses, and phone numbers are selected, and 30 customers are chosen at random from those who have had recent service. Jason chooses the delayed version forms method for reliability testing (see Chapter 11). Jason administers the expectation scale followed by the numerical scale to half of the participants and the numerical scale followed by the expectation scale to the other half. Each half sample experiences a time delay. No "order-of-presentation" affects are found. Subsequently, Jason correlates the numerical satisfaction scores with the expectation scores and plots the results, shown in Exhibit 12-13.

On the problem resolution question, the participants' scores from the satisfaction scale and those from the expectation scale...
The Sources and Collection of Data

The correlation index is .90 (1.00 is a perfect positive correlation). This reveals the equivalence of both scales. On another reliability issue, stability, a check of the test-retest reliability over a one-week interval produces a correlation index of .93 for the expectation scale, but the MindWriter satisfaction scale has a lower index ($r = .75$). This implies that the expectation scale is more likely to produce stable and consistent results. Finally, as shown in the plot, the data for the resolution question measured on both scales are linear (they cluster around a straight line). It is logical to conclude that the expectation scale, which was originally thought to have advantages, performs well in initial tests and is a good replacement for MindWriter’s existing scale.

The decision is made. They will use the new hybrid expectation scale for the MindWriter research on the CompleteCare service program.

The instructions, the technique can be used to describe products, services, behavioral intentions, and a host of other applications. In the example below, participants are asked to complete a structured sort of cards containing the names of magazines. The scale values and the number of cards in each pile are predetermined, although the distribution in this case represents a normal statistical distribution.

What magazines do you want Singapore Airlines to carry for its in-flight service?

<table>
<thead>
<tr>
<th>Most Preferred</th>
<th>Least Preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 9 8 7 6 5 4 3 2 1 0</td>
<td>(scale value)</td>
</tr>
<tr>
<td>3 4 7 10 13 16 13 10 7 4 3</td>
<td>(number of cards per pile)</td>
</tr>
</tbody>
</table>

Cumulative Scales

Total scores on cumulative scales have the same meaning. Given a person’s total score, it is possible to estimate which items were answered positively and negatively. A pioneering scale of this type was the scalogram. Scalogram analysis is a procedure for determining whether a set of items forms a unidimensional scale. A scale is unidimensional if the responses fall into a pattern in which endorsement of the item reflecting the extreme position results in endorsing all items that are less extreme.

Assume we are surveying opinions regarding a new style of running shoe. We have developed a preference scale of four items:

1. The Airsole is good-looking.
2. I will insist on Airsole next time because it is great-looking.
3. The appearance of Airsole is acceptable to me.
4. I prefer the Airsole style to other styles.

Participants indicate whether they agree or disagree. If these items form a unidimensional scale, the response patterns will approach the ideal configuration shown in Exhibit 12-14. Item 2 is the most
Exhibit 12-14  Ideal Scalogram Response Pattern

<table>
<thead>
<tr>
<th>Item</th>
<th>2</th>
<th>4</th>
<th>1</th>
<th>3</th>
<th>Participant Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>-</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td></td>
<td>X</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>X</td>
<td>1</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

*X = agree; - = disagree.

The extreme position of the four attitude statements. A participant who agrees with item 2 will agree with all four items. The items are ordered in the scalogram left to right from most to least extreme. If each agreement renders a score of 1, a score of 4 indicates all statements are agreed upon and represents the most favorable attitude. Persons with a score of 3 should disagree with item 2 but agree with all others, and so on. According to scalogram theory, this pattern confirms that the universe of content (attitude toward the appearance of this running shoe) is scalable.

The scalogram and similar procedures for discovering underlying structure are useful for assessing attitudes and behaviors that are highly structured, such as social distance, organizational hierarchies, and evolutionary product stages. The scalogram is used much less often today, but retains potential for specific applications.

summary

1 Managers know that the measurement of attitudes is an important aspect of strategy and often the best tool available because attitudes reflect past experience and shape future behavior. Attitudes are learned, stable predispositions to respond to oneself, other persons, objects, or issues in a consistently favorable or unfavorable way. Attitudes are generally thought to be composed of three components: affective, cognitive, and behavioral intentions.

2 Selecting and constructing a measurement scale requires the consideration of several factors that influence the reliability, validity, and practicality of the scale. Two broad research objectives are to measure characteristics of the individuals who participate in studies and to use participants as judges of the objects or indicants presented to them. Measurement scales fall into one of four general response types: rating, ranking, categorization, and sorting. The properties of data are classified in increasing order of power—nominal, ordinal, interval, or ratio—which determines how a particular measurement scale's data will be analyzed statistically. Measurement scales are either unidimensional or multidimensional. A balanced rating scale has an equal number of categories above and below the midpoint, whereas an unbalanced rating scale has an unequal number of favorable and unfavorable response choices. An unforced-choice rating scale provides participants with an opportunity to express no opinion when they are unable to make a choice among the alternatives offered. A forced-choice scale requires that they select one of the offered alternatives. The ideal number of points for a rating scale should match the stimulus presented and extract information proportionate to the complexity of the attitude object. The value of rating scales depends on the assumption that a rater can and will make good judgments. Errors of central tendency, halo effect, and leniency adversely affect a precise understanding of the measurement.

3 Rating scales have several uses, design features, and requirements. The simple category scale offers two mutually exclusive response choices. The multiple-choice, single-response scale offers the rater several options, including "other." The multiple-choice, multiple-response scale (also called a checklist) allows the rater to select one or several alternatives, thereby providing a cumulative feature.
The Likert scale consists of a series of statements, and the participant is asked to agree or disagree with each statement. Summation is possible with this scale although not necessary and in some instances undesirable.

The semantic differential (SD) scale measures the psychological meanings of an attitude object. Researchers use this scale for studies of brand and institutional image. The method consists of a set of bipolar rating scales, usually with 7 points, by which one or more participants rate one or more concepts on each scale item. The Stapel scale is used as an alternative to the semantic differential, especially when it is difficult to find bipolar adjectives that match the investigative question. Participants select a plus number for the characteristic that describes the attitude object. Ratings range from +5 to −5, and participants select a number that describes the object very accurately to very inaccurately.

Numerical scales have equal intervals that separate their numeric scale points. Verbal anchors serve as the labels for the extreme points. Numerical scales are often 5-point scales but may have 7 or 10 points. A multiple rating list scale is similar to the numerical scale but accepts a circled response from the rater, and the layout allows visualization of the results.

A scale that helps the researcher discover proportions is the constant-sum scale. The participant distributes 100 points among up to 10 categories. The graphic rating scale was originally created to enable researchers to discern fine differences. Raters check their response at any point along a continuum. Other graphic rating scales use pictures, icons, or other visuals to communicate with children or others whose limited vocabulary prevents the use of scales anchored with words.

Ranking scales allow the participant to compare two or more objects and make choices among them. Frequently, the participant is asked to select one as the “best” or the “most preferred.” When there are only two choices, as with the paired-comparison scale, the participant can express attitudes unambiguously by choosing between two objects. The forced ranking scale lists attributes that are ranked relative to each other. This method is faster than paired comparisons and more user-friendly. Often the researcher is interested in benchmarking. This calls for a standard by which training programs, processes, brands, point-of-sale purchases, or people can be compared. The comparative scale is ideal for such comparisons if the participants are familiar with the standard.

Q-sorts are a form of scaling that requires sorting of a deck of cards into piles that represent points along a continuum. The purpose of sorting is to get a conceptual representation of the sorter’s attitude toward the attitude object and to compare the relationships between people. Given a person’s total score, it is possible to estimate which items were answered positively and negatively on cumulative scales. A pioneering cumulative scale was the scalogram, a procedure for determining whether a set of items forms a unidimensional scale.
Terms in Review
1 Discuss the relative merits of and problems with:
   a Rating and ranking scales.
   b Likert and differential scales.
   c Unidimensional and multidimensional scales.

Making Research Decisions
2 Assume you are Menu Foods and you planned a major research study just prior to the largest pet food recall in our history. You plan to proceed with the study and feel you must add one or more questions to measure the consumer's confidence that your firm will be able to recover. Draft a scale for each of the following types that will measure that confidence.
   a Constant-sum scale.
   b Likert-type summated scale.
   c Semantic differential scale.
   d Stapel scale.
   e Forced ranking scale.

3 An investigative question in your employee satisfaction study seeks to assess employee "job involvement." Create a measurement question that uses the following scales:
   a A graphic rating scale.
   b A multiple rating list.

Which scale do you recommend and why?

4 You receive the results of a paired-comparison preference test of four soft drinks from a sample of 200 persons. The results are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Koak</th>
<th>Zip</th>
<th>Pabze</th>
<th>Mr. Peepers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Koak</td>
<td>-</td>
<td>50*</td>
<td>115</td>
<td>35</td>
</tr>
<tr>
<td>Zip</td>
<td>150</td>
<td>-</td>
<td>160</td>
<td>70</td>
</tr>
<tr>
<td>Pabze</td>
<td>85</td>
<td>40</td>
<td>-</td>
<td>45</td>
</tr>
<tr>
<td>Mr. Peepers</td>
<td>165</td>
<td>130</td>
<td>155</td>
<td>-</td>
</tr>
</tbody>
</table>

*Read as 50 persons preferred Zip to Koak.

a How do these brands rank in overall preference in this sample?

b Develop an interval scale for these four brands.

5 One of the problems in developing rating scales is the choice of response terms to use. Below are samples of some widely used scaling codes. Do you see any problems with them?
   a Yes—Depends—No
   b Excellent—Good—Fair—Poor
   c Excellent—Good—Average—Fair—Poor
   d Strongly Approve—Approve—Uncertain—Disapprove—Strongly Disapprove

6 You are working on a consumer perception study of four brands of bicycles. You will need to develop measurement questions and scales to accomplish the tasks listed below. Be sure to explain which data levels (nominal, ordinal, interval, ratio) are appropriate and which quantitative techniques you will use.
   a Prepare an overall assessment of all the brands.
   b Provide a comparison of the brands for each of the following dimensions:
      (1) Styling
      (2) Durability
      (3) Gear quality
      (4) Brand image

7 Below is a Likert-type scale that might be used to evaluate your opinion of the educational degree program in which you are enrolled. There are five response categories: Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, and Strongly Disagree. If Strongly Agree (SA) represents the most positive attitude, how would you value the items below? Record your answers to the items.
   a This program is not very challenging. SA AND SD
   b The general level of teaching is good. SA AND SD
   c I really think I am learning a lot from this program. SA AND SD
   d Students' suggestions are given little attention here. SA AND SD
   e This program does a good job of preparing one for a career. SA AND SD
   f This program is below my expectations. SA AND SD

In what two different ways could such responses be used? What would be the purpose of each?

Bringing Research to Life
8 What is the basis of Jason and Myra's argument for the need of an arbitrary scale to address customer expectations?

From Concept to Practice
9 Using the response strategies within Exhibit 12-1 or 12-10, which would be appropriate and add insight to understanding the various indicators of student demand for the academic program in which the students are enrolled?

From the Headlines
10 According to BusinessWeek, the U.S. workforce is becoming a temporary workforce. Even full-time workers are often contract employees without health, vacation, or other benefits. What measurement scale(s) would be appropriate to measure this trend?
You will find a description of each case in the Case Abstracts section of this textbook. Check the Case Index to determine whether a case provides data, the research instrument, video, or other supplementary material. Written cases are downloadable from the text website (www.mhhe.com/cooper11e). All video material and video cases are available from the Online Learning Center. The film reel icon indicates a video case or video material relevant to the case.