Marketing Scales

In: Essentials of Marketing Research: Putting Research into Practice

By: Kenneth E. Clow & Karen E. James

Pub. Date: 2014
Access Date: April 8, 2019
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781412991308
Online ISBN: 9781483384726
DOI: https://dx.doi.org/10.4135/9781483384726
Print pages: 284-321

Marketing Scales

Learning Objectives

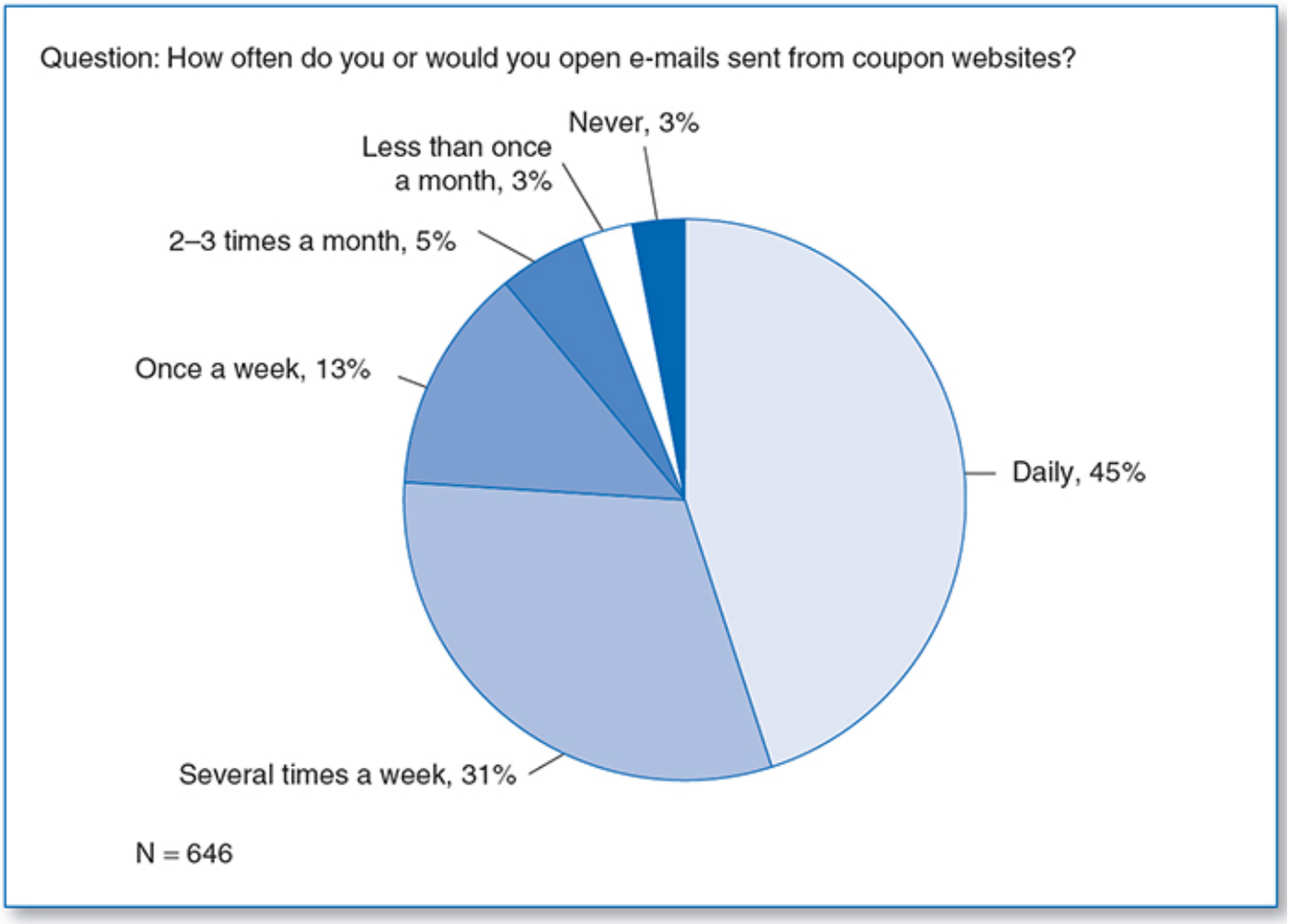
After reading this chapter, you should be able to

1. Discuss the concept of attitude measurement.
2. Explain the concept of using scales for attitude measurement.
3. Identify and describe the various comparative scales.
4. Identify and describe the various noncomparative scales.
5. Identify and describe scales that can be either comparative or noncomparative.
6. Discuss the considerations involved in selecting marketing scales.
7. Explain ways researchers can ensure the reliability and validity of scales.

Introduction

Marketing scales are used extensively by marketing researchers to measure a wide array of beliefs, attitudes, and behaviors. They can be used to measure beliefs individuals may have about brands, people, or objects. They can be used to measure people's feelings. They can also be used to measure past or current behavior as well as future intentions. Figure 10.1 shows the result of a single-item marketing scale that was used to measure how often individuals do or would open e-mails sent from coupon websites. Out of the sample of 646 respondents, 45% said they do or would open the e-mails daily, and another 31% said they do or would open the e-mails several times a week. Only 3% claimed they would never open them.

Figure 10.1 Graph of a Behavioral Marketing Scale



Source: Author-created with data from ReRez (www.rerez.com), December 2, 2011.

While scales can be used to measure behavior, they are especially important in the measurement of consumer attitudes, since attitudes cannot be observed. Various types of scales are presented in this chapter. Some scales ask individuals to make comparisons, while other scales do not. The chapter concludes with a discussion of how to ensure reliability and validity of marketing scales.

Measuring Attitudes

Objective 10.1: Discuss the concept of attitude measurement.

Attitudes are relatively enduring predispositions to respond to an object—such as a brand, a spokesperson, an ad, an event, or a store—in a consistent fashion. You may recall from a study of consumer behavior that an attitude consists of three components: cognitive, affective (liking), and behavioral. The cognitive component represents the belief and knowledge part of attitude. Examples of items designed to measure beliefs or

knowledge may be “The Ford Fusion was rated *Motor Trend*'s midsize car of the year” or “Pepsi has a sweeter taste than Coke.” The affective component of attitude is the feelings and emotions. Statements such as “I love my new Dell computer” or “I hate the taste of grapefruit” reflect the affective or feeling component of attitude. The behavioral component is the action or intentions aspect of attitude. It may involve measuring the intent to purchase a pair of Guess jeans at a retail store or the actual purchase of a Mounds candy bar.

Counting the number of candy bars sold from a point-of-purchase (POP) display is rather easy to do. So is comparing sales from different types of displays, different POP locations, and among stores using different types and locations of POP displays. Measuring consumer attitudes that influence those purchase decisions is more problematic. Since attitudes cannot be seen and exist only in the minds of consumers, researchers have to look at alternative methods of measurement. A common method is to use some type of scale. Some scales contain a single item, as in Figure 10.1, while others may contain multiple items. While ordinal scales are often used, researchers would prefer using interval scales since they produce higher-order data that can be subjected to more robust statistical tests. However, attitude scales that produce interval data are more difficult to construct.

Because an attitude consists of multiple dimensions, is abstract, is nonobservable, and rests in the minds of consumers, measuring attitude is challenging. To do so, researchers develop attitude constructs and measurement scales. How individuals respond to the items or questions making up the scale provides an understanding of their attitude.

Scale Development

Objective 10.2: Explain the concept of using scales for attitude measurement.

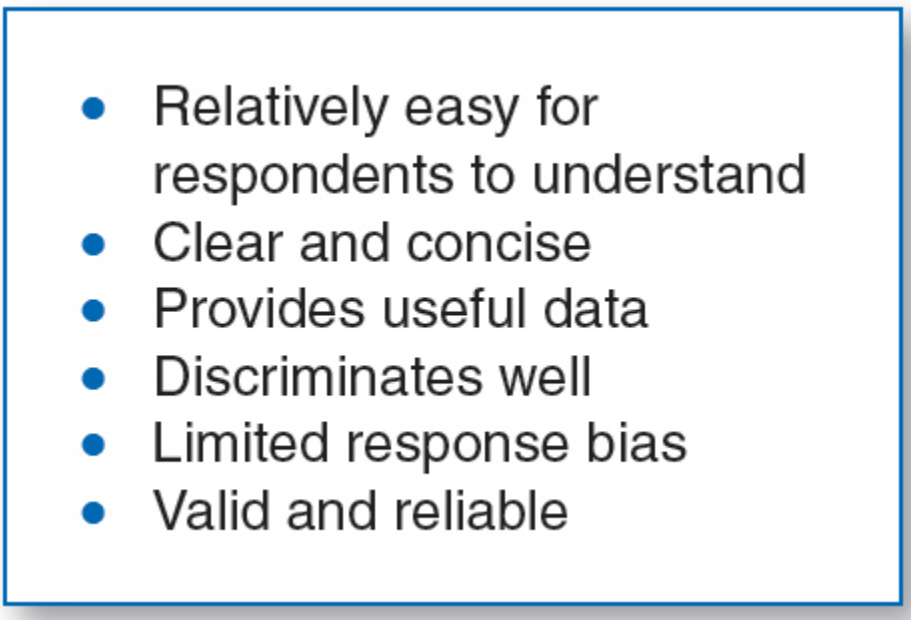
Since consumer attitudes cannot be observed, researchers develop scales to measure them. **Scaling** is the process of assigning numerical values or properties to subjective or abstract concepts. Attitude might be measured along a continuum using a 7-point scale with 7 indicating a very positive attitude and 1 indicating a very negative attitude. Thus, a value of 5 would indicate a more positive attitude than a value of 4, 3, 2, or 1. Alternatively, customer satisfaction attitudes might be assessed by selecting from among “Very Dissatisfied,” “Dissatisfied,” “Neutral,” “Satisfied,” and “Very Satisfied” scaled response categories.

Scales can be unidimensional or multidimensional. **Unidimensional scales** measure only one attribute or concept, such as a general attitude toward an advertisement. A researcher may use several items to measure the construct, but all of the items in a unidimensional scale measure a single concept. **Multidimensional scales** are designed to measure multiple dimensions or facets of a concept, an idea, or an attitude. Measuring store image involves multiple dimensions, such as atmospherics, aesthetics, product selection, and price.

Characteristics of a Good Scale

Developing or using good existing scales is important if a concept, such as attitude or satisfaction, is going to be measured with any degree of precision. Figure 10.2 identifies some of the characteristics of a good scale.

Figure 10.2 Characteristics of a Good Scale



Scales should be relatively easy for respondents to understand. Wording is important. It is advisable to use language that is used by respondents, but also important to ensure the scale items are clear and concise. Clarity and language familiarity are important, because these factors help to ensure that respondents understand the question, and interpret it correctly, thus minimizing measurement error. The scale needs to provide useful data, so in addition to being clear and concise, the items need to discriminate well among different attitudes held by respondents. If a 5-point scale is used to measure attitude toward a brand, and if 95% of the respondents check the same response category, then the question does not discriminate adequately because it doesn't identify differences in attitudes. Scale items can be developed or borrowed from existing scales that produce a wide range of responses. In rare cases when a given scale does not discriminate well among respondents, the problem may be not with the scale but with the sample selected. If avid Apple users are surveyed, they are very likely to rate the Apple brand as a 5 on a 5-point rating scale.

Good scales limit response bias. Asking individuals to evaluate fast-food restaurants in terms of food quality may produce response bias because respondents know the food is not the healthiest, even though they really like eating at fast-food restaurants. Thus, they may provide answers that are not completely honest. The reverse can also occur. If the survey is being taken in Japan and the respondents think the interviewer or survey sponsor wants to show the positive side of American food served in Japan, they may respond with a more positive attitude than they really believe.

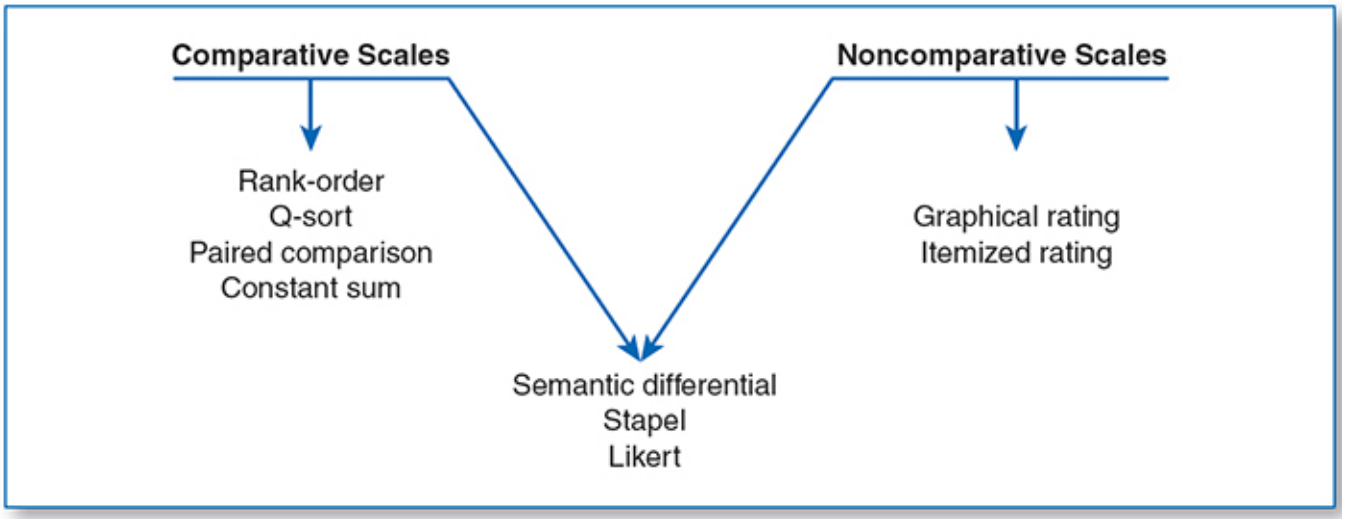
Lastly, scales need to be valid and reliable. The validity and reliability tests discussed in Chapter 9 can be

used to ensure a scale is valid and reliable. Alternatively, researchers can use scales that have already been established by prior researchers and that have been validated through proper research methods. It must be kept in mind, however, that if well-established scales are modified, used with a unique sample that is different from the general population, or used in a context that is significantly different, it may not produce valid and reliable results.

Marketing Scales

Scales can be divided into two primary categories: comparative scales and noncomparative scales. With **comparative scales** respondents are asked to evaluate or rate a brand, a product, an object, a concept, or a person relative to other brands, products, objects, concepts, or individuals or to an ideal item. **Noncomparative scales** involve respondents making judgments about a brand, a product, an object, a concept, or a person without reference to another item or an ideal item. Figure 10.3 identifies the most common scales used by marketing researchers. Typical comparative scales include rank-order, Q-sort, paired comparison, and constant sum. Common noncomparative scales are graphical rating and itemized rating. The semantic differential, Stapel, and Likert scales can be either comparative or noncomparative depending on how the question is worded.

Figure 10.3 Frequently Used Marketing Scales



Rank-Order Scales

Objective 10.3: Identify and describe the various comparative scales.

When researchers want to evaluate brands in relation to competing brands, rank-order scales are often used. **Rank-order scales** involve respondents comparing two or more objects, concepts, or persons and ranking them in some type of order sequence. Because respondents are asked to make comparisons, rank-order

scales are classified as comparative scales. They are relatively easy for respondents to answer and tend to mimic reality somewhat because consumers often will rank brands, products, or attributes mentally when faced with purchase decisions. Figure 10.4 illustrates a typical rank-order question.

Figure 10.4 Sample Rank-Order Questions

All of the following restaurants are located in the Oakview Mall area. Please rank the restaurants in terms of your personal preference from 1 being your most preferred, to 2 being your second most preferred, and so forth, to 7 being your least preferred. Each number can only be used once, and each restaurant listed below must be ranked.

_____ Chili's

_____ Jade Garden

_____ LongHorn Steakhouse

_____ O'Charley's

_____ Olive Garden

_____ Pueblo Viejo

_____ Red Lobster

While rank-order scales are relatively easy to administer, they do have some disadvantages (see Figure 10.5). First, the list of alternatives may not be categorically all-inclusive. In the case of the sample question shown in Figure 10.4, a researcher can identify all of the restaurants in the Oakview Mall area fairly easily, so this is not a problem. But, if respondents are asked instead to rank computers, a brand that a consumer prefers may not be on the list as it is much more difficult to identify all brands of computers. Ranking attributes or criteria used in making purchase decisions is even more difficult in terms of including all possible items. If the key criteria used by a consumer in making a decision are not listed, and thus not ranked, the resulting data are biased. It is also important to understand that for ranking data to be meaningful, all respondents must rank the exact same items relative to one another. This means that an "other" option with a fill-in-the-blank next to it cannot be used in a ranking question.

Figure 10.5 Disadvantages of Rank-Order Questions

- List may not be categorically exhaustive
- Respondent may not have knowledge or experience with all items listed
- Difficult to rank middle items in a long list
- Criteria used in the ranking may not be clear
- Produces ordinal data, not interval

A second disadvantage is that a consumer may have no experience with one or more items in the list to be ranked, and as a result may be unable to make meaningful evaluations. With the sample scale shown, a respondent may have no experience with Jade Garden or Pueblo Viejo and may only have eaten at an Olive Garden in another town. Thus, ranking those three restaurants may produce meaningless data.

A third disadvantage occurs when individuals are asked to rank a long list of items. The longer the list, the more difficult it is for individuals to distinguish between the items “in the middle.” Consumers know what they love, and they know what they hate, but differentiating between ranked items toward which they are relatively neutral in a meaningful way is difficult.

Fourth, ranking data is often of limited value because researchers do not have any knowledge of why the restaurants were ranked in a specific order. It could be based on price, on personal experiences, or on what someone told the respondents about a particular restaurant. Thus, rank-order questions may result in large amounts of measurement error. More importantly, the reason behind the rankings may provide more insight than the rankings themselves.

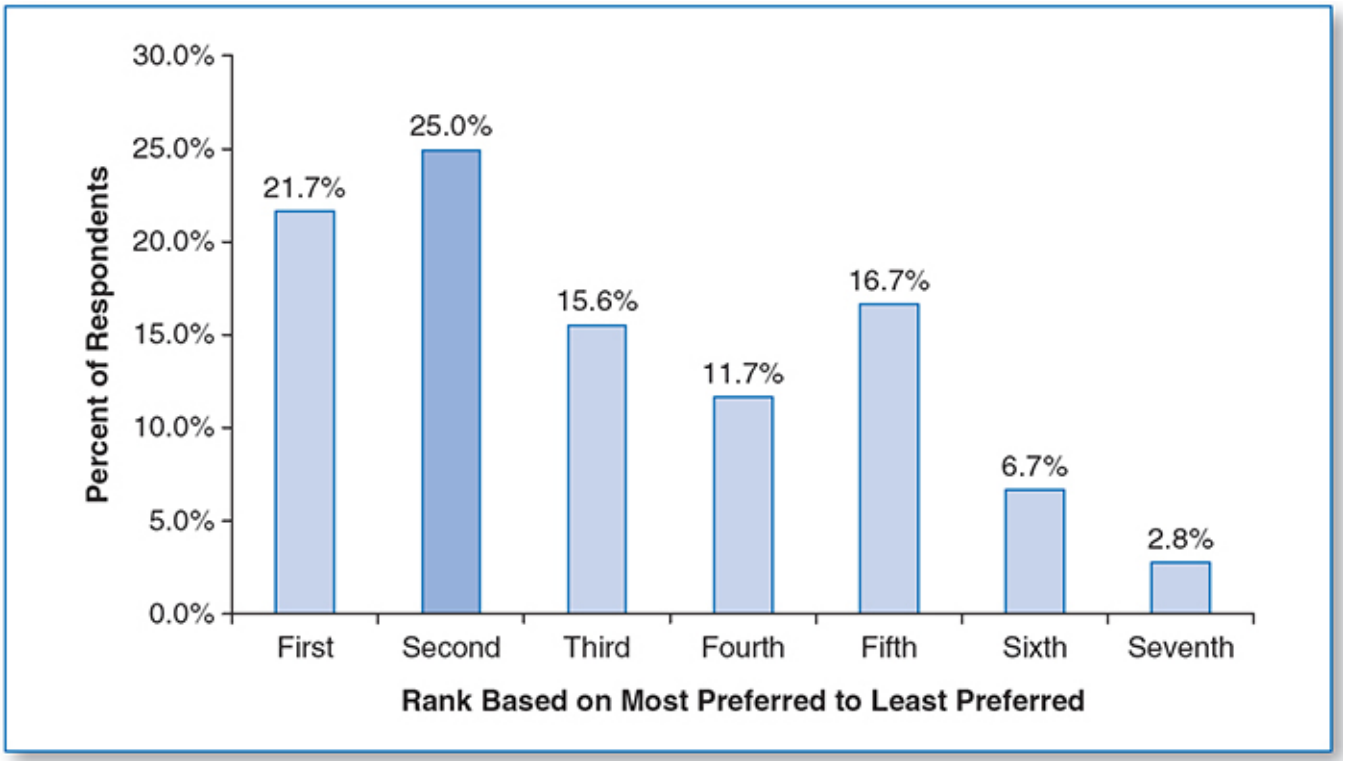
The last disadvantage relates to the type of data produced by rank-order scales—ordinal data. While ordinal data indicate an order, they do not indicate the distance between the rankings. Suppose both Maria and Josh ranked Red Lobster as 1 and Olive Garden as 2. Suppose, for Maria, the two were very close, and she had

a difficult time deciding which to rank 1 and which to rank 2. But, for Josh, the decision was easy. In his mind there was a large gap between Red Lobster and Olive Garden. The magnitude of this difference in preference is not captured by ordinal data. Similarly, ranking questions forces individuals to rank items they may not consider to be viable alternatives. Thus, while Josh may have ranked Jade Garden as 5, he may have no intention of ever eating at any of the restaurants ranked 5 through 7. The forced nature of ranking questions can be misleading in terms of gauging preferences.

Rank-order questions are best used in situations in which the consumer is highly familiar with all items to be ranked and when the number of items to be ranked is relatively low. While questions featuring five items to be ranked are optimal, ultimately, there should be fewer than ten. In many cases, an itemized rating question or constant sum scale may yield more useful information and should be considered as a viable alternative.

In terms of reporting results of a rank-order scale, Figure 10.6 shows the rankings for Olive Garden for a survey of 180 respondents. Notice that the highest percentage, 25%, ranked Olive Garden as their second most preferred. It was ranked either first or second by a combined 46.7% of the respondents, indicating it is a popular choice for dining in the Oakview Mall area. Very few ranked Olive Garden near the bottom at sixth or seventh.

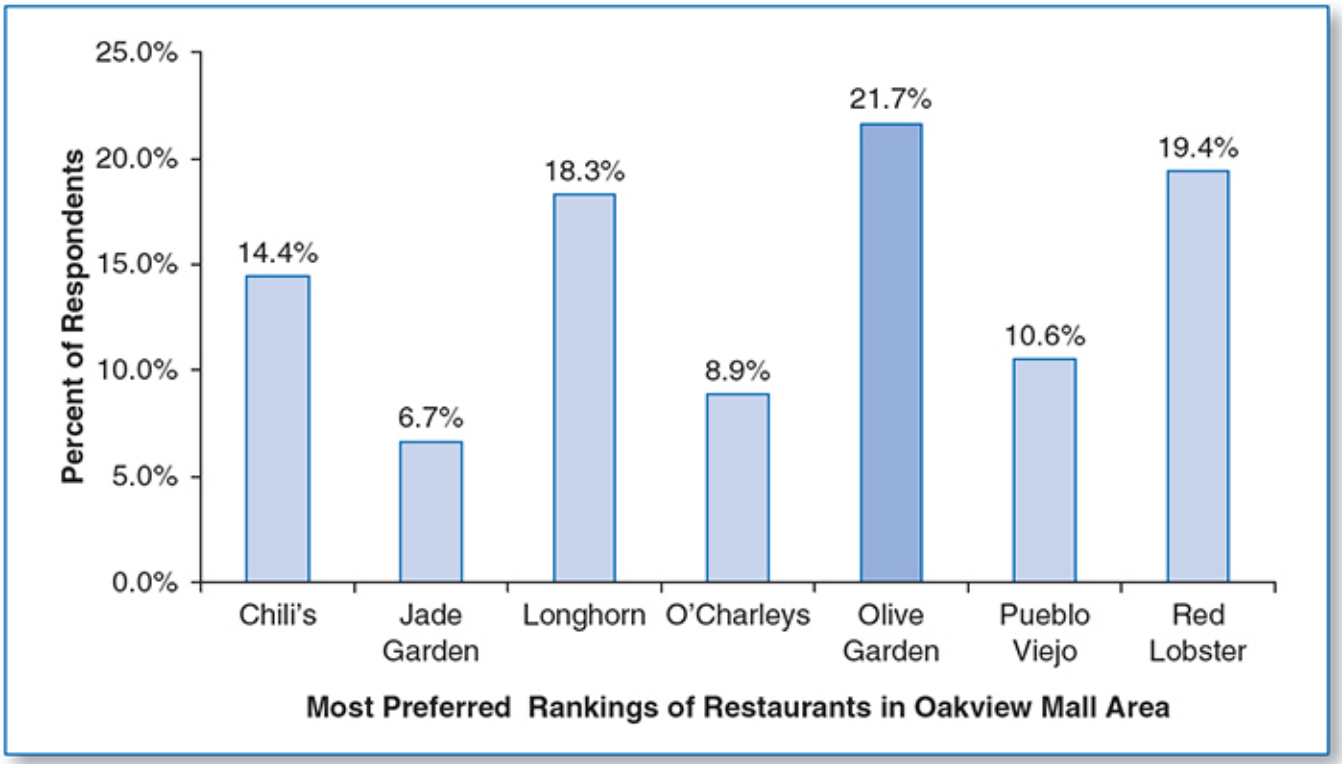
Figure 10.6 Results for Olive Garden



Another way the researcher may want to report the findings is to identify the top choice. Based on the graph in Figure 10.6, it would be easy to conclude that Olive Garden is the second most preferred restaurant in the Oakview Mall area. The manager of Olive Garden would likely be interested in which restaurant ranked first. Figure 10.7 shows the percent of respondents who ranked each of the restaurants as their “most preferred.”

Surprisingly, Olive Garden was ranked as the most preferred by the highest number of individuals, 21.7%. Red Lobster was second at 19.4%, and LongHorn Steakhouse was third at 18.3%. In examining results of rank-order scales and ordinal data, it is important for researchers to analyze the data in multiple ways. Frequencies, percentages, mode, and median can be used.

Figure 10.7 Most Preferred Brands



Q-Sort

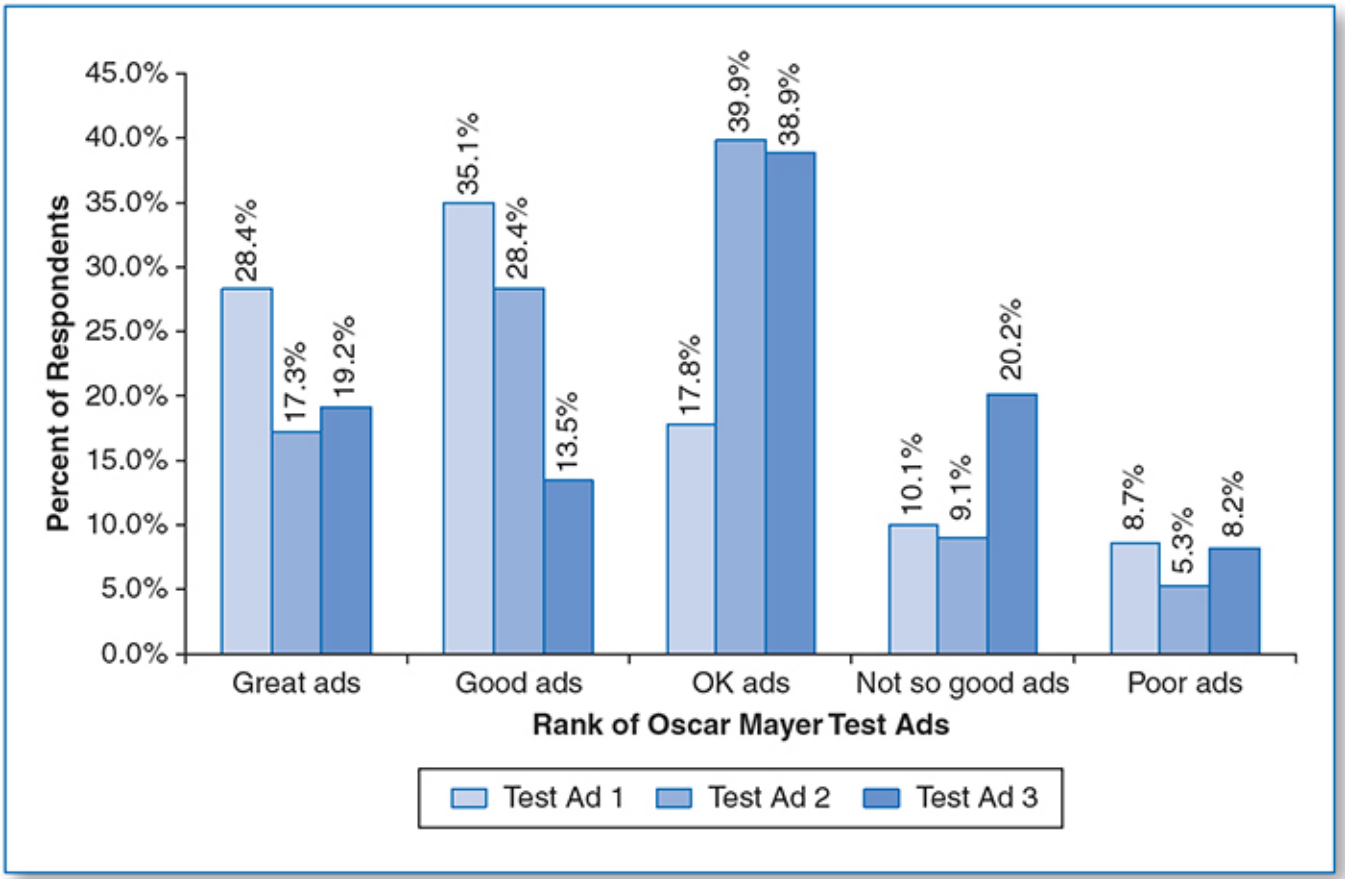
Q-sorting is a comparative technique whereby respondents rank a set of objects into a prespecified number of categories along a particular attribute or criterion. Rank-order scales are good for a small number of items while Q-sorting works better for a large group of items, such as 50 to 100. Ranking 50 items using rank-ordering would be impossible for an individual to do, but can be done using Q-sorting. With this method individuals are asked to place the items in piles based on some criteria.

Suppose Oscar Mayer wanted to test three new print ads that have been designed by its advertising agency. It gathers a large number of print ads from recent issues of magazines that fit the Oscar Mayer target market. The three new Oscar Mayer ads are placed in a stack randomly with 57 other ads, making 60 ads total. Respondents are asked not to study the ad, but just to look at it as if they were thumbing through a magazine. In the first sorting round, they would begin by picking the top 5—the "great" ads. Next, they would be instructed to select the 5 "poorest" ads from the remaining pile. In the second stage, respondents would first select the 10 next best ads and place them in the "good" ad pile, then identify the 10 "next worst" ads, which would be placed in the "not so good" pile. The 30 ads that remain after the second stage sort would be left in a fifth "OK ad" pile. Switching from one extreme to the other makes the sorting task easier and

faster than sorting great, good, OK, not so good, and poor ads in order. It is likely that the ads placed in the "great" pile would be ads that caused the respondents to stop and look closer. The ads in the "worst" pile would be those that were quickly ignored or were offensive in some fashion. The stacks between the two extremes would indicate some degree of interest. Using this method, a respondent could quickly evaluate 60 ads. Researchers with Oscar Mayer would be able to see how their test ads ranked compared to ads currently in circulated magazines.

Like rank-order scales, Q-sort scales produce ordinal data. The results of the Oscar Mayer Q-sort are shown in Figure 10.8. Test Ad 1 was placed in the first stack by 28.4% of the respondents and in the second stack by 35.1%. So about 63% of the respondents saw Test Ad 1 as either great or good, compared to the other 59 print ads they sorted. Notice Test Ads 2 and 3 did not produce as good of results as Test Ad 1. Less than 50% of the respondents placed the ads in the great or good groups. Clearly, Test Ad 1 is the one that Oscar Mayer should use based on this research, though other ad pretesting measures may be appropriate before making the final decision.

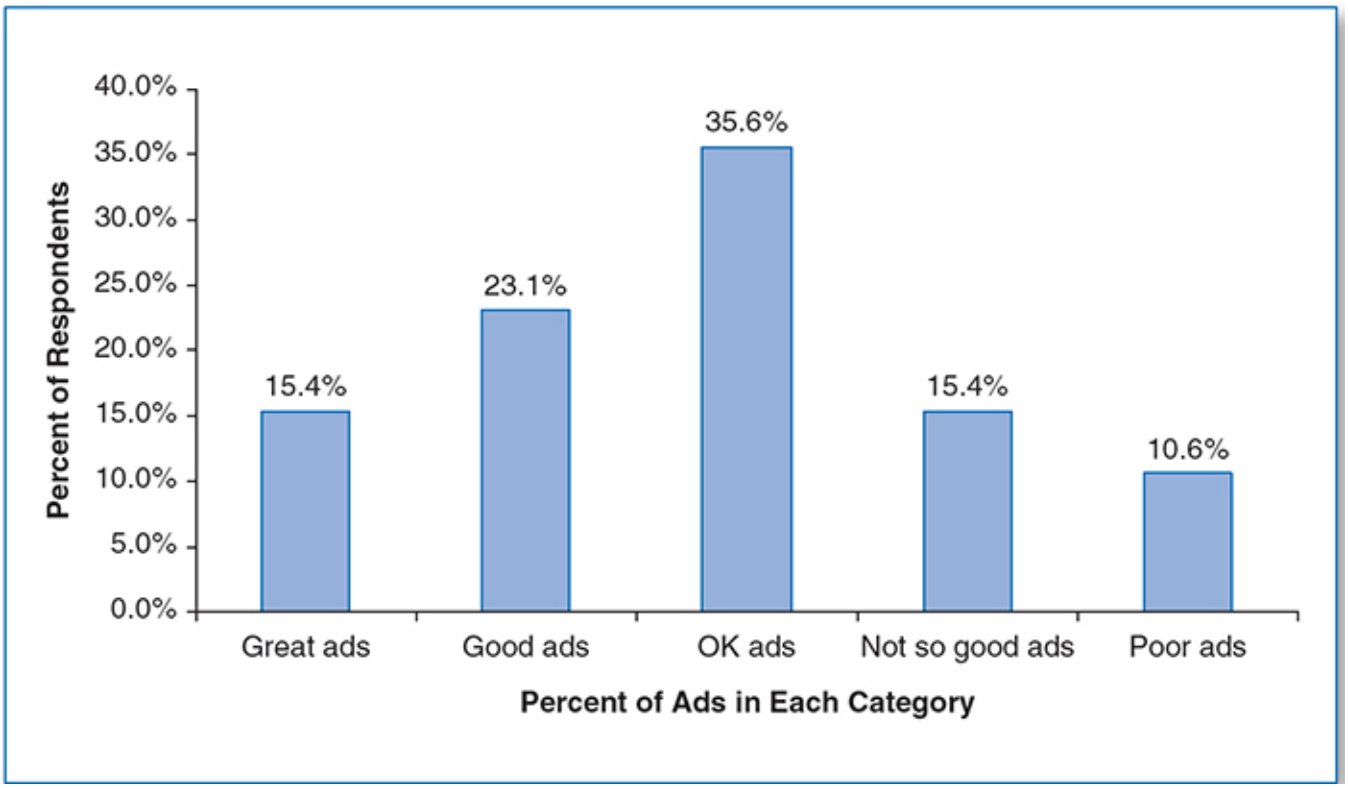
Figure 10.8 Q-Sort Results of Oscar Mayer Test Ads



With a Q-sort, the results tend to display a normal distribution curve. However, instructions given by the researcher, the types of items being sorted, and the category specifications can alter the data distribution. For example, respondents in the Oscar Mayer test could have been instructed to select all of the "great" ads—regardless of the number—and so on, rather than being forced to fill preset quotas for great, good, OK,

not so good, and poor categories. Had this been the case, the total distribution of ads for the Oscar Mayer Q-sort experiment could be similar to that shown in Figure 10.9, in which the distribution is skewed slightly positive. This information is beneficial to researchers since it tells them that as a group the respondents ranked more ads in the great and good categories than would be typical for a normal distribution curve.

Figure 10.9 Data Q-Sort Distribution



Paired Comparisons

Rather than ask individuals to rank order a set of items, researchers may use a series of paired comparisons. With the **paired comparison scale**, respondents choose one of two items in a set based on a specific criterion or attribute. Figure 10.10 illustrates a paired comparison scale examining four criteria consumers use in purchasing a laptop computer. The key advantage of the paired comparison scale is that it is typically easier for respondents to choose between two items than it is to rank a series of items. The paired comparison scale also tends to overcome order bias that may be created in listing the items for a rank-order scale. Respondents may be influenced by the order in which items are listed in the rank-order scale, even if the list is alphabetical.

Figure 10.10 Sample Paired Comparison Scale

$$(n)\,(n-1)\,/\,2$$

A major problem with the paired comparison scale is that all possible combinations must be listed. For a small set of items, such as the four used in Figure 10.10, only six paired combinations are needed. But, as the number of items to be evaluated increases, the number of paired combinations increases geometrically according to the following formula:

For each of the following pairs of criteria in purchasing a laptop computer, indicate which item in each pair is most important to you by placing a checkmark on the appropriate line.

_____ Price or Size of computer _____

_____ Physical appearance or Technical specifications _____

_____ Price or Physical appearance _____

_____ Physical appearance or Size of computer _____

_____ Technical specifications or Price _____

_____ Size of computer or Technical specifications

For instance, for five items it would take 10 pairs, for six items it would take 15 pairs, and for seven items it would take 21 pairs. Asking someone to evaluate 15 or 21 pairs of items can be quite taxing, producing respondent fatigue. By the end of the exercise, respondents may start checking responses and not spend much time thinking about the paired items being evaluated.

While ordinal data are produced from a paired comparison scale, reporting the results can be challenging. Figure 10.11 provides the results of the sample comparison scale on the four purchase criteria for laptop computers. Of the 240 respondents who completed the exercise, 80% said that price was more important than the size of the computer, 65% said price was more important than the physical appearance of the computer, and 76.7% indicated that price was more important than technical specifications. Clearly, price is important. The size of the computer may be the least important since it was selected less than the other three comparison alternatives. Physical appearance was more important than either size or technical specifications. Lastly, technical specifications were more important than size.

Figure 10.11 Results from a Paired Comparison Scale

Purchase Criteria*	Price	Size	Appearance	Specs
Price		80.0%	65.0%	76.7%
Size of computer	24.2%		42.9%	38.8%
Physical appearance	35.0%	57.1%		60.4%
Technical specifications	23.3%	61.3%	39.6%	

* Percentages indicate percent of sample that chose the row criterion over the column criterion.

Constant Sum

With the rank-order, Q-sort, and paired comparisons, relative distance between rankings cannot be determined and can vary substantially among respondents. To overcome this disadvantage, researchers can use a constant sum comparative scale. The **constant sum scale** asks respondents to allocate points among

various attributes or brands to indicate their importance or preference relative to one another. Typically, researchers will ask respondents to divide 100 points, but any number of points can be used, such as 10 or 20. Because the total points must add to 100 (or another specified number), the number of items to be ranked must remain small, usually under 10. Figure 10.12 illustrates a constant sum scale for evaluating the set of restaurants in the Oakview Mall area.

Figure 10.12 Example of a Constant Sum Scale

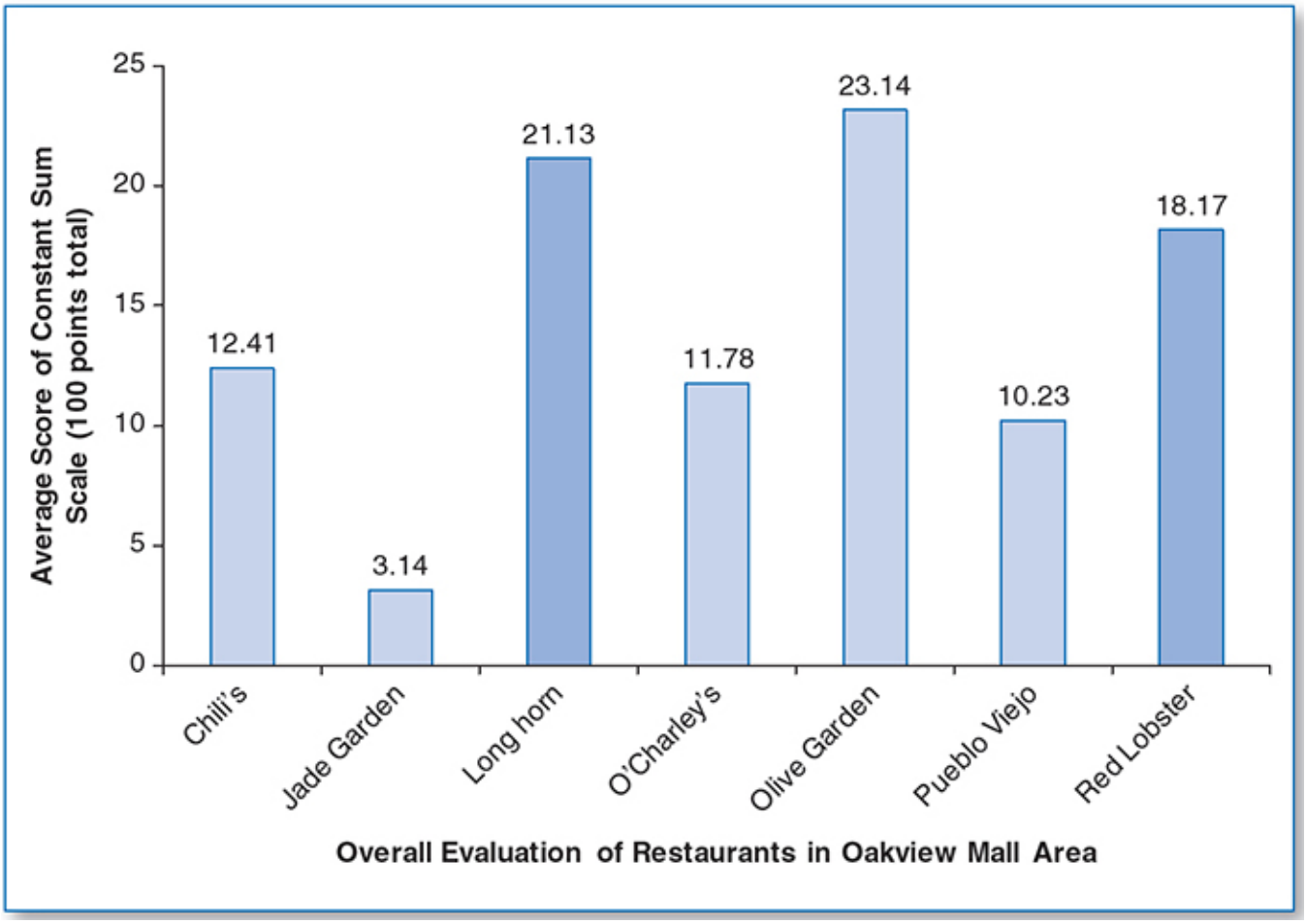
Listed below are the restaurants located in the Oakview Mall area. Please allocate 100 points among the seven restaurants based on your overall preference for each restaurant. The more points you assign to a restaurant, the higher the overall preference. The lower the number of points you assign, the lower the overall preference. A restaurant that is preferred twice as much as another restaurant should have twice as many points. It is possible to assign zero points to a restaurant if it is not at all preferred. The total number of points should add to 100.

Chili's	_____
Jade Garden	_____
LongHorn Steakhouse	_____
O'Charley's	_____
Olive Garden	_____
Pueblo Viejo	_____
Red Lobster	_____
Total number of points	100

The primary advantage of the constant sum scale over rank-order and the other comparative scales is that the relative distance between rankings can be determined. Refer back to the example of both Maria and Josh ranking Red Lobster and Olive Garden 1 and 2 using the rank-order scale. Recall that for Maria the two restaurants were very close while for Josh they were not. Using the constant sum scale, Maria may give Red Lobster a rating of 30 and Olive Garden a 28. Josh, on the other hand, may give Red Lobster a 35 and Olive Garden a 20. With the constant sum it is even possible to give the same number to two items. If Maria really could not decide between Red Lobster and Olive Garden, she could give them both the same score.

Another advantage of constant sum scales is that they produce ratio data, which are the highest order of data. In addition to rankings, researchers can compare the relative magnitude of the rankings. Suppose instead of using a rank-order scale researchers had used a constant sum scale to evaluate the seven restaurants in the Oakview Mall area. Figure 10.13 shows the results.

Figure 10.13 Results of Using a Constant Sum Scale



Results are very similar to the rank-order scale in Figure 10.7, but now relative magnitude can be compared. Notice that Olive Garden is still rated the highest, but using the constant sum scale, LongHorn Steakhouse is second and Red Lobster third. Because ratio data are produced, it is possible to say that the respondents' overall evaluation of Olive Garden is about twice as high as that for Chili's (23.14 compared to 12.41). The overall evaluation of Olive Garden is about seven times greater than for Jade Garden, a local Chinese restaurant. The low ranking for Jade Garden reflects the fact that constant sum scales allow respondents to assign a zero to items that are not at all preferred. Similarly, if asked to evaluate the relative importance of factors influencing a purchase decision, the constant sum scale would allow consumers to indicate which if any factors were not at all influential.

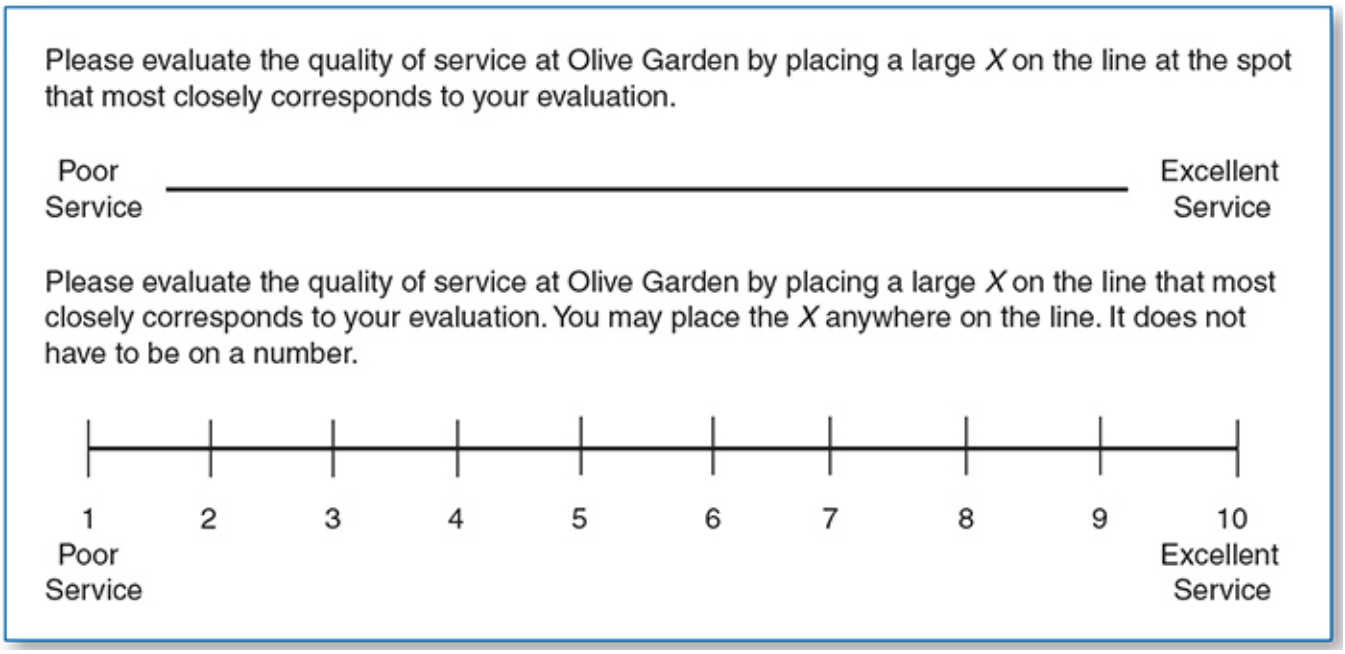
While the constant sum scale offers the advantage of producing ratio data and allows researchers to see relative distances between rankings, its primary disadvantage is the proper allocation of the points. Respondents may find it difficult to allocate 100 points among seven different choices. Unless the exercise is done on a computer that automatically adds the scores, respondents can easily make a mistake resulting in a total that does not add to 100. Using a smaller scale, such as 10, is easier for respondents to add the values, but a scale of 10 does not allow as much discrimination among choices as would 100.

Graphical Rating Scales

Objective 10.4: Identify and describe the various noncomparative scales.

Graphical rating scales are noncomparative scales that allow respondents to place a response anywhere on a continuous line. Respondents are not making any comparisons, but are asked to make a judgment or an evaluation. The scale is normally anchored at each end with antonyms or words with highly different meanings, such as *poor* and *excellent* or *friendly* and *unfriendly*. Figure 10.14 illustrates two different types of graphical rating scales. In the first example respondents can place an *X* anywhere on the line that they feel corresponds to their evaluation of service quality. The second scale has numbers, but respondents do not have to place their *X* on the line that corresponds to a number. They can place it between numbers if they want.

Figure 10.14 Examples of Graphical Rating Scales

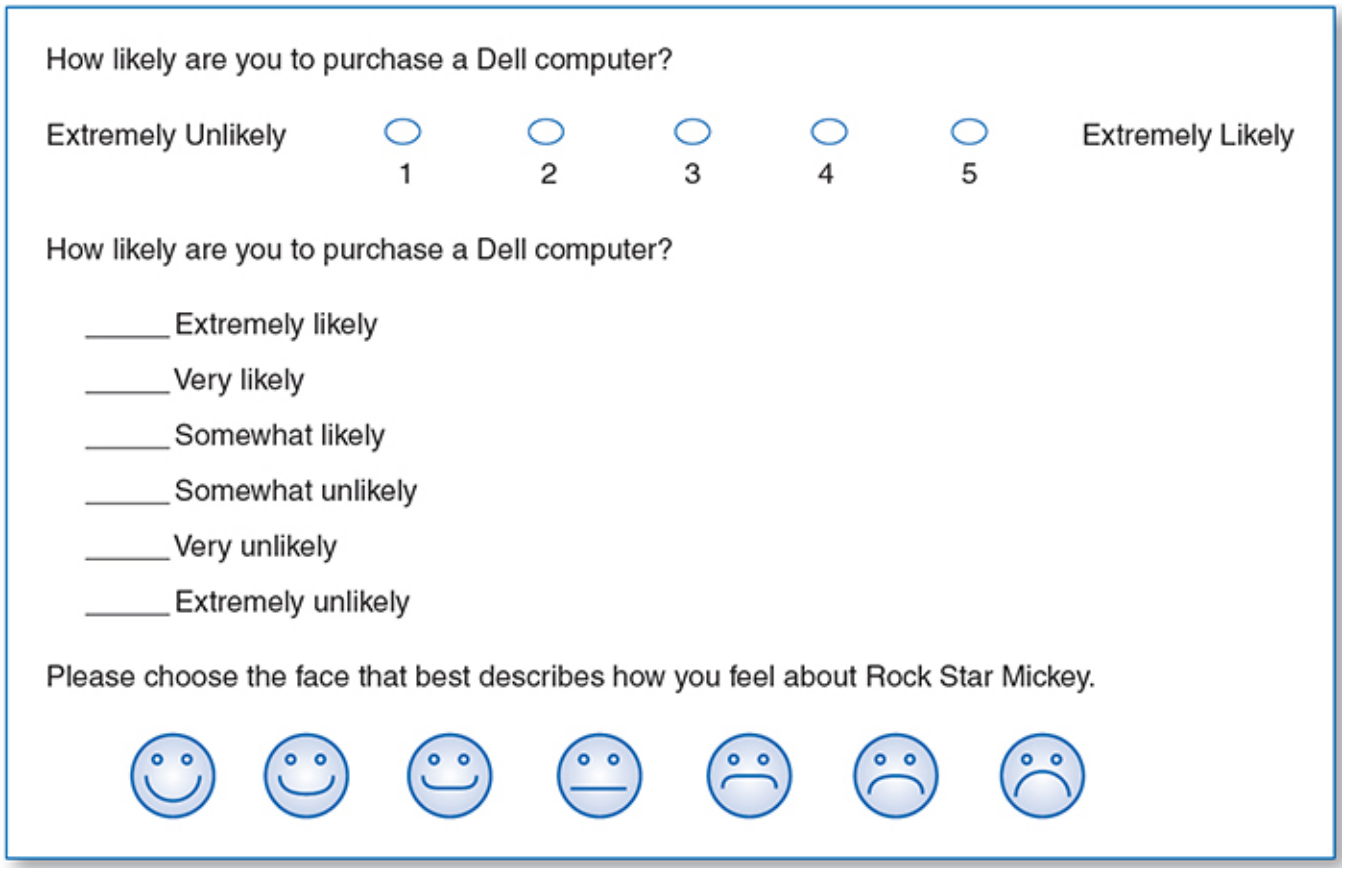


Graphical rating scales have the advantage of allowing respondents to provide a response along a continuum. The challenge is in converting the response to a number. If the line is 6 inches long, researchers may equate each inch with a number, thus a scale of 1 to 6. Alternatively, they could divide the line into 12 parts and have a scale of 1 to 12. Because of the continuous nature of a graphical scale, it will produce interval data. It cannot be ratio data since there is no absolute zero. Descriptive results would be reported in terms of a mean and standard deviation. Thus, for the graphical scales in Figure 10.14, the mean for the top scale using a 6-point scale may be 4.2, and for the bottom using a 10-point scale it may be 7.1.

Itemized Rating Scales

With an **itemized rating scale,** respondents choose a response from a select number of items or categories. It is also a noncomparative rating scale. Figure 10.15 illustrates three possible itemized rating scales. In the first question, respondents are asked how likely they are to purchase a Dell computer. They have five different possible answers they can check from “extremely unlikely” to “extremely likely.” In the second example, rather than have a 5-point scale with two extreme anchors, it has six categories, each with a distinctive label.

Figure 10.15 Examples of Itemized Rating Scales



Itemized rating scales may also use picture-based response options instead of words, as shown in the third example. This type of graphic itemized rating scale is commonly used in the medical profession for measuring patient pain, especially in emergency situations when patients may not be able to articulate how they feel. In marketing research, the use of the smiley/sad face scale may be perceived as insulting when administered to literate adults. Yet the same scale could be very appropriate for young children or mentally disadvantaged adults who can easily point to the face that best describes how they feel about the question as read by an interviewer.

The primary advantages of the itemized rating scale are it is easy for respondents to understand and easy for researchers to code. No subjective judgment is needed to identify what number should be given to a person's response. As a result, itemized rating scales are used extensively in marketing research. They can be used in all types of survey methods from telephone surveys to online surveys.

Most itemized rating scales produce interval data because it is assumed there is equal distance between

each of the categories or items in the response. In the examples given in Figure 10.15, researchers assume the distance between a 2 response and a 3 response is the same as the distance between a 4 and a 5. Therefore, the appropriate descriptive statistics would be a mean and standard deviation. For the first question, the mean may be 3.72 with a standard deviation of 1.04. For the second the mean may be 4.25 with a standard deviation of 0.96. Researchers may also want to know how many individuals checked each response, especially if the question was in response to an e-mail offer or a direct response offer.

Some itemized rating scales produce ordinal-level data (not interval), because the distance between response options cannot assumed to be equal. Figure 10.16 lists several examples. Importance, satisfaction, and other key attitudes can be assessed using itemized rating scales, but the manner in which the response categories are labeled will influence whether the resulting data are ordinal or interval in nature. When ordinal data are collected, means cannot be computed. Instead, data analysis should be limited to reporting frequencies and percents and, when appropriate, the mode or median.

Figure 10.16 Examples of Ordinal Itemized Scale Categories

Scales Assessing Frequency:

Very often	Often	Sometimes	Rarely	Never
All of the time	Most of the time	Some of the time	Just now and then	
All of the time	Very often	Often	Sometimes	Hardly ever

Scales Assessing Quality:

Excellent	Good	Fair	Poor	
Very good	Fairly good	Neither good nor bad	Not very good	Not good at all

Scale Assessing Uniqueness:

Extremely unique	Very unique	Somewhat unique	Slightly unique	Not at all unique
Very different	Somewhat different		Slightly different	Not at all different

Semantic Differential Scales

Objective 10.5: Identify and describe scales that can be either comparative or noncomparative.

A semantic differential scale can be either comparative or noncomparative depending on how the question is worded. The **semantic differential scale** involves a finite number of choices anchored by dichotomous words or phrases. Most semantic differential scales have 5 or 7 points, which allows for a neutral position. In some cases the neutral position is eliminated, and a 4- or 6-point scale is used to force respondents to choose one side or the other of the scale. The key to good semantic differential scales is choosing the anchor phrases or words that will produce discriminate answers among respondents. Scale anchors should be bipolar, meaning

that the anchors are perceived as “opposites” by respondents. Figure 10.17 provides an example of semantic differential scales for Home Depot.

Figure 10.17 Examples of a Semantic Differential Scale

Please evaluate the last purchase you made at Home Depot and the experience you had at the retail store.

Unfriendly staff	○	○	○	○	○	Friendly staff
Staff not very helpful	○	○	○	○	○	Staff very helpful
Poor selection	○	○	○	○	○	Excellent selection
Store unclean	○	○	○	○	○	Store clean
Poor value	○	○	○	○	○	Excellent value
Slow checkout	○	○	○	○	○	Fast checkout
Unsuccessful trip	○	○	○	○	○	Successful trip

When properly constructed, a major advantage of the semantic differential scale is its ability to discriminate differences in the direction and intensity of attitudes.[1] The key is choosing anchor phrases or words that reflect opposite meanings. This isn't always an easy task. For example, most people would agree that *love* and *hate* are bipolar opposites. But, what would be the appropriate opposite of *angry*? If you answered *happy*, then what word would you choose as the opposite of *sad*? If the anchors are chosen well, the semantic differential is relatively easy for respondents to understand and therefore easy to answer. If the anchors are not chosen well, it will lead to respondent confusion and scores that tend to drift to the midpoint. Also, if anchors are too strong, then the majority of responses will tend to be in the middle. However, if anchors are too weak, then all of the answers will be at one extreme or the other.

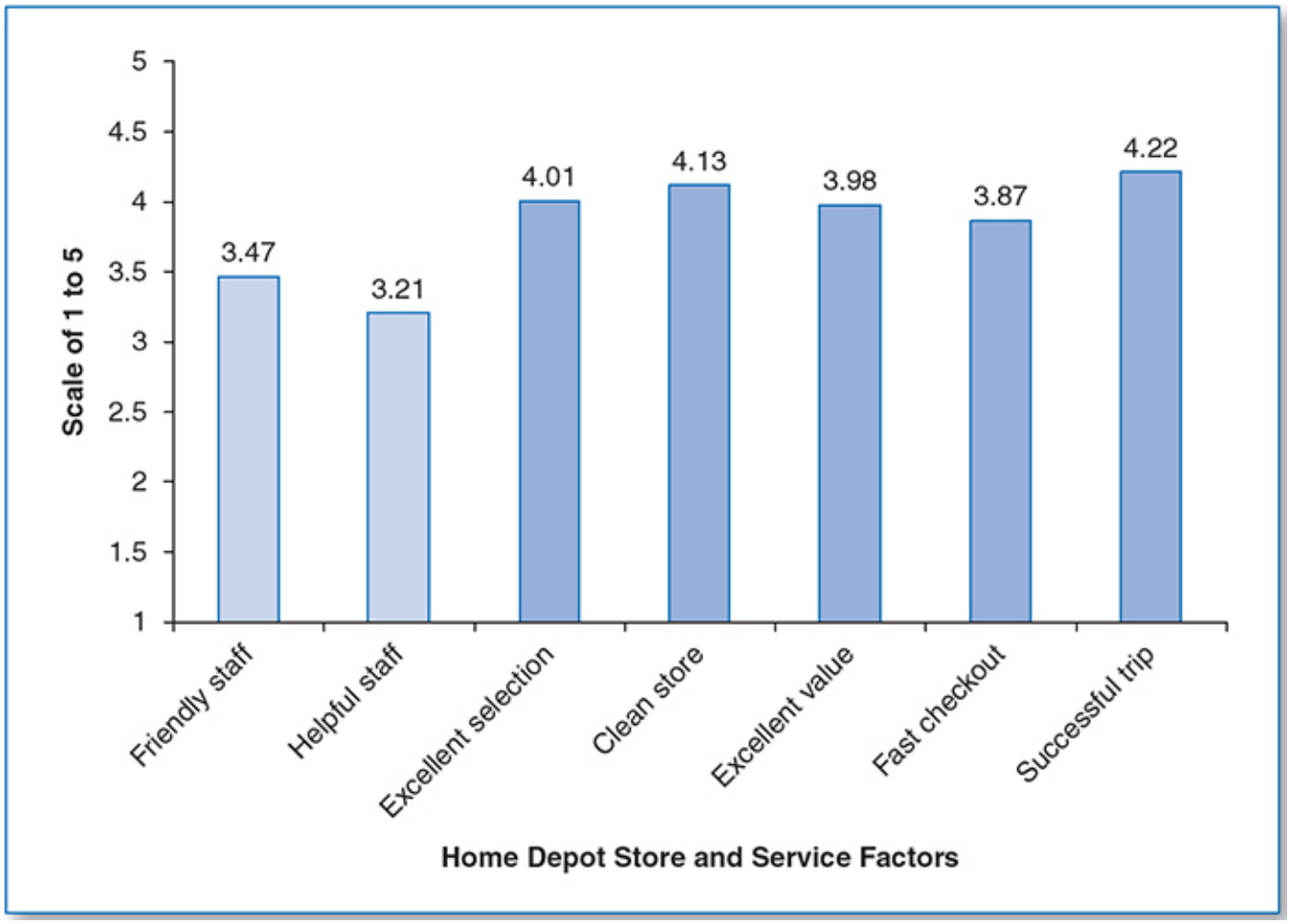
Semantic differential scales are excellent for assessing brand personality or image, and can be quickly answered by respondents. The results tend to be reliable. They eliminate problems that can occur with question phrasing when using itemized or Likert scales.[2] Semantic differential scales are also used frequently to assess attitudes toward an advertisement, a celebrity endorser, or a retail store.

A danger with using the semantic differential scale is the **halo effect,** which occurs when respondents have an overall feeling about the topic being surveyed and that overall perception influences their response to individual items so that all of the answers are relatively close to the same. Little or no discrimination among individual items occurs. For instance, if Carla's overall experience at Home Depot was positive, she may just go through and mark the fourth circle on each item. While it is possible that those responses may reflect her actual experience, it is also highly likely she did not think about each item sufficiently. It is unlikely that her experience would be a 4 for every single item. Past research using semantic differentials has shown that halo effect is more likely to occur when all favorable evaluations are placed on the left-hand side of the scale.[3]

The semantic differential scale shown in Figure 10.17 is a noncomparative scale because survey respondents are asked to evaluate their experience with Home Depot. No comparisons are made. If researchers want to compare Home Depot to Lowe's along the seven dimensions, then the scale can be modified asking respondents to evaluate Home Depot in comparison to Lowe's. In addition, the anchors can be modified. For instance, the first item can have anchors “friendlier staff than at Lowe's” and “less friendly staff than at Lowe's.”

Semantic differential scales produce interval data. Researchers assume there is equal distance between each of the points on the scale. Therefore, the appropriate descriptive measures would be a mean and standard deviation. Figure 10.18 shows the results of the semantic differential scale about Home Depot. The two lowest ratings in this particular survey were for friendly staff and helpful staff.

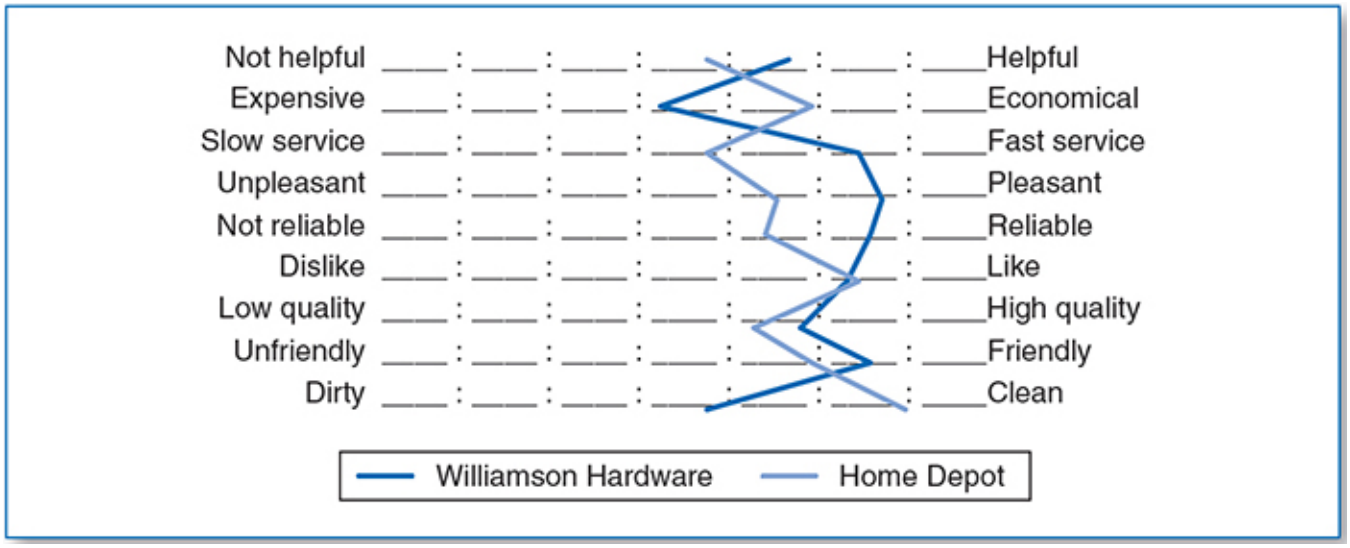
Figure 10.18 Results of Home Depot Semantic Differential



Semantic differentials are often used to develop brand and image profiles. Figure 10.19 shows an image profile for Williamson Hardware, a local retailer, compared to the national chain Home Depot. Instead of comparing a brand or business to a particular competitor, respondents can be asked to compare Williamson Hardware to other hardware stores in general—that is, against the industry. But, for this survey, individuals evaluated Williamson Hardware along nine dimensions and also Home Depot along the same nine dimensions. The mean scores on each item for each brand were then plotted and connected to provide a

visual representation of the resulting brand image profiles. A review of the results shows Williamson Hardware was evaluated higher on the dimensions of helpfulness, fast service, pleasantness, reliability, quality, and friendliness. On the negative side, Williamson Hardware was seen as being more expensive and not as clean. A paired sample t-test could be used to test whether the differences in the perceptions of cleanliness between Home Depot and Williamson Hardware are significant, meaning that they are not due to random sampling error. Should significant differences be found, the store manager can take steps to improve the cleanliness of the store, and thereby enhance the store's image.

Figure 10.19 Image Profile of Williamson Hardware



When multi-item measures are used to assess an individual's attitude toward an advertisement or a brand, the common practice is to create an overall average by summing the mean scores for each item. This summed score is then divided by the number of items in the scale to find the mean attitude score of the entire measure. This overall mean score for the scale or attitude dimension is then used in subsequent data analyses.

Stapel Scales

The **Stapel scale** is similar to the semantic differential, but uses only one anchor and both a positive and a negative numeric scale. Sample Stapel scales are shown in Figure 10.20. In this survey respondents are asked to evaluate Mel's Diner along the four dimensions. As it is currently worded, it is a noncomparative scale. It can be converted to a comparative scale by asking respondents to compare Mel's Diner to another restaurant, to restaurants in general, or to the ideal restaurant.

Figure 10.20 Examples of Stapel Scales

Please evaluate Mel's Diner along the following dimensions. Positive numbers indicate a positive evaluation of the criterion while negative numbers indicate a negative evaluation.

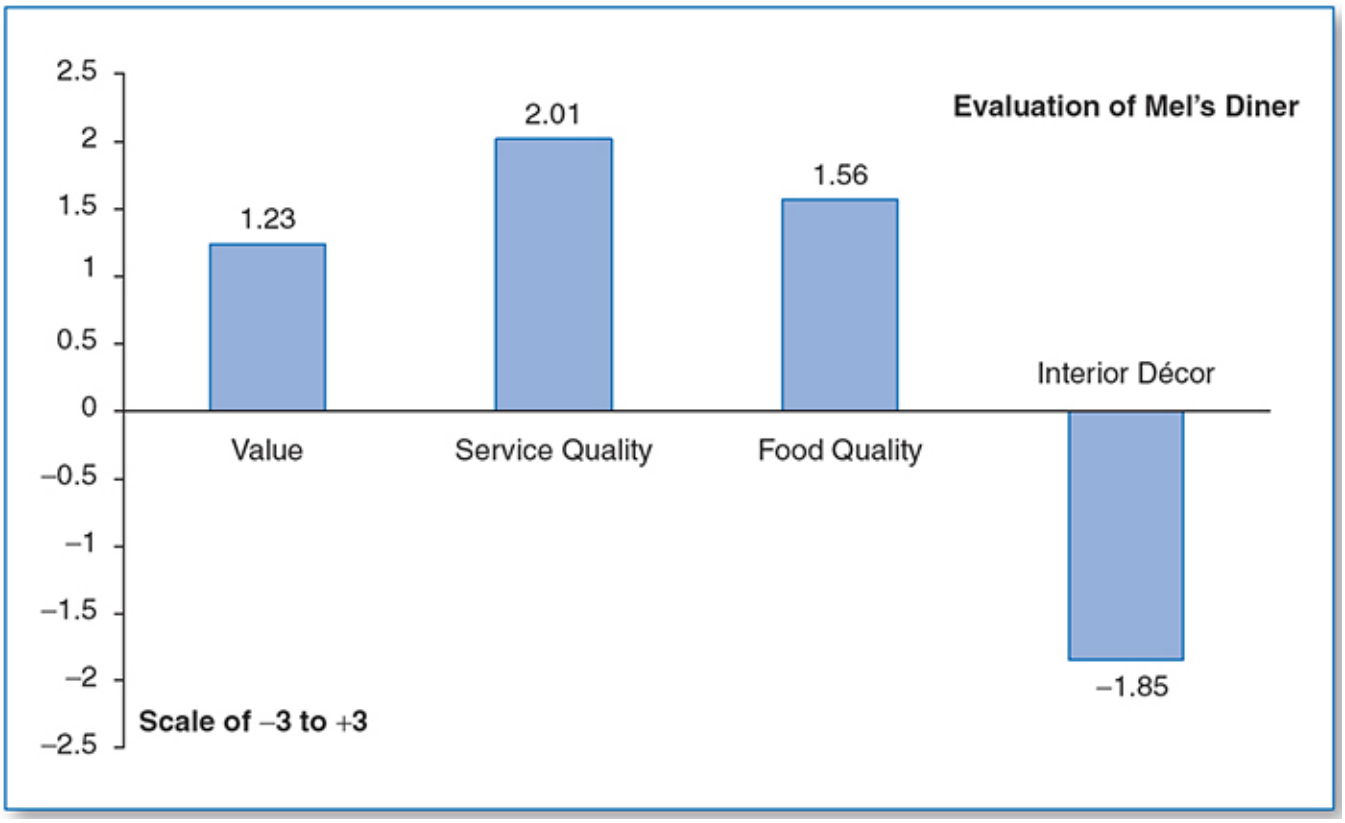
+3	+3	+3	+3
+2	+2	+2	+2
+1	+1	+1	+1
Value	Service Quality	Food Quality	Interior Decor
-1	-1	-1	-1
-2	-2	-2	-2
-3	-3	-3	-3

One of the challenges of the semantic differential scale is choosing bipolar anchors. The advantage of the Stapel scale is that only one anchor is needed. By using both positive and negative numbers, the scale produces both positive and negative evaluations. It is easy for respondents to use since they can signify negative evaluations by circling or checking a negative response.

The major disadvantage is the impact the one-word anchor can have. The word or phrase used as an anchor can be expressed in positive terms, in neutral terms, or in a negative format. Whichever format is used will impact the results. Using the anchor "value" produces different results than using anchors such as "poor value" or "good value." An additional disadvantage of the Stapel scale is that some respondents find it difficult to understand, and may circle the anchor itself, rather than the positive or negative number that expresses the degree to which the anchor reflects their positive or negative opinion. Misunderstandings of this nature are more likely to occur when non-neutral anchors are used. Finally, Stapel scales may take up more space on self-administered surveys, and thus be less desirable than semantic differentials. In practice, Stapel scales are not frequently used by marketing research practitioners.

Stapel scales generate interval data, and thus descriptive measures such as means and standard deviations are appropriate. Results for the Stapel scale are shown in Figure 10.21. The interior décor was rated negatively, at a −1.85. But, the restaurant provides a high level of service quality, 2.01, and food quality, 1.56. Value also had a positive evaluation. So despite the poor interior décor, Mel's Diner stays in business because of the high level of customer service and to some extent good food quality and good value for the money paid.

Figure 10.21 Results from Stapel Scale



Likert Scales

Likert scales are among the most popular for marketing research. **Likert scales** list a series of statements, and respondents are asked to indicate their level of agreement or disagreement with each statement. One reason Likert scales are used frequently is their ease of construction. Likert scales avoid the difficulty of finding the right anchor words or phrases that are commonly encountered when creating a semantic differential or Stapel scale. Likert scales are easy for respondents to read and understand. They are also easy to answer as subjects can respond to a whole series of statements that utilize the same scale. Figure 10.22 contains some Likert statements about fashion.

Figure 10.22 Example of a Likert Scale

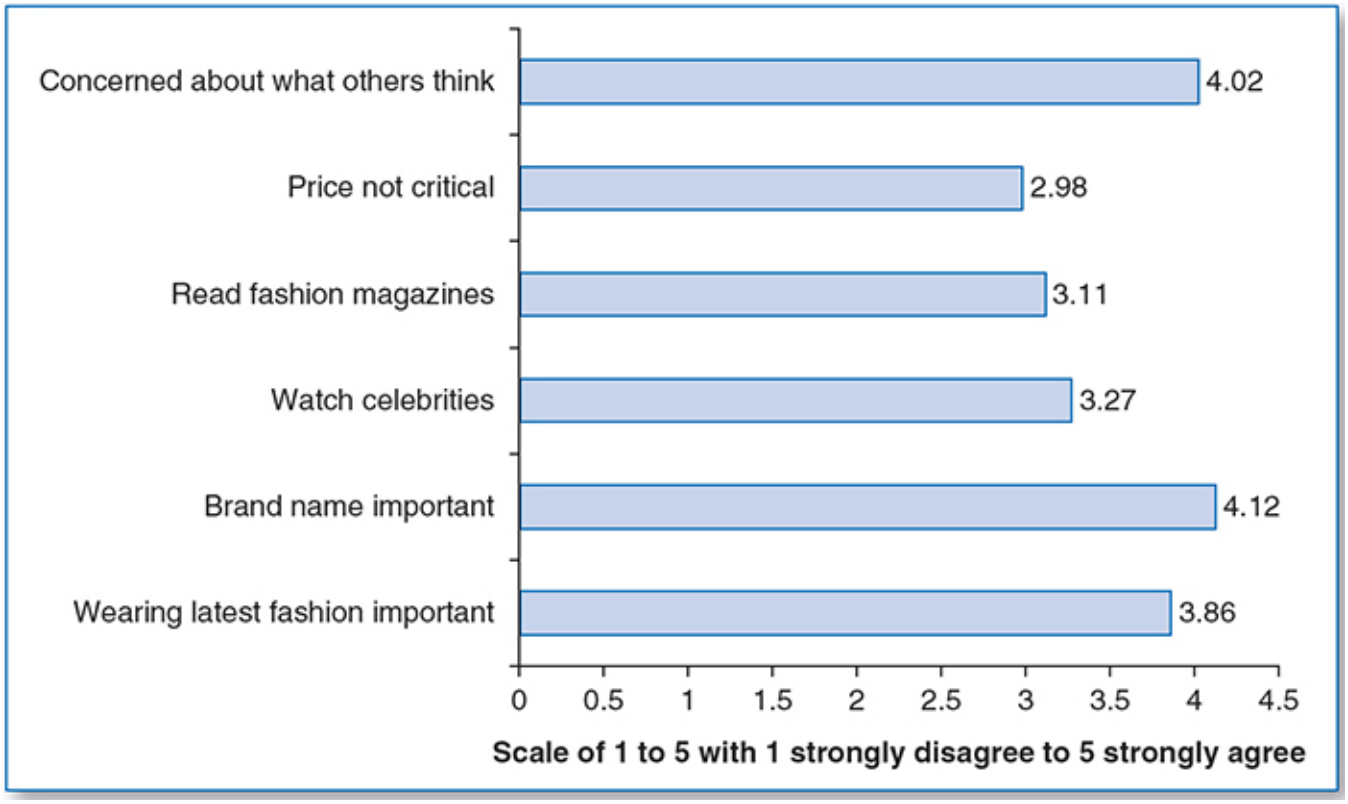
Please indicate your level of agreement or disagreement with each of the following statements.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. Wearing the latest fashion is important to me.	○	○	○	○	○
2. The brand name is important to me.	○	○	○	○	○
3. I watch what celebrities wear.	○	○	○	○	○
4. I read magazines such as *Glamour* regularly.	○	○	○	○	○
5. In purchasing clothes, price is not a critical factor. ..	○	○	○	○	○
6. I am concerned about what others think of me.	○	○	○	○	○

Likert scales can be either comparative or noncomparative. The scale shown in Figure 10.22 is noncomparative. A comparative Likert scale may ask individuals to respond to a series of statements about a particular retail store, such as the Gap, as it compares to other clothing retail stores or a specific store such as rue21. An example of a comparative Likert statement would be “rue21 offers higher-quality clothing than the Gap.”

Likert scales tend to be either 5-point or 7-point scales, with the 5-point scale being the most common. Because Likert scales produce interval data, means and standard deviations are the appropriate descriptive measures. This allows the researcher to conduct various tests for differences in attitudes based on demographic factors such as age, gender, or race, as well as other classification variables. Figure 10.23 graphs the results of the Likert scale about fashion. However, when multi-item Likert scales are used to assess an individual's attitude, the common practice is to sum or average the scores of all item means composing the attitude, and use this summated or average measure in data analyses rather than individual items.

Figure 10.23 Results of Likert Scale About Fashion



Likert scales often include both reversed (negatively phrased) and nonreversed (positively phrased) items. Doing so is thought to reduce response bias due to acquiescent respondents who are likely to be agreeable and rate everything positively[4] Mixing positive and negative items is also important because some argue it helps to alert lazy or less attentive respondents that the question content varies.[5] Unfortunately, reversing scale items has been found to diminish the reliability of scales and may result in respondents selecting the exact opposite of their true attitude.[6]

Scale Category Considerations

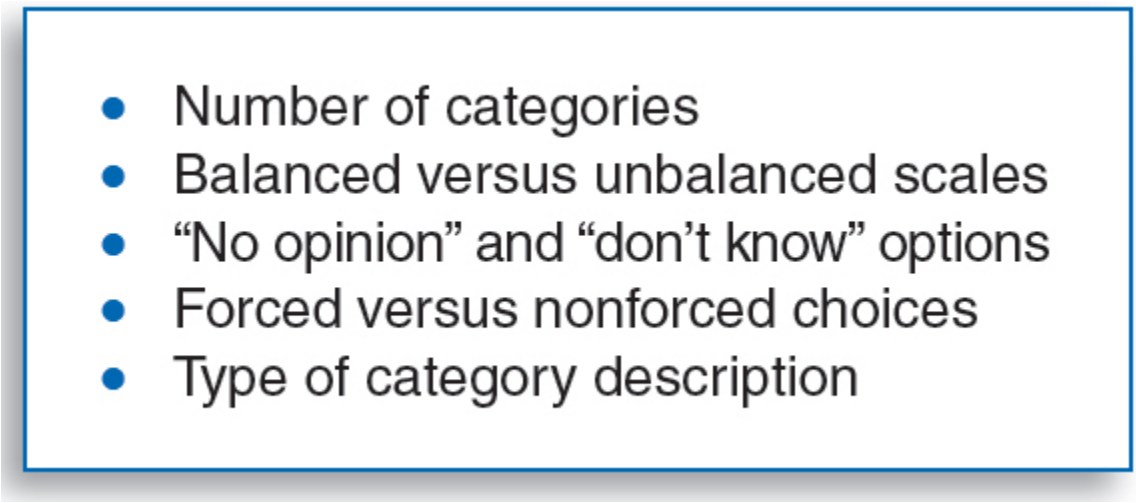
Objective 10.6: Discuss the considerations involved in selecting marketing scales.

Choosing the best scale for a research project requires an understanding of the research objective, target market, information needs, and mode of administration. If the objective involves comparisons with other brands or products, then one of the comparative scales should be used, or semantic differential, Stapel, or Likert scales should be used to assess multiple brands or products. The education level and reading ability of the target audience should also be considered. Some scales, such as the constant sum scale, require a higher level of education and math skills than do some of the other scales. To encourage

respondents to complete the survey and to provide honest, beneficial answers, the scales need to be easy for the respondents to read, understand, and complete, a factor that often works against the Stapel scale. Researchers must also consider the information needs of the research study, especially in terms of the data that each scale will produce. Ordinal data may be sufficient for some research studies, while others may need interval or ratio data. Mode of administration has an impact on scale selection. Some scales are easier to use with telephone surveys than others. If the respondents can see the scale, as with a web or paper survey, then more complicated scales can be used than if they can only hear the scale as it is read to them.

When it comes to actually creating the marketing scale, researchers must think about the number of categories, the use of balanced versus unbalanced scales, “no opinion” and “don't know” options, forced versus nonforced choices, and the extent of category description (see Figure 10.24). These decisions can have a significant impact on how respondents answer questions and the usefulness of the data that are produced.

Figure 10.24 Considerations in Creating Marketing Scales



Number of Categories

In developing scales, researchers must decide on the number of categories. A 2-point or 3-point scale lacks discrimination ability. Respondents tend to feel uncomfortable when using such a scale as it limits their freedom of expression. A Likert scale that has only three categories of “disagree,” “neutral,” and “agree” will not provide the researcher with a great deal of information. While it may technically be an interval scale, the mean will not be of much value. Frequency counts of how many agree and how many disagree may be more useful. By expanding the number of categories to 5, 7, or 10, greater detail is obtained, and the scale increases in discriminatory power. While 5- and 7-point scales produce comparable means, 10-point scales have been shown to result in slightly lower means.[7]

As the number of categories increases from 3 to 9, the reliability of the measure improves.[8] Thus, differences in responses are more likely to be found as the number of categories increases. But, having a scale with 10 or

more points may be too many for some respondents. It may be difficult for them to distinguish between values, such as 6 and 7 or 7 and 8. Reliability may also suffer as scales with a large number of categories have been found to be less reliable than those with fewer categories.[9] As a result, most marketing researchers recommend scales with 5 to 7 points as being optimal.[10] Often the decision on how many scale points comes down to researcher preference and the importance of the decision that needs to be made.

Another consideration in terms of number of categories is whether the scale should have an even or odd number of choices. When an odd number is used, such as 5 or 7, respondents can choose a neutral or middle position. With an even number of choices, respondents do not have that option. They must make a choice on one side or the other of the scale. It is a difficult decision. In some cases, individuals may not have an opinion, and so having a neutral position is valid. Forcing them to choose one side or the other may distort the data. However, the downside of having an odd number of choices is that it provides an easy escape for individuals who do not want to state an opinion.

Balanced Versus Unbalanced Scales

In addition to the number of categories, researchers must decide on whether to use a balanced or an unbalanced scale. A **balanced scale** will have the same number of positive choices as negative choices, and the phrasing of each negative choice typically mirrors the phrasing of each positive choice. An **unbalanced scale** is weighted toward one of the anchors, either positive or negative. Figure 10.25 illustrates each of the scales in relation to an individual's attitude toward wearing the latest fashions.

Figure 10.25 Examples of Balanced and Unbalanced Scales

Balanced scale

How important is it to you to wear the latest fashions?

____ very unimportant ____ somewhat unimportant ____ neutral ____ somewhat important ____ very important

Unbalanced scale

How important is it to you to wear the latest fashions?

____ very unimportant ____ somewhat unimportant ____ somewhat important ____ very important ____ extremely important

In the first example, the itemized rating scale has two negative response categories, a neutral position, and two positive response categories. It is balanced. In the second example, there are two negative response categories, no neutral position, and three positive response categories. It is unbalanced. The researcher could still use a neutral position in the unbalanced category if he or she desires. In most cases, a balanced scale is preferred to an unbalanced. But, if past research shows that most people lean toward a negative or positive position, then an unbalanced scale may be a better choice because it will provide greater

discrimination.

There are also situations where the target audience may dictate which scale is used. If the fashion question was asked of the general population, the balanced scale would be appropriate, and expanding the number of categories to a 7-point scale may produce even better results. If the population being studied is teenage girls, then an unbalanced scale may be better since most female teenagers are concerned about personal appearance and fashions.

“No Opinion” and “Don't Know” Options

Another decision that must be made is whether to include a “no opinion” or “don't know” option. As stated earlier, if an odd number of points is used in a scale, respondents tend to think of the center point as being neutral. They may also see it as a “don't know” or “no opinion” choice if none is provided. This is problematic, as researchers are not really sure if the respondent is neutral about the issue or doesn't really have an opinion. To avoid this situation and reduce measurement error, a “no opinion” or “don't know” option can be placed to the right of the scale. By placing it on the right side, respondents are first encouraged to express an opinion. Usually just one of the options is included, but both options can be used if researchers feel that there is a significant difference between someone who doesn't want to express an opinion and someone who truly has no experience or knowledge with which to answer the question.

The downside of including a “no opinion” or “don't know” option is that its presence often encourages respondents to use it. If a “no opinion” option is used on the Likert questions in Figure 10.22, a number of respondents may check it rather than indicate a level of agreement or disagreement with the statement. This is more likely to occur if an even-numbered scale is used in which a neutral position is not available. The same situation can occur with the semantic differential scales in Figure 10.17, if a “no opinion” option is available, though most semantic differentials do not include this category.

In surveys where respondents may not have any experience with the brand or product being studied, a “don't know” option is valuable. Suppose the respondent completing the questions about fashions (see Figure 10.22) was from France or another country and was not familiar with *Glamour* magazine. The respondent has to make a decision. She could skip the question and not answer it. She could answer it in terms of magazines in general, or guess what is meant by the question since she has no familiarity with *Glamour*. She might assume it is a magazine that features celebrities, tying it in with the previous item. By having a “don't know” option, the respondent can check it, and the researcher is not left with biased data.

Forced Versus Nonforced Choices

With semantic differential and Likert scales that use an odd number of response categories, respondents can check the middle option and not have to take a stand on an issue. On the Likert scales about fashions shown in Figure 10.22, an individual can choose “neutral.” While individuals can certainly be neutral about

the importance of fashion, brand names of clothes, pricing, and so forth, it is also likely that they do have an opinion. They may just choose the easy route of checking “neutral,” or if the researcher has a “no opinion” category, they may check it.

To force individuals to take a stand on issues, such as fashions, researchers can use an even number of points such as the 6-point scale in Figure 10.26 that does not have a neutral point. If there are no other options such as “no opinion,” the respondents must make a stand, either negative or positive. Thus, they have to either disagree or agree with the statement “I like wearing brand-name clothing.” The only other option open to the respondent is to refuse to answer the question and leave it blank. Most individuals, however, when forced to make a stand on an issue or a statement, will do so. Thus, by eliminating “neutral,” “no opinion,” and “don't know” options, researchers force respondents to either the positive or negative side of the statement or question.

Figure 10.26 Likert Scale Using Even Number of Points

For each of the following statements, indicate your level of agreement or disagreement by placing an *X* in the open box that most closely represents your feelings.

I like wearing brand-name clothing.	Strongly Disagree__|__|__|__|__|__Strongly Agree
I prefer to purchase clothes that are on sale.	Strongly Disagree__|__|__|__|__|__Strongly Agree
Wearing name-brand clothing makes me feel accepted by others.	Strongly Disagree__|__|__|__|__|__Strongly Agree
Advertising influences my decision on what brand to purchase.	Strongly Disagree__|__|__|__|__|__Strongly Agree
I use coupons when I buy clothes.	Strongly Disagree__|__|__|__|__|__Strongly Agree
I follow clothing trends so I can purchase the latest styles.	Strongly Disagree__|__|__|__|__|__Strongly Agree
I normally buy clothes that are on a sales rack.	Strongly Disagree__|__|__|__|__|__Strongly Agree
Price is important to me when selecting clothes.	Strongly Disagree__|__|__|__|__|__Strongly Agree
The brand name is important in my selection of clothes.	Strongly Disagree__|__|__|__|__|__Strongly Agree

Type of Category Description

Researchers must also decide whether each category in a scale should be labeled with a verbal description or a number, or whether only the scale anchors should be labeled. Figure 10.26 illustrates the latter situation. Alternatively, a Likert scale could be composed as shown in Figure 10.22 in which each category is labeled. A final example that combines both verbal and numerical category descriptions is shown in Figure 10.27.

Figure 10.27 Verbal and Numerical Descriptions in a Likert Scale

Please indicate your level of agreement or disagreement with each of the following statements by circling the number that best corresponds with your attitude. 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree.

	SD	D	N	A	SA
1. Wearing the latest fashion is important to me.	1	2	3	4	5

Using numbers to represent categories precodes the questionnaire and saves the researcher valuable time once the data have been collected, because the results of each survey can be immediately entered into the data file. Thus, manual coding of each survey item is avoided. Interestingly, the numbers chosen for scales can significantly influence the resulting ratings. Multiple studies have compared scales containing all positive numbers, such 0–10, 1–9, or 1–5 with scales in which numbers were negatively and positively balanced (–5 to +5, −4 to +4, −2 to +2). In each instance, scales using negatively and positively balanced numeric response categories produced more positive evaluations.[11]

Limiting the use of verbal labels to anchors is desirable when the researcher wants to be certain that interval-level data are collected. However, describing each category can be helpful when the goal is to better explain the options available to respondents. Obviously, the education level of respondents should be considered when making this decision. Respondents with lower levels of education are more likely to benefit from category descriptions that are applied to every option.

An additional consideration when choosing category descriptions relates to the strength of the anchors chosen when only anchors appear, and individual category descriptors are not provided. Suppose a question asked consumers to “Indicate your satisfaction level with your current cell phone service provider.” How should the anchors be phrased? Satisfied versus dissatisfied? Very satisfied versus very dissatisfied? Extremely satisfied versus extremely dissatisfied? While the choice may be influenced somewhat by the number of categories used in the scale, it is also important to understand that the adjectives chosen to accompany a category descriptor can influence respondents' choices. For instance, respondents would be less inclined to choose the first or last category when “extremely satisfied/dissatisfied” anchors exist than they would when the “very satisfied/dissatisfied” or “satisfied/dissatisfied” anchors are used.

Validity and Reliability Measurement

Objective 10.7: Explain ways researchers can ensure the reliability and validity of scales.

In developing marketing scales, researchers have three basic options. First, they can use scales that have

already been established by other researchers. The *Marketing Scales Handbook* by Bruner, Hensel, and James is an excellent source of scales that have already been established. Additional scales can be found at http://marketingscales.com. A major advantage of using these scales is that many have already gone through rigorous validity and reliability tests. Thus, results from these scales are likely to have a high level of validity and reliability.

A second option is to use established scales, but modify them to fit the product or situation being studied. By starting with scales that have gone through validity and reliability tests, it is likely that the new, revised scales will also have a higher level of validity and reliability.

The final option is for researchers to develop their own scales. Scale development is difficult and time consuming, and a great deal of academic research has been devoted to the scale development process, so those wishing to develop their own scales are best served by following established procedures.[12] Whether existing scales are modified or new scales are developed from scratch, researchers need to check the validity and reliability of their scales. Chapter 9 presented various methods of measuring reliability and ensuring validity.

Multiple items (or indicators) are often used to measure consumer attitudes or other marketing-related variables (or constructs), such as brand image or attitude toward an ad. Using more than one item to measure a single construct provides a more accurate picture than using just one-item indicators. Consider the Likert statements shown in Figure 10.26 that were taken from a survey about purchasing clothes. Researchers who developed the survey believe the nine statements selected measure two constructs: branding and price. To check if this is true, researchers can use two statistical methods: correlation analysis and factor analysis.

Correlation Analysis

Correlation is the degree to which one variable changes with another. If the correlation involves only two variables, it is called bivariate correlation. A common statistical test to measure bivariate correlation is the Pearson product-moment correlation. The Pearson test examines the two variables to see the amount of change in one variable compared to the amount of change in the other variable. Pearson correlations can vary from +1 to −1. A +1 score would mean an identical change. If variable *A* increased by 2, then variable *B* would increase by 2. If variable *A* declined by 4, variable *B* declined by 4. On the Likert scales shown in Figure 10.26, if two statements had a perfect +1 correlation, then the answers to both questions would always be the same.

A score of −1 would indicate an inverse relationship. If variable *A* increased by 2, then variable *B* would decline by 2 units. If variable *A* declined by 4, then variable *B* would increase by 4. A score of 0 would indicate no correlation at all. Changes in variable *A* had no relationship to the change in variable *B*. Scores of +1, −1, or 0 are extremely unlikely with a valid sample of respondents.

Referring back to the Likert statements about purchasing clothes, Pearson correlations that are positive

indicate respondents tended to give the same answers to two different questions. The higher the value (i.e., the closer to 1 the correlation), the more often the same response was given to the two questions. A Pearson correlation that is negative indicates that respondents tended to give opposite answers. Thus, if the respondents strongly agreed with Statement A and it had a negative correlation with Statement B, they would tend to disagree with Statement B.

For simplification purposes, the Pearson correlation and significance level of the first two Likert statements are shown with all of the statements for a sample of 194 respondents in Figure 10.28. The values shown in the first "Correlation" column indicate how well the statement in each row is correlated with the statement "I like wearing brand-name clothing." The significance level of the correlation is reported next. The last two columns contain the Pearson correlation and significance level with the statement "I prefer to purchase clothes that are on sale."

Figure 10.28 Partial Bivariate Correlation Matrix

Likert statements	I like wearing brand-name clothing.		I prefer to purchase clothes that are on sale.	
	Correlation	P-value	Correlation	P-value
Like wearing brand-name clothes	1		–0.310	0.000
Shop clothes on sale	–0.310	0.000	1	
Wearing brands, feel accepted	0.712	0.000	–0.276	0.000
Advertising influences	0.359	0.000	–0.317	0.000
Use coupons	–0.222	0.002	0.497	0.000
Wearing latest fashions important	0.379	0.000	–0.342	0.000
Purchase from sales racks	–0.072	0.321	0.545	0.000
Price is important	–0.084	0.243	0.570	0.000
Brand name important	0.687	0.000	–0.468	0.000

Notice the Pearson correlation between "I like wearing brand-name clothing" and "I prefer to purchase clothes that are on sale" is −0.310, which indicates an inverse correlation. The more the respondents liked wearing brand-name clothing, the less they preferred purchasing clothes off of the sales rack. The p-value of .000 indicates that this is a significant inverse relationship that is unlikely to have been found by chance. Thus, we can be confident that the results indicate a true inverse relationship in the population since it was true for this sample. The correlation between the first and third statements is 0.712, indicating a strong positive correlation between the statements "I like wearing brand-name clothing" and "Wearing name-brand clothing makes me feel accepted by others." The 0.712 can be thought of as a percent of common answers. Thus, 71.2% of the respondents checked the same spot on the Likert scale for those two questions. In terms of the

first and second statement correlations, it would indicate that 31% of the time respondents checked just the opposite answer. If they checked “strongly agree” for “I like wearing brand-name clothing,” then 31% of the time they checked “strongly disagree” for “I prefer to purchase clothes that are on sale.”

A review of the Pearson correlation matrix shows that the first item is highly correlated with the third statement and the last statement, “The brand name is important in my selection of clothes.” It is positively related to the statements about advertising and importance of wearing the latest fashions, but the correlations are not as strong. It is inversely correlated to the statements about purchasing clothes on sale and using coupons. This would indicate that individuals who felt brand names were important tended to put less emphasis on sales and coupons. But, notice that for the statements “I normally buy clothes that are on a sales rack” and “Price is important to me when selecting clothes” the correlation is negative. But neither correlation is significant, which indicates very little correlation between these two statements and “I like wearing brand-name clothing” in the sample under study. So, from the correlation analysis, the researcher can say there is no correlation between individuals who like wearing brand names and their desire to shop from sales racks and the importance of price to them.

A check of the final two columns shows just the opposite for the statement about shopping for clothes that are on sale. That statement is positively correlated with other statements about pricing and inversely correlated with statements about branding and fashions. All are significant.

Factor Analysis

A statistical procedure that is often used by researchers with multi-item attitude scales that are meant to represent multiple facets of an individual's attitude is factor analysis. A **factor analysis** involves analyzing a set of indicators (items) to determine underlying constructs dimensions) by reducing a larger number of items into a smaller subset of factors. Through factor analysis a researcher can determine which questions are measuring facets of the same component, or attitudinal dimension. For example, a researcher may start with 30 items that measure retail store image. Through factor analysis, these 30 items may be reduced to just 3 or 4 factors that measure different aspect of store image, such as atmospherics, layout, image, pricing, and product selection. The correlation matrix discussed in the previous section provides some information, but a factor analysis is a more comprehensive, statistical method that provides better information. A discussion of factor analysis can be found at the textbook's website along with instructions on how to conduct a factor analysis in SPSS.

Global Concerns

In developing scales to be used in other countries, researchers must be cognizant of differences in culture. A literal translation from English (or whatever language the original questionnaire was developed in) can yield scales that may actually have a different meaning, or at least not carry the same connotation. Engaging translators who understand the nuances of the language and culture can be extremely valuable. Double

translation, in which one person translates the survey into the foreign language and another translates this survey back into English, is often used to verify that the meaning between surveys is consistent.

Still, translation is a costly and time-consuming process. Researchers studying international populations that are bilingual (in English) are often tempted to use surveys written in English rather than in the native language of the population being studied. While doing so eliminates translation costs and saves time, recent research has shown that the quality of data suffers due to an anchor contraction effect. The **anchor contraction effect (ACE)** is best defined as a systematic form of response bias in which international subjects "report more intense emotions" when answering questions in English, as opposed to when they answer the same questions in their native language. Thus, ACE introduces bias into the data.[13] Given the growing number of multilingual individuals worldwide, the increasingly global focus of the business environment, and the widespread influence of the Internet, this is of serious concern. ACE could artificially inflate product or business ratings in a significant fashion.

People do not always respond to scales in the same manner. Culture can impact the types of responses that are given. For instance, many individuals believe that the Likert scale is universal and easily understood by people of any language and culture. Such is not the case. A Likert scale can yield different results depending on the culture context where it is used. Research has shown that using scales with numbers anchored only by "strongly agree" and "strongly disagree" can yield different results than using verbal words for each point on the scale.[14] The problem is that there is no conclusive evidence on which Likert scale is best because results were highly dependent on the topic being studied. With some topics, the numeric Likert scale performed better, but for other topics the verbal Likert scale was better. It is important for market researchers to carefully compare results across multiple countries since the scale used may impact results. It's possible that differences found may be due to the scale rather than true differences in attitudes or opinions. To avoid this problem, researchers will often use numerical scales that have only anchors at each end. This approach appears to present the least problems in comparing results across cultures.[15]

As an alternative, researchers may benefit from using more semantic differential scales in international research. Semantic differential scales are less prone to response bias, when compared with Likert scales. A study of U.S. and South Korean respondents found no statistically significant differences in extreme responses when the semantic differential was used, while both U.S. and South Korean respondents exhibited greater levels of extreme responses when Likert scales were employed.[16]

Another concern is with Hispanic cultures. In measuring attribute importance and brand performance, Hispanics have a tendency to rate both on the upper, positive side of the scale. With a 5-point scale, nearly all responses are in the 3 to 5 range. With a 10-point scale, almost all responses are in the 6 to 10 range. These ratings tend to be higher than those of the general population. A study by Jeffry Savitz found that on the average Hispanics rate products about 6% higher than non-Hispanics. So comparing Hispanic populations to the general population or non-Hispanic cultures may be interpreted incorrectly. Researchers may see a significant difference, but that difference may be due to a cultural trend of Hispanics to rate more positively,

rather than due to an actual difference in evaluation.[17] One approach that can be used with Hispanics to obtain more useful information is to use an unbalanced scale that has more positive points than negative. This will allow for greater discrimination in answers. It does not, however, allow for more accurate comparisons to non-Hispanics.

Statistics Review

Working with established scales improves the reliability and validity of the construct being measured. However, it is always a good idea to measure the reliability of the scale being used since the sample respondents and survey conditions vary. The best statistical tool to measure a scale's reliability is Cronbach's alpha.[18]

A survey was developed to examine the source credibility of models used in print advertisements. One of the scales used measured the model's perceived expertise to endorse the product being advertised. A 5-item scale was used. The results of the Cronbach's alpha statistical test are shown in Figure 10.29. The Cronbach's alpha score for the 5-item scale is .837. Typically, Cronbach's alpha reliability scores above .700 are considered good scales.

Figure 10.29 Initial Cronbach Alpha Reliability Measures

Reliability Statistics

Cronbach's Alpha	N of Items
.837	5

Item–Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Source Expertise 1: Experience	14.13	18.447	.753	.770
Source Expertise 2: Qualifications	13.79	19.799	.729	.779
Source Expertise 3: Expert	14.40	21.530	.626	.808
Source Expertise 4: Skills	13.93	24.525	.404	.862
Source Expertise 5: Knowledgeability	13.75	20.687	.697	.789

The fourth column of the bottom table shows the correlation of each question with the overall correlation with the other items. Notice the correlation of the fourth question is only .404, which indicates only about 40% of the time do respondents give the same answer to this question as the others. If the fourth question was deleted from the scale, the overall Cronbach's alpha score would improve to .862. This number is given in the last column of the table. If the Cronbach's alpha reliability test was run again in SPSS with Question 4 deleted, the Cronbach's alpha score would be .862. This second SPSS analysis with Question 4 deleted is shown in Figure 10.30.

Figure 10.30 Cronbach Alpha Score with Question 4 Deleted

Reliability Statistics

Cronbach's Alpha	N of Items
.862	4

Item -Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Source Expertise 1: Experience	10.55	13.036	.760	.803
Source Expertise 2: Qualifications	10.22	14.129	.743	.810
Source Expertise 3: Expert	10.83	15.682	.629	.855
Source Expertise 5: Knowledgeability	10.18	14.878	.712	.823

The four items are a reliable measure of the expertise construct. An examination of the last column of the output shows the scale's reliability cannot be improved by dropping any additional questions. The Cronbach's alpha score of .862 shows that approximately 86% of the time, the responses to the four questions are the same indicating the questions are measuring the same construct or dimension.

Dealing with Data

Source credibility of models in advertisements consists of five different dimensions. The “Statistics Review” examined one of the dimensions—expertise. The other dimensions are perceived trustworthiness of the model in the ad, attractiveness, similarity, and liking. Trustworthiness involves honesty and dependability. Attractiveness consists of the model's perceived physical beauty, elegance, and sex appeal. Similarity measures if the model is perceived to have similar values, tastes, and preferences as the respondent. Liking refers to the degree to which the model appears friendly, approachable, and likeable to the

respondent.

Access the SPPS data file "Chapter 10 Dealing with Data" at http://www.sagepub.com/clowess. The variables for each of the remaining four source credibility dimensions are provided. Use SPSS to obtain a Cronbach's alpha score for each of the following dimensions: Final Source Trust, Final Source Attractiveness, Final Source Similarity, and Final Source Liking. When you are finished, answer the following questions.

1. What is the Cronbach's alpha score for the "trustworthiness" dimension? Can the scale be improved by deleting a question? If so, which question? Based on this initial printout, what would be the Cronbach's alpha if the item was deleted? Run a new Cronbach's alpha with the item deleted. What was the new Cronbach's alpha score?
2. What is the Cronbach's alpha score for the "attractiveness" dimension? Can the scale be improved by deleting a question? If so, which question? Based on this initial printout, what would be the Cronbach's alpha if the item was deleted? Run a new Cronbach's alpha with the item deleted. What was the new cronbach's alpha score? Can the reliability measure be improved by dropping a second question? Why or why not?
3. What is the Cronbach's alpha score for the "similarity" dimension? Can the scale be improved by deleting a question? If so, which question? Based on this initial printout, what would be the Cronbach's alpha if the item was deleted?
4. What is the Cronbach's alpha score for the "liking" dimension? Can the scale be improved by deleting a question? If so, which question? Based on this initial printout, what would be the Cronbach's alpha if the item was deleted? Run a new Cronbach's alpha with the item deleted. What was the new Cronbach's alpha score?

Summary

Objective 1: Discuss the Concept of Attitude Measurement

Measuring consumer attitudes that influence purchase decisions is challenging because it is not visible and only exists in the minds of consumers. Attitude consists of three components: cognitive, affective, and behavioral. The cognitive component represents the

belief and knowledge part of attitude. The affective component of attitude is the feelings and emotions. The behavioral component of attitude is the action or intentions aspect of attitude. Because attitude consists of multiple dimensions, is abstract, and is in the minds of consumers, researchers typically develop multi-item scales to measure it.

Objective 2: Explain the Concept of Using Scales for Attitude Measurement

Scaling is the process of assigning numerical values or properties to subjective or abstract concepts. Scales can be unidimensional or multidimensional. Unidimensional scales measure only one attribute or concept, such as attitude toward an advertisement. Multidimensional scales are designed to measure multiple dimensions or facets of a concept, an idea, or an attitude. Developing good scales is important if a concept is going to be measured with any degree of precision. Scales should be relatively easy for respondents to understand, making it advisable to use language that is used by respondents. Scale items should be developed that produce a wide range of responses yet limit response bias. Lastly, scales need to be valid and reliable. Researchers can use scales that have already been established by prior researchers and that have been validated through proper research methods. Scales can be divided into two primary categories: comparative scales and noncomparative scales.

Objective 3: Identify and Describe the Various Comparative Scales

With comparative scales respondents are asked to evaluate or rate a brand, a product, an object, a concept, or a person relative to other brands, products, objects, concepts, or individuals or to an ideal item. Typical comparative scales include rank-order, Q-sort, paired comparison, and constant sum. Rank-order scales involve respondents comparing two or more objects, concepts, or persons in some type of sequential order. They are relatively easy for respondents and tend to mimic reality. Q-sorting has respondents rank a set of objects into a prespecified number of categories. Q-sorting works well for a large number of items. With a paired comparison scale, respondents are asked to choose one of two items based on some criteria specified by the researcher. The number of comparisons grows quickly since all possible pairs must be evaluated. The constant sum scale has respondents allocating points (often 100) among various alternatives based on some prespecified criteria.

Objective 4: Identify and Describe the Various Noncomparative Scales

Noncomparative scales involve respondents making judgments about a brand, a product,

an object, a concept, or a person without reference to another or an ideal item. Common noncomparative scales are graphical rating and itemized rating. Graphical rating scales allow respondents to place a response anywhere on a continuum. It may or may not have numbers. With itemized rating scales, respondents choose from a set of alternatives.

Objective 5: Identify and Describe Scales That Can Be Either Comparative or Noncomparative

The semantic differential, Stapel, and Likert scales can be either comparative or noncomparative depending upon how the question is worded. Semantic differential scales involve a finite number of choices anchored by bipolar words or phrases. Semantic differential scales are easy for respondents to understand. The challenge, however, is choosing words or phrases that are opposites. Stapel scales have only one anchor, and respondents can indicate a negative or positive possession of the trait or characteristic being examined. Likert scales are commonly used in marketing and involve respondents indicating the degree to which they agree or disagree with statements. Likert scales are easy to understand and relatively easy to construct.

Objective 6: Discuss the Considerations Involved in Selecting Marketing Scales

Choosing the best scale for a marketing study requires an understanding of the research objective, target market, information needs, and mode of administration. When it comes to actually creating the marketing scale, researchers need to consider the number of category responses, balanced versus unbalanced scales, “no opinion” and “don't know” options, forced versus nonforced choices, and the type of category description. Scales can have as few as 2 choices and as many as 10, or even more. However, typically marketers use from 5-point to 7-point scales. A balanced scale will have the same number of positive choices as negative, while an unbalanced scale will be weighted in one direction. Researchers have to decide if they want to include a “no opinion” or “don't know” option. Allowing respondents to choose this option may encourage them to select it rather than state a view. Using an even number of categories forces individuals to choose either the negative or positive side since no neutral position exists. The last consideration is whether to use verbal descriptors for every category, verbal descriptors at the anchor points only, and/or numeric descriptors for categories. The choice made can influence respondents' answers.

Objective 7: Explain Ways Researchers Can Ensure the Reliability and Validity of Scales

In using scales, researchers have three options: use a scale that has already been developed, adapt a scale that has already been developed, or create a new scale. With the last two options it is important to ensure reliability and validity of the scale. One method of doing this is through a correlation analysis. Items that have a high correlation are measuring the same construct and therefore fit together in a scale. Another option is a factor analysis. This procedure will examine data to determine underlying constructs by reducing a large number of items into smaller subsets of factors. Items that are related will be placed together within the same factor.

Glossary of Key Terms

Anchor contraction effect (ACE): systematic form of response bias in which international subjects report more intense emotions when answering questions in English, as opposed to when they answer the same questions in their native language

Attitudes: relatively enduring predispositions to respond to an object in a consistent fashion

Balanced scale: has the same number of positive response choices as negative choices

Comparative scale: respondents are asked to evaluate or rate a brand, a product, an object, a concept, or a person relative to other brands, products, objects, concepts, or individuals or to an ideal item

Constant sum scale: respondents allocate points among various alternatives so the total sums to a specified amount designated by the researcher

Correlation: the degree to which one variable changes with another

Factor analysis: analysis that reduces a larger number of items into a smaller subset of factors based on similarity

Graphical rating scale: noncomparative scale that allows respondents to place a response anywhere on a continuous line

Halo effect: when respondents have an overall feeling about the topic being surveyed and that overall perception influences their response so that all of the

answers are relatively close to the same

Itemized rating scale: respondents choose a response from a select number of items or categories

Likert scale: series of statements to which respondents indicate their level of agreement or disagreement with the statement

Multidimensional scale: measures multiple dimensions or facets of a concept, an idea, or an attitude

Noncomparative scale: respondents make judgments about a brand, a product, an object, a concept, or a person without reference to another item or an ideal item

Paired comparison scale: respondents choose one of two items in a set based on a specific criterion or attribute

Q-sort: comparative technique whereby respondents rank a set of objects into a prespecified number of categories along a particular attribute or criterion

Rank-order scale: respondents compare two or more objects, concepts, or persons by ranking them in some type of ordered sequence

Scaling: process of assigning numerical values or properties to subjective or abstract concepts

Semantic differential scale: involves a finite number of choices anchored by dichotomous words or phrases

Stapel scale: uses only one anchor and both positive and negative options

Unbalanced scale: response categories are weighted toward one of the anchors, either positive or negative

Unidimensional scale: measures only one attribute or concept

Critical Thinking Exercises

1. Suppose you worked for a firm that sold products to other businesses over the Internet as well as through an external sales force. What types of attitudes might be important to assess if you surveyed your customers? Identify at least five separate attitudes that would be relevant to your firm.
2. Using the five attitude constructs you developed in response to Question 1, identify whether each attitude would likely be unidimensional or

multidimensional. If multidimensional, what dimensions might be part of the overall attitude? Think carefully about the attitude in question as you do so, and justify your recommendations.

3. Critique the following scale in terms of the characteristics of a good scale that were discussed in the chapter. Make certain that you evaluate both the scale item and its response categories. Is a unidimensional scale appropriate to assess the beauty salon? Why or why not?
   Please rate Debbie's Beauty Salon using the following scale:

Outstanding	Excellent	Very Good	Good	Fair	Poor	The Worst

4. Compare and contrast rank-order scales with itemized rating scales. What are the advantages and disadvantages of each type of scale relative to one another?
5. Using examples other than those described in the text, identify two specific research scenarios in which a Q-sort could be used. Explain what objects would be sorted and what labels would be assigned to each category. How could the results of the research be used by marketers? Should respondents be allowed to select as many items for each category as they want? Why or why not?
6. You have been asked to develop a scale that assesses the relative importance of factors that influence the purchase of a smartphone. Begin by listing the product attributes or other factors that you believe could be influential. If the choice was between a ranking scale, a paired comparison scale, and a constant sum scale, which would you use and why? If the choice was between an itemized rating scale, a Likert scale, and a semantic differential scale, which would you use and why? Develop the scale type you ultimately would select given all possibilities, and defend your recommendation.
7. A sample of 200 individuals completed a series of paired comparison statements to assess preference for four national fast-food chain restaurants specializing in hamburgers. Interpret the results shown below and present them in a table, using percentages. Then create a graph with the percentages. The number of individuals selecting each restaurant for a paired comparison is shown in parentheses.

(160) Sonic	(40) McDonald's

(84) Wendy's	(116) McDonald's
(94) Sonic	(106) Burger King
(78) Wendy's	(122) Burger King
(134) Sonic	(66) Wendy's
(92) Burger King	(108) McDonald's

8. If a key research objective sought to develop brand image profiles for a local bank and a key competitor, what type of scale should be used? Why? Create a multi-item scale that measures a bank's brand image. Defend the scale category choices you made in terms of number of categories, balanced versus unbalanced scales, “no opinion” and “don't know” options, forced versus nonforced choices, and type of category description.
9. Create a Likert scale or an itemized rating scale to determine student attitudes toward the university bookstore. Defend the scale category choices you made in terms of number of categories, balanced versus unbalanced scales, “no opinion” and “don't know” options, forced versus nonforced choices, and type of category description.
10. Given the relative advantages and disadvantages associated with providing a “don't know” response option, would you recommend its inclusion in a question that asks people who live within a given zip code to rate the effectiveness of the local school board? Why or why not?

Continuing Case Study: Lakeside Grill

Rather than develop their own questions to measure the quality of service at Lakeside Grill, the student group wanted to see if there was a scale that had already been established. The students were ecstatic when Brooke found a scale called SERVQUAL, which was developed by Leonard Berry and his associates.[19] After examining the scale, the team chose 12 questions they thought would be good measures of service at Lakeside Grill. The scales measured empathy, reliability, and tangibles. Empathy is the degree to which employees give attention to customers and understand their needs; reliability

measures the degree to which Lakeside Grill is responsible and can be depended to perform as promised; and the tangible aspect relates to the degree to which the facilities and employees are visually appealing. The questions are listed in Figure 10.31. The students pretested the scale using college students.

Figure 10.31 Lakeside Grill Pretest Questions

- Employees of Lakeside Grill give you individual attention. (E1)
- Lakeside Grill's facility is visually appealing. (T1)
- Lakeside Grill gets your order right the first time. (R1)
- Employees of Lakeside Grill understand your specific needs. (E2)
- Lakeside Grill insists on error-free customer service. (R2)
- Lakeside Grill's employees are well dressed and appear neat. (T2)
- Lakeside Grill has employees who give you personal attention. (E3)
- Tables and place settings are visually appealing at Lakeside Grill. (T3)
- Lakeside Grill has convenient hours for all of its customers. (E4)
- When Lakeside Grill promises to have food ready by a certain time, it does so. (R3)
- The exterior appearance of Lakeside Grill is attractive and inviting. (T4)
- When you have a problem, Lakeside Grill shows an interest in correcting it. (R4)

Scales → Empathy (E)
Reliability (R)
Tangibles (T)

The students used SPSS to run a Cronbach's alpha test for the three scales. Complete results are located at http://www.sagepub.com/clowess in the PDF file "Chapter 10 Lakeside Grill Output." The raw data are titled "Chapter 10 Lakeside Grill Data." The empathy scale had an initial Cronbach's alpha score of .743. Dropping the fourth question, which dealt with convenient hours, improved the Cronbach's alpha score to .806. The initial Cronbach's alpha score for the tangible scale was .868. The initial score for the reliability scale was .765. The score could not be improved by dropping any questions. So the group decided to keep all of the questions in that scale.

Critique Questions:

1. Was the decision to use students to pretest the questions dealing with service quality a good decision? Why or why not?
2. How would you evaluate the Cronbach's alpha scores for each of the three scales? Would you agree with the final decision to drop the question about convenient hours? Why or why not?
3. What are the advantages and disadvantages of using these scales on

the group's survey of Lakeside Grill versus the students creating their own questions?

Marketing Research Portfolio

The “Marketing Research Portfolio” for Chapter 10 requires that students first download a questionnaire and data set from http://www.sagepub.com/clowess. Using the questionnaire, students are asked to identify the specific type of comparative and noncomparative scales used in the survey. Next, they are asked to critically evaluate the scales used in the survey against the factors used when selecting a scale. Finally, students use SPSS to run scale reliabilities on the data set, and check to see whether any scale's reliability can be improved by dropping one or more items.

Student Study Site

Visit the Student Study Site at http://www.sagepub.com/clowess to access the following additional materials:

- eFlashcards
- Web Quizzes
- SAGE Journal Articles
- Web Resources

http://dx.doi.org/10.4135/9781483384726.n10