

Bringing Research to Life

They boarded the sleek corporate jet in Palm Beach and were taken aft to meet with the general manager of MindWriter, who was seated at a conference table that austere held one sheaf of papers and a white telephone.

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

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

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

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

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

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

"When we held focus groups with your customers, they continually referred to the need for your product service to 'meet expectations' or 'exceed expectations.' The hundredth time we heard this we realized . . ."

"It's our company credo. 'Under-promise and exceed expectations.'"

"Well, virtually none of the scales developed for customer satisfaction deal with expectations. We want a scale that ranges in five steps from 'Met few expectations' to 'Exceeded expectations,' but we don't know what to name the in-between intervals so that the psychological spacing is equal between increments. We think 'Met many expectations' and 'Met most expectations' and 'Fully met expectations' will be OK, but we want to be sure."

"You are not being fussy here, are you, Jason?"

"No. Because of the way you are running your service operation, we want great precision and reliability."

"Justify that, please, Myra."

"Well, Jean-Claude, besides setting up our own repair force, we have contracted with an outside organization to provide repairs in certain areas, with the intention after six months of comparing the performance of the inside and outside repair organizations and giving the future work to whoever performs better. We feel that such an important decision, which involves the job security of MindWriter employees, must have full credibility."

"I can accept that. Good." The manager scribbled his signature on the contract. "You'll receive this contract in three days, after it has wended its way past the paper pushers. Meantime, we'll settle for a handshake. Nice job, so far, Myra. You seem to have gotten a quick start with MindWriter. Congratulations, Jason."

"We can put them down in Orlando," said the steward.

"No," said Jean-Claude. "We are only five minutes out. Turn the plane around and put these folks out where they got on. They can start working this afternoon . . . Gosh, is that the beach out there? It looks great. I've got to get some sun one of these days."

"You do look pale," said Myra, sympathetically.

"*Fais gaffe, tu m'fais mal!*" he muttered under his breath.

The Nature of Measurement Scales

When you develop measurement questions for your research study, you often may choose between standardized scales and custom-designed ones. When what you measure is concrete (for example the length of an assembly line), we usually choose a standardized measure (like measuring the assembly line with an electronic range finder or tape measure). When what we want to measure is a more abstract and complex construct (like customer attitudes about a product service program), standardized measures may neither exist nor provide a close enough fit to a particular manager's scenario. In these situations, developing a customized scale to measure the construct is the only option. Otherwise, we are left measuring a construct with a tool designed for something else. This would be like measuring the length of the assembly line with our forearm instead of visible laser beam technology.

This chapter covers procedures that will help you understand measurement scales so that you might select or construct measures that are appropriate for your research. We concentrate here on the problems of measuring more complex constructs, like attitudes and opinions.

Scaling Defined

Scaling is a "procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some of the characteristics of numbers to the properties in question."¹

What Is Scaled? Procedurally, we assign numbers to indicants of the properties of objects. Thus, one assigns a number scale to the various levels of heat and cold and calls it a thermometer. If you want to measure the temperature of the air, you know that a property of temperature is that its variation leads to an expansion or contraction of mercury. A glass tube with mercury provides an indicant of temperature change by the rise or fall of the mercury in the tube.

In another context, you might devise a scale to measure the durability (property) of paint. You secure a machine with an attached scrub brush that applies a predetermined amount of pressure as it scrubs. You then count the number of brush strokes that it takes to wear through a 10-mil thickness of paint. The scrub count is the indicant of the paint's durability. Or you may judge a person's supervisory capacity (property) by asking a peer group to rate that person on various questions (indicants) that you create.

Scale Selection

Scaling may be reviewed in several ways, but here we cover those approaches that are of greatest value for management research.² Selection or construction of a measurement scale requires decisions in six key areas:

- Study objective
- Response form
- Degree of preference
- Data properties
- Number of dimensions
- Scale construction

MANAGEMENT



Study Objective Researchers face two general study objectives:

- To measure certain characteristics of the respondents who complete the study.
- To use respondents as judges of the objects or indicants presented to them.

Assume you've been contracted by the city of Miro Beach to conduct a study supposedly of voters' approval or disapproval of one or more regulatory programs. In the first type of study, your scale would measure the voters' political orientation as conservative or liberal. You might combine each person's answers to form an indicator of that person's political orientation. The emphasis in this first study objective is on measuring attitudinal differences among people. With the second study objective, you might use the same data but in this case you are truly interested in how satisfied people are with different governmental programs. In this study objective, your true interest is in the differences in the acceptance level of one or more regulatory programs.

Response Form Measurement scales are of three types: *rating*, *ranking*, and *categorization*. A **rating scale** is used when respondents score an object or indicant without making a direct comparison to another object or attitude. For example, they may be asked to evaluate the styling of a new automobile on a five-point rating scale. **Ranking scales** constrain the study participant to make comparisons among two or more indicants or objects. Respondents may be asked to choose which one of a pair of cars has more attractive styling. They could also be asked to order the importance of comfort, ergonomics, performance, and price for the target vehicle. **Categorization** asks respondents to put themselves or property indicants in groups or categories. Asking auto show respondents to identify their gender or ethnic background or to indicate whether a particular prototype car design would attract a youthful or mature clientele would require a categorization response strategy.

Degree of Preference Measurement scales may involve *preference* measurement or *nonpreference* evaluation. In the former, each respondent is asked to choose the object he or she favors or the solution he or she would prefer. In the latter, respondents are asked to judge which object has more of some characteristic or which solution takes the most resources, without reflecting any personal preference toward objects or solutions.

Data Properties Measurement scales also may be viewed in terms of the data properties generated by each scale. Chapter 8 indicated that data are classified as nominal, ordinal, interval, or ratio. The assumptions underlying each data type determine how a particular measurement scale's data can be handled statistically.

Number of Dimensions Measurement scales are either *unidimensional* or *multidimensional*. With a **unidimensional scale**, one seeks to measure only one attribute of the respondent or object. One measure of employee potential is promotability. It is a single dimension. Several items may be used to measure this dimension and, by combining them into a single measure, a manager may place employees along a linear continuum of promotability. **Multidimensional scaling** recognizes that an object might be better described in an attribute space of n dimensions rather than on a unidimensional continuum. The employee promotability variable might be better expressed by three distinct dimensions—managerial performance, technical performance, and teamwork.

Scale Construction We can classify measurement scales by the methods used to build them. Five construction approaches are used in research practice:

- **Arbitrary:** A scale is custom-designed to measure a property or indicant.
- **Consensus:** Judges evaluate the items to be included.
- **Item analysis:** Measurement scales are tested with a sample of respondents.
- **Cumulative:** Scales are chosen for their conformity to a ranking of items with ascending and descending discriminating power.
- **Factoring:** Scales are constructed from intercorrelations of items from other studies.³

Arbitrary scales may measure the concepts for which they have been designed, but the researcher has no advance evidence of a particular scale's validity and reliability. Nevertheless, researchers commonly choose this construction approach. *Consensus scales* are developed by a panel of judges who evaluate the items to be included based on topical relevance and lack of ambiguity.

In *item analysis*, after administering the test, a total score is calculated for each scale. Individual items (a scale or part of a scale) are then analyzed to determine which best discriminate between persons or objects with high total scores and low total scores.

In the *cumulative* approach, the endorsement of an item that represents an extreme position results in the endorsement of all items of less extreme positions.

Finally, in *factoring* common factors account for the relationships. The relationships are measured statistically through factor analysis or cluster analysis.

The business researcher studies both the type of measurement scale and the scale's construction when selecting an appropriate scale. These topics form the basis for the remainder of the chapter.

Response Methods

In Chapter 8, we said that questioning is a widely used stimulus for measuring concepts and constructs. A manager may be asked his or her views concerning an employee. The response is, "a good machinist," "a troublemaker," "a union activist," "reliable," or "a fast worker with a poor record of attendance." These answers, because they represent such different frames of reference for evaluating the worker and thus lack comparability, would be of limited value to the researcher.

Two approaches improve the usefulness of such replies. First, various properties may be separated and the respondent asked to judge each specific facet. Here, the researcher would substitute several distinct questions for a single one. Second, the researcher can replace the free-response reply with structuring devices. To quantify dimensions that are essentially qualitative, rating or ranking scales are used.

MANAGEMENT



Rating Scales

One uses rating scales to judge properties of objects without reference to other similar objects. These ratings may be in such forms as "like-dislike," "approve-indifferent-disapprove," or other classifications using even more categories.

Number of Scale Points There is little conclusive support for choosing a three-point scale over scales with five or more points. Some researchers think that more points on a rating scale provide an opportunity for greater sensitivity of measurement and extraction of variance. The most widely used scales range from three to seven

points, but it does not seem to make much difference which number is used—with two exceptions.⁴ First, a larger number of scale points is needed to produce accuracy when using single-dimension versus multiple-dimension scales. Second, in cross-cultural measurement, the culture may condition respondents to a standard metric—a 10-point scale in Italy, for example.

Alternative Scales Examples of rating scales are shown in Exhibit 9-1. This exhibit amplifies the overview presented in this section.⁵ Later in the chapter, construction techniques for some commonly used rating scales are presented.

The **simple category scale** (also called a *dichotomous scale*) offers two mutually exclusive response choices. In Exhibit 9-1 they are yes and no but they could just as easily be important and unimportant, agree and disagree, or another set of discrete categories had the question been different. This response strategy is particularly useful for demographic questions or where a dichotomous response is adequate.

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

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

The **Likert scale** is the most frequently used variation of the *summated rating scale*. Summated scales consist of statements that express either a favorable or unfavorable attitude toward the object of interest. The respondent is asked to agree or disagree with each statement. Each response is given a numerical score to reflect its degree of attitudinal favorableness, and the scores may be totaled to measure the respondent’s attitude. In our example, the respondent chooses one of five levels of agreement. The numbers indicate the value to be assigned to each possible answer with 1 the least favorable impression of Internet superiority and 5 the most favorable. These values are normally not printed on the instrument but are shown in Exhibit 9-1 to indicate the scoring system. Between 20 and 25 properly constructed questions about an attitude object would be required for a reliable Likert scale.

Likert scales help us compare one person’s score with a distribution of scores from a well-defined sample group. This measurement scale is useful for a manager when the organization plans to conduct an experiment or undertake a program of change or improvement. The researcher can measure attitudes before and after the experiment or change, or judge whether the organization’s efforts have had the desired effects. This scale produces interval data.

The **semantic differential scale** measures the psychological meanings of an attitude object. Managers use this scale for brand image and other marketing studies of institutional images, political issues and personalities, and organizational studies. It is based on the proposition that an object can have several dimensions of connotative meaning. The meanings are located in multidimensional property space, called *semantic space*. The method consists of a set of bipolar rating scales, usually with seven

EXHIBIT 9-1 Sample Rating Scales

Simple Category Scale
(dichotomous)
data: nominal

"I plan to purchase a MindWriter laptop in the next 12 months."
 Yes
 No

Multiple Choice Single-Response Scale
data: nominal

"What newspaper do you read most often for financial news?"
 East City Gazette
 West City Tribune
 Regional newspaper
 National newspaper
 Other (specify: _____)

Multiple Choice Multiple-Response Scale (checklist)
data: nominal

"Check any of the sources you consulted when designing your new home:"
 Online planning services
 Magazines
 Independent contractor/builder
 Developer's models/plans
 Designer
 Architect
 Other (specify: _____)

Likert Scale Summated Rating
data: interval

"The Internet is superior to traditional libraries for comprehensive searches."
 STRONGLY AGREE (5) AGREE (4) NEITHER AGREE NOR DISAGREE (3) DISAGREE (2) STRONGLY DISAGREE (1)

Semantic Differential Scale
data: interval

Lands' End Catalog
 FAST _____ : _____ : _____ : _____ : _____ : _____ : SLOW
 HIGH QUALITY _____ : _____ : _____ : _____ : _____ : _____ : LOW QUALITY

Numerical Scale
data: ordinal or* interval

EXTREMELY FAVORABLE 5 4 3 2 1 EXTREMELY UNFAVORABLE
 Employee's cooperation in teams _____
 Employee's knowledge of task _____
 Employee's planning effectiveness _____

Continued

points, by which one or more respondents rate one or more concepts on each scale item. In the example in Exhibit 9-1, two sets of bipolar pairs are shown, one from the traditional source and one adapted to the research purpose. Based on the construction requirements discussed later, we might choose 10 scale items to score the "Lands' End Catalog."

EXHIBIT 9-1 Concluded

Multiple Rating List Scale
data: interval

"Please indicate how important or unimportant each service characteristic is:"

	IMPORTANT					UNIMPORTANT	
Fast reliable repair	7	6	5	4	3	2	1
Service at my location	7	6	5	4	3	2	1
Maintenance by manufacturer	7	6	5	4	3	2	1
Knowledgeable technicians	7	6	5	4	3	2	1
Notification of upgrades	7	6	5	4	3	2	1
Service contract after warranty	7	6	5	4	3	2	1

Fixed Sum Scale
data: ratio

"Taking all the supplier characteristics we've just discussed and now considering cost, what is their relative importance to you (dividing 100 units between):"

Being one of the lowest cost suppliers

All other aspects of supplier performance

Sum



Stapel Scale
data: ordinal or* interval

	(Company Name)			
+5	+5			+5
+4	+4			+4
+3	+3			+3
+2	+2			+2
+1	+1			+1
Technology Leader	Exciting Products		World-Class Reputation	
-1	-1		-1	-1
-2	-2		-2	-2
-3	-3		-3	-3
-4	-4		-4	-4
-5	-5		-5	-5

Graphic Rating Scale
data: ordinal or* interval or ratio

"How likely are you to recommend CompleteCare to others?" (place an X at the position along the line that best reflects your judgment)

VERY LIKELY VERY UNLIKELY

(alternative with graphic)

* In chapter 8 we noted that researchers differ in the ways they treat data from certain scales. If you are unable to establish the linearity of the measured variables or you cannot be confident that you have equal intervals, it is proper to treat data from these scales as ordinal.

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

rating. It is a standardized technique that is easily repeated but escapes many problems of response distortion found with more direct methods.

Numerical scales have equal intervals that separate their numeric scale points. The verbal anchors serve as the labels for the extreme points. Numerical scales are often 5-point scales, as shown in the exhibit, but may have 7 or 10 points. The respondent writes a number from the scale next to each item. If numerous questions about employee performance were included in the example, the scale would provide both an absolute measure of importance and a relative measure (ranking) of the various items rated. The scale's linearity, simplicity, and production of ordinal or interval data make it popular for managers and researchers.

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

A scale that helps the researcher discover proportions is the **fixed sum scale**. In the example, two categories are presented that must sum to 100. Up to 10 categories may be used, but both respondent precision and patience suffer when too many stimuli are proportioned and summed. A respondent's ability to add is also taxed in some situations; thus this is not a response strategy that can be effectively used with children or the uneducated. The advantage of the scale is its compatibility with percent (100 percent) and the fact that continuous data (versus discrete categories) can be compared for the alternatives. The scale is used to record attitudes, behavior, and behavioral intent. It produces interval data.

The **stapel scale** is used as an alternative to the semantic differential, especially when it is difficult to find bipolar adjectives that match the investigative question. In the example in Exhibit 9-1 there are three attributes of corporate image. The scale is composed of the word (or phrase) identifying the image dimension and a set of 10 response categories for each of the three attributes. Fewer response categories are sometimes used. Respondents select a plus number for the characteristic that describes the named company. The more accurate the description, the larger is the positive number. Similarly, the less accurate the description, the larger is the negative number chosen. Ratings range from ± 5 to -5 , very accurate to very inaccurate. Like the semantic differential, stapel scales usually produce interval data.

The **graphic rating scale** was created to enable researchers to discern fine differences. Theoretically, an infinite number of ratings is possible if the respondent is sophisticated enough to differentiate and record them. The respondent checks his or her response at any point along a continuum. Usually, the score is a measure of length (millimeters) from either end point. The results are usually treated as interval data. The difficulty is in coding and analysis. This response strategy requires more time than scales with predetermined categories. Other graphic rating scales use pictures, icons, or other visuals to communicate with the rater and represent a variety of data types. Graphic scales are often used with children, whose more limited vocabulary prevents the use of scales anchored with words.

Errors to Avoid with Rating Scales The value of rating scales for measurement purposes depends on the assumption that a person can and will make good judgments. Before accepting respondents' ratings, we should consider their tendencies to make errors of three types: (1) leniency, (2) central tendency, and (3) halo effect.

Leniency. The error of leniency occurs when a respondent is either an "easy rater" or a "hard rater." The latter is an error of *negative leniency*. Raters are inclined to score

people higher whom they know well and with whom they are ego involved. There is also the opposite—where acquaintances are rated lower because one is aware of the tendency toward positive leniency and attempts to counteract it. A way to deal with *positive leniency* is to design the rating scale to anticipate it. An example might be an asymmetrical scale that has only one unfavorable descriptive term and four favorable terms (poor—fair—good—very good—excellent). The scale designer expects that the mean ratings will be near “good” and that there will be a symmetrical distribution about that point.

Central Tendency. Raters are reluctant to give extreme judgments, and this fact accounts for the error of **central tendency**. This is most often seen when the rater does not know the object or property being rated. To counteract this type of error try the following:

MANAGEMENT



- Adjust the strength of descriptive adjectives.
- Space the intermediate descriptive phrases farther apart.
- Provide smaller differences in meaning between the steps near the ends of the scale than between the steps near the center.
- Use more points in the scale.

Halo. The **halo effect** is the systematic bias that the rater introduces by carrying over a generalized impression of the subject from one rating to another. You expect the student who does well on the first question of an examination to do well on the second. You conclude a report is good because you like its form, or you believe someone is intelligent because you agree with him or her. Halo is a pervasive error. It is especially difficult to avoid when the property being studied is not clearly defined, not easily observed, not frequently discussed, involves reactions with others, or is a trait of high moral importance.⁷ One way to counteract the halo effect is to rate one trait at a time for all subjects or to have one trait per page.

Rating scales are widely used in management research and generally deserve their popularity. The results obtained with careful use compare favorably with other methods.

Ranking Scales

In ranking scales, the subject directly compares two or more objects and makes choices among them. Frequently, the respondent is asked to select one as the “best” or the “most preferred.” When there are only two choices, this approach is satisfactory, but it often results in “ties” when more than two choices are found. For example, assume respondents are asked to select the most preferred among three or more models of a product. In response, 40 percent choose model A, 30 percent choose model B, and 30 percent choose model C. Which is the preferred model? The analyst would be taking a risk to suggest that A is most preferred. Perhaps that interpretation is correct, but 60 percent of the respondents chose some model other than A. Perhaps all B and C voters would place A last, preferring either B or C to it. This ambiguity can be avoided by using some of the techniques described in this section.

MANAGEMENT



Using the **paired-comparison scale**, the respondent can express attitudes unambiguously by choosing between two objects. Typical of paired comparisons would be the sports car preference example in Exhibit 9–2. The number of judgments required in a paired comparison is $[(n)(n - 1)/2]$, where n is the number of stimuli or objects to be judged. When four cars are evaluated, the respondent evaluates six paired comparisons $[(4)(3)/2 = 6]$.

EXHIBIT 9-2 Ranking Scales

Paired-Comparison Scale

data: ordinal

"For each pair of two-seat sports cars listed, place a check beside the one you would most prefer if you had to choose between the two."

- | | |
|---|---|
| <input type="checkbox"/> BMW Z3 | <input type="checkbox"/> Chevrolet Corvette |
| <input type="checkbox"/> Porsche Boxster | <input type="checkbox"/> Porsche Boxster |
| <input type="checkbox"/> Chevrolet Corvette | <input type="checkbox"/> Porsche Boxster |
| <input type="checkbox"/> BMW Z3 | <input type="checkbox"/> Dodge Viper |
| <input type="checkbox"/> Chevrolet Corvette | <input type="checkbox"/> Dodge Viper |
| <input type="checkbox"/> Dodge Viper | <input type="checkbox"/> BMW Z3 |

Forced Ranking Scale

data: ordinal

"Rank the radar detection features in your order of preference. Place the number 1 next to the most preferred, 2 by the second choice, and so forth."

- User programming
- Cordless capability
- Small size
- Long-range warning
- Minimal false alarms

Comparative Scale

data: ordinal

"Compared to your previous mutual fund's performance, the new one is:"

- | | | | | |
|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| SUPERIOR | | ABOUT THE SAME | | INFERIOR |
| <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| 1 | 2 | 3 | 4 | 5 |

SNAPSHOT

Mastering Leadership in Education

In an attempt to respond to the less-than-stellar performance of their students on standardized tests, many states are mandating new educational standards for teachers. Ohio is one such state. Starting with year 2002 graduates of education programs, all teachers of kindergarten through high school will need to earn a master's degree within five to seven years of licensure in order to maintain their teaching certification. Wittenberg University, a nationally ranked, private liberal arts institution in Ohio, has been training teachers for 150 years through a bachelor of arts in education. The faculty in its education department recently mailed a survey to more than 2,000

teachers in its five-county market area to determine the attractiveness of Wittenberg as a source for the required master's degree. This mail survey generated nearly 800 responses and indicated the market is receptive to attending Wittenberg for the required master's degree in education. Respondents' enthusiasm was tempered only by concerns about price. Looking at the instrument provided with the case "Mastering Teacher Leadership" in the Cases section of the text, did the survey designers scale the items to be measured correctly?

www.wittenberg.edu

EXHIBIT 9-3 Response Patterns of 200 Union Members' Paired Comparisons on Five Suggestions for Bargaining Proposal Priorities

Paired-comparison data may be treated in several ways. If there is substantial consistency, we will find that if *A* is preferred to *B*, and *B* to *C*, then *A* will be consistently preferred to *C*. This condition of transitivity need not always be true but should occur most of the time. When it does, take the total number of preferences among the comparisons as the score for that stimulus. Assume a union bargaining committee is considering five major demand proposals. The committee would like to know how the union membership ranks these proposals. One option would be to ask a sample of the members to pair-compare the personnel suggestions. With a rough comparison of the total preferences for each option, it is apparent that *B* is the most popular.

	Suggestion				
	A	B	C	D	E
A	—	164*	138	50	70
B	36	—	54	14	30
C	62	146	—	32	50
D	150	186	168	—	118
E	130	170	150	82	—
Total	378	666	510	178	268
Rank order	3	1	2	5	4
M_p	0.478	0.766	0.610	0.278	0.368
Z_j	-0.060	0.730	0.280	-0.590	-0.340
R_j	0.530	1.320	0.870	0.000	0.250

Interpret this cell, 164 members preferred suggestion *B* (column) to suggestion *A* (row).

In another example we might compare two bargaining proposals available to union negotiators (see Exhibit 9-3). Generally, there are more than two stimuli to judge, resulting in a potentially tedious task for respondents. If 15 suggestions for bargaining proposals are available, 105 paired comparisons would be made.

Reducing the number of comparisons per respondent without reducing the number of objects can lighten this burden. You can present each respondent with only a sample of the stimuli. In this way, each pair of objects must be compared an equal number of times. Another procedure is to choose a few objects that are believed to cover the range of attractiveness at equal intervals. All other stimuli are then compared to these few standard objects. If 36 automobiles are to be judged, four may be selected as standards and the others divided into four groups of eight each. Within each group, the eight are compared to each other. Then the 32 are individually compared to each of the four standard automobiles. This reduces the number of comparisons from 630 to 240.

Paired comparisons run the risk that respondents will tire to the point that they give ill-considered answers or refuse to continue. Opinions differ about the upper limit, but five or six stimuli are not unreasonable when the respondent has other questions to

MANAGEMENT



answer. If the data collection consists only of paired comparisons, as many as 10 stimuli are reasonable.

While a paired comparison provides ordinal data, there are methods for converting it to interval data. The Law of Comparative Judgment involves converting the frequencies of preferences (such as in Exhibit 9-3) into a table of proportions that are then transformed into a Z matrix by referring to the table of areas under the normal curve.⁸ Guilford's *composite-standard* method is another alternative.⁹

The **forced ranking scale** shown in Exhibit 9-2 lists attributes that are ranked relative to each other. This method is faster than paired comparisons and is usually easier and more motivating to the respondent. With five items, it takes 10 paired comparisons to complete the task, and the simple forced ranking of five is easier. Also, ranking has no transitivity problem where A is preferred to B, and B to C, but C is preferred to A.

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

Often the manager or researcher is interested in benchmarking. This calls for a standard by which other programs, processes, brands, points of sale, or people can be compared. The **comparative scale** is ideal for such comparisons if the respondents are familiar with the standard. In the Exhibit 9-2 example, the standard is the respondent's previous mutual fund. The new fund is being assessed relative to it. The provision to compare yet other funds to the standard is not shown in the example but is nonetheless available to the researcher.

Some researchers treat the data produced by comparative scales as interval data since the scoring reflects an interval between the standard and what is being compared. We would treat the rank or position of the item as ordinal data unless the linearity of the variables in question could be supported.

None of the ranking methods covered is particularly useful when there are many items. The method of **successive intervals** is sometimes used to sort the items (usually one per card) into piles or groups representing a succession of values. From the sort, an interval scale can then be developed.¹⁰ This procedure is not used frequently and then only in unique studies.

Measurement Scale Construction

Earlier, we discussed scales by the techniques used to construct them. Of the five techniques, three are used frequently: the *arbitrary* approach, *item analysis*, and *factoring*. They are emphasized in this section along with a preview of *multivariate* scales (described in more detail in Chapter 19). *Consensus* and *cumulative* methods receive less attention because they are time-consuming to construct or have fewer management applications. They are briefly mentioned because of their influence on current methods.

Arbitrary Scaling

We design **arbitrary scales** by collecting several items that we believe are unambiguous and appropriate to a given topic. Some are chosen for inclusion in the instrument. To illustrate, consider a company image study. We choose a sample of items that we believe are the components of company image:

How do you regard (Company X's) reputation?

- | | | |
|------------------------------------|-----------|------------|
| 1. As a place to work? | Bad _____ | Good _____ |
| 2. As a sponsor of civic projects? | Bad _____ | Good _____ |
| 3. For ecological concern? | Bad _____ | Good _____ |
| 4. As an employer of minorities? | Bad _____ | Good _____ |

We might score each of these from 1 to 5, depending on the degree of favorableness reported. The results may be studied in several ways. Totals may be made by individual items, by company, by companies as places to work, for ecological concern, and so on. Totals for each company or for individuals may be calculated to determine how they compare to others. Based on a total for these four items, each company would receive from 4 to 20 points from each respondent. These data may also be analyzed from a respondent-centered point of view. Thus, we might use the attitude scores of each individual to study differences among individuals.

Arbitrary scales are easy to develop, inexpensive, and can be designed to be highly specific. They provide useful information and are adequate if developed skillfully. There are also weaknesses. The design approach is subjective. The researcher's insight and ability offer the only assurance that the items chosen are a representative sample of the universe of content (the totality of what constitutes "company image"). We have no evidence that respondents will view all items with the same frame of reference.

While arbitrary scales are often used, there has been a great effort to develop construction techniques that overcome some of their deficiencies. An early attempt was consensus scaling.

Consensus Scaling

Consensus scaling requires items to be selected by a panel of judges and then evaluated on (1) relevance to the topic area, (2) potential for ambiguity, and (3) the level of attitude they represent. A widely known form of this approach is the **Thurstone equal-appearing interval scale**. Also known as the *Thurstone scale*, this approach resulted in an interval rating scale for attitude measurement. Often 50 or more judges evaluate a large number of statements expressing different degrees of favorableness toward an

Assume you are asked by Galaxy Department Stores to study the shopping habits and preferences of teen girls. Galaxy is seeking a way to compete with specialty stores that are far more successful in serving this market segment. Galaxy is considering the construction of an intrastore boutique catering to these teens. What measurement issues would determine your construction of measurement scales?



object. There is one statement per card. The judges sort each card into 1 of 11 piles representing their evaluation of the degree of favorableness that the statement expresses. The judge's agreement or disagreement with the statement is not involved. Of the 11 piles, 3 are identified to the judges by labels of "favorable" and "unfavorable" at the extremes and "neutral" at the midpoint. The eight intermediate piles are unlabeled to create the impression of equal-appearing intervals between the three labeled positions.

This method of scale construction is rarely used in applied management research these days. Its cost, time, and staff requirements make it impractical. The importance of this historic method, however, is its influence on the Likert and semantic differential scales.

Item Analysis Scaling

Item analysis scaling is a procedure for evaluating an item based on how well it discriminates between those persons whose total score is high and those whose total score is low. The most popular scale using this approach is the summated or Likert scale.

Item analysis involves calculating the mean scores for each scale item among the low scorers and high scorers. The item means between the high-score group and the low-score group are then tested for significance by calculating t values. Finally, the 20 to 25 items that have the greatest t values (significant differences between means) are selected for inclusion in the final scale.¹¹

Likert-type scales are relatively easy to construct compared to the equal-appearing interval scale.¹² The first step is to collect a large number of statements that meet two criteria: (1) Each statement is believed to be relevant to the attitude being studied. (2) Each is believed to reflect a favorable or unfavorable position on that attitude. People similar to those who are going to be studied are asked to read each statement and to state the level of their agreement with it, using a five-point scale. A scale value of 1 might indicate a strongly unfavorable attitude; 5, a strongly favorable attitude (see Exhibit 9-1).

Each person's responses are then added to secure a total score. The next step is to array these total scores and select some portion representing the highest and lowest total scores, say, the top 25 percent and the bottom 25 percent. These two extreme groups represent people with the most favorable and least favorable attitudes toward the topic being studied. The extremes are the two criterion groups by which we evaluate individual statements. Through a comparative analysis of response patterns to each statement by members of these two groups, we learn which statements consistently correlate with low favorability and which correlate with high favorability attitudes.

This procedure is illustrated in Exhibit 9-4. In evaluating response patterns of the high and low groups to the statement "I consider my job exciting," we secure the results shown. After finding the t values for each statement, we rank-order them and select those statements with the highest t values. As an approximate indicator of a statement's discrimination power, Edwards suggests using only those statements whose t value is 1.75 or greater, provided there are 25 or more subjects in each group.¹³ To safeguard against response-set bias, we should word approximately one-half of the statements to be favorable and word the other half to be unfavorable.

The Likert scale has many advantages that account for its popularity. It is easy and quick to construct. Each item that is included has met an empirical test for discriminating ability. Since respondents answer each item, it is probably more reliable and it provides a greater volume of data than many other scales.

EXHIBIT 9-4 Evaluating a Scale Statement by Item Analysis

Response Categories	Low Total Score Group				High Total Score Group			
	X	f	fX	fX^2	X	f	fX	fX^2
Strongly agree	5	3	15	75	5	22	110	550
Agree	4	4	16	64	4	30	120	480
Undecided	3	29	87	261	3	15	45	135
Disagree	2	22	44	88	2	4	8	16
Strongly disagree	1	15	15	15	1	2	2	2
Total		73	177	503		73	285	1,183
		n_L	ΣX_L	ΣX_L^2		n_H	ΣX_H	ΣX_H^2

Steps:

1. For the statement "I consider my job exciting," we select the data from the bottom 25 percent of the distribution (low total score group) and the top 25 percent (high total score group). There are 73 people in each group. The remaining 50 percent in the middle of the distribution is not considered for this analysis. For each of the response categories, the scale's value (X) is multiplied by the frequency or number of respondents (f) who chose that value. These values produce the product (fX). This number is then multiplied by X (fX^2). For example, there are 3 respondents in the low score group who scored a 5 (strongly agreed with the statement): (fX) = $5 \times 3 = 15$; (fX^2) = $15 \times 5 = 75$.
2. The frequencies, products, and squares are summed.
3. A mean score for each group is computed.
4. Deviation scores are computed, squared, and summed as required by the formula.
5. The data are tested in a modified t -test that compares the high and low scoring groups for the item. Notice the mean scores in the numerator of the formula.
6. The calculated value is compared with a criterion, 1.75. If the calculated value (in this case, 8.92) is equal to or exceeds the criterion, the statement is said to be a good discriminator of the measured attitude. (If it is less than the criterion, we would consider it a poor discriminator of the target attitude and delete it from the measuring instrument.) We then select the next item and repeat the process.

Cumulative Scaling

Total scores on **cumulative scales** have the same meaning. Given a person's total score, it is possible to estimate which items were answered positively and negatively. A pioneering scale of this type was the **scalogram**. Scalogram analysis is a procedure for determining whether a set of items forms a unidimensional scale.¹⁴ A scale is unidimensional if the responses fall into a pattern in which endorsement of the item reflecting the extreme position results also in endorsing all items that are less extreme.

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

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

Respondents indicate whether they agree or disagree. If these items form a unidimensional scale, the response patterns will approach the ideal configuration shown in Exhibit 9-5.

A score of 4 indicates all statements are agreed upon and represents the most favorable attitude. Persons with a score of 3 should disagree with item 2 but agree with all

EXHIBIT 9-5 Ideal Scalogram Response Pattern

Item				Respondent Score
2	4	1	3	
X	X	X	X	4
—	X	X	X	3
—	—	X	X	2
—	—	—	X	1
—	—	—	—	0

X = Agree

— = Disagree

others, and so on. According to scalogram theory, this pattern confirms that the universe of content (attitude toward the appearance of this running shoe) is scalable.

The scalogram and similar procedures for discovering underlying structure are useful for assessing behaviors that are highly structured, such as social distance, organizational hierarchies, and evolutionary product stages.¹⁵ Although used less often today, the scalogram retains potential for managerial applications.

Factor Scaling

Factor scales include a variety of techniques that have been developed to address two problems: (1) how to deal with the universe of content that is multidimensional and (2) how to uncover underlying (latent) dimensions that have not been identified by exploratory research.

S N A P S H O T

A Survey of Controversy: RU486

The Food and Drug Administration's approval of the sale of mifepristone (RU486), "the first dedicated medical abortion pill regimen," on September 28, 2000, heralded a period of some concern in medical practices as well as private and public health facilities. Fearing a "dramatic transformation of the abortion landscape," many physicians claimed they would stay on the sidelines. One year later, a landmark study by the Henry J. Kaiser Family Foundation (KFF) reveals whether those fears were realized.

KFF hired Princeton Survey Research Associates (PSRA) to conduct a phone survey of 790 health care providers, including 595 gynecologists and 195 family practitioners, internists, and general practitioners between May and August 2000. PSRA randomly drew the sample of physicians from the American Medical Association's master file. Doctors were asked to reveal the degree to which they had performed surgical abortion in the previous five years or their reasons for not providing this service (personal convictions, hospital policy, etc.). Each interview also included

questions to measure physician familiarity with medical abortion regimen, as well as its perceived safety and effectiveness. Finally, interviewers asked physicians whether they had prescribed mifepristone since its FDA approval, whether they had previously participated in clinical trials of the drug during its FDA approval process, their reasons for prescribing or not prescribing mifepristone, and their future intentions for prescribing mifepristone plus their reasons for acting as they predicted.

Given the controversial nature of the subject and physicians' expressed concerns, what measurement issues should have been considered and which scales would have been appropriate for this first-year benchmark study? To see the measurement questions used, see the KFF website.

www.kff.org

www.psra.com

These techniques are designed to intercorrelate items so their degree of interdependence may be detected. There are many approaches that the advanced student will want to explore, such as latent structure analysis (of which the scalogram is a special case), factor analysis, cluster analysis, and metric and nonmetric multidimensional scaling. We limit the discussion in this section to the semantic differential (SD), which is based on factor analysis.¹⁶

Osgood and his associates developed the *semantic differential method* to measure the psychological meanings of an object to an individual.¹⁷ They produced a long list of adjective pairs useful for attitude research. Searching *Roget's Thesaurus* for such adjectives, they located 289 pairs. These were reduced to 76 pairs that were formed into rating scales. They chose 20 concepts that evoked the psychological meanings they wished to probe. The concepts from this historical study illustrate the wide applicability of the technique to persons, abstract concepts (such as leadership), events, institutions, and physical objects.¹⁸

By factor analyzing the data, they concluded that semantic space is multidimensional rather than unidimensional. Three factors contributed most to meaningful judgments by respondents: (1) evaluation, (2) potency, and (3) activity. The evaluation dimension usually accounts for one-half to three-fourths of the extractable variance. (The evaluation dimension is the only dimension possessed by Likert scales.) Potency and activity are about equal and together account for a little over one-fourth of the extractable variance. Occasionally, the potency and activity dimensions combine to form "dynamism." Results of the *Thesaurus* study are shown in Exhibit 9-6.

The SD scale should be adapted to each research problem. SD construction involves the following steps.

1. Select the concepts. The concepts are nouns, noun phrases, or nonverbal stimuli such as visual sketches. Concepts are chosen by judgment and reflect the nature of the investigative question. In the MindWriter study, one concept might be "Call Center accessibility." Or in a study to evaluate multiple candidates for an executive position in an industry association, the concept might be a candidate, "Darnell Williams."

2. Select the original bipolar word pairs or pairs you adapt to your needs. If the traditional Osgood items are used, several criteria guide your selection. The first is the factor(s) composition.

- You need at least three bipolar pairs for each factor to use evaluation, potency, and activity. Scores on these individual items should be averaged, by factor, to improve their test reliability.
- The scale must be relevant to the concepts being judged. Choose adjectives that allow connotative perceptions to be expressed. Irrelevant concept-scale pairings yield neutral midpoint values that convey little information.
- Scales should be stable across subjects and concepts. A pair such as "large-small" may be interpreted by some to be denotative when judging a physical object such as an "automobile" but may be used connotatively in judging abstract concepts such as "quality management."
- Scales should be linear between polar opposites and pass through the origin. A pair that fails this test is "rugged-delicate," which is nonlinear on the evaluation dimension. When used separately, both adjectives have favorable meanings.¹⁹

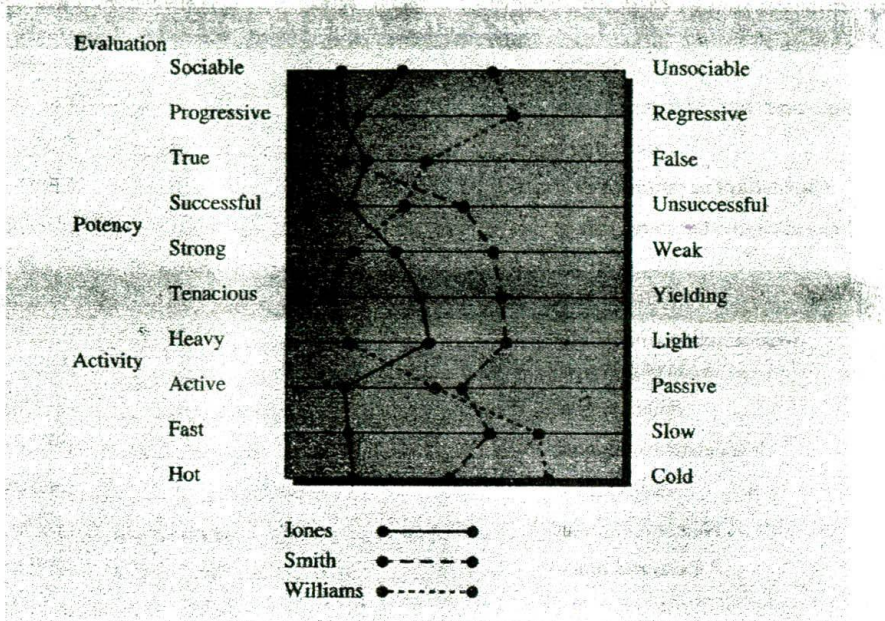
In Exhibit 9-7 we see the scale being used by a panel of corporate leaders to rate candidates for an industry leadership position. The selection of concepts in this case is simple; there are three candidates, plus a fourth—the ideal candidate.

We provide an illustration of factor analysis and multidimensional scaling in Chapter 19.

MANAGEMENT

Tip

EXHIBIT 9-8 Graphic Representation of SD Analysis



The nature of the problem determines the selection of dimensions and bipolar pairs. Since the person who wins this position must influence business leaders, we decide to use all three factors. The candidate must deal with many people, often in a social setting; must have high integrity; and must take a leadership role in encouraging more progressive policies in the industry. The position will also involve a high degree of personal activity. Based on these requirements, we choose 10 scales to score the candidates from 7 to 1. The negative signs in the original scoring procedure (-3, -2, -1, 0, +1, +2, +3) were found to produce coding errors. Exhibit 9-7 illustrates the scale used for the research. The letters along the left side, which show the relevant factor, would be omitted from the actual scale, as would the numerical values shown. Note also that the evaluation, potency, and activity scales are mixed, and about half are reversed to minimize the halo effect. To analyze the results, the set of evaluation (E) values is averaged, as are those for the potency (P) and activity (A) dimensions.

The data are plotted in Exhibit 9-8. Here the adjective pairs are reordered so evaluation, potency, and activity descriptors are grouped together with the ideal factor reflected by the left side of the scale. Profiles of the three candidates may be compared to each other and to the ideal.

Adapting SD Scales to the Management Question One study explored a retail store image using 35 pairs of words or phrases classified into eight groups. These word pairs were especially created for the study. Excerpts from this scale are presented in Exhibit 9-9. Other categories of scale items were "general characteristics of the company," "physical characteristics of the store," "prices charged by the store," "store personnel," "advertising by the store," and "your friends and the store." Since the scale pairs are closely associated with the characteristics of the store and its use, one could develop image profiles of various stores.

EXHIBIT 9-9 Adapting SD Scales for Retail Store Image Study

Convenience of Reaching the Store from Your Location		
Nearby	_____	Distant
Short time required to reach store	_____	Long time required to reach store
Difficult drive	_____	Easy drive
Difficult to find parking place	_____	Easy to find parking place
Convenient to other stores I shop	_____	Inconvenient to other stores I shop
Products Offered		
Wide selection of different kinds of products	_____	Limited selection of different kinds of products
Fully stocked	_____	Understocked
Undependable products	_____	Dependable products
High quality	_____	Low quality
Numerous brands	_____	Few brands
Unknown brands	_____	Well-known brands

SOURCE: Robert F. Kelly and Ronald Stephenson, "The Semantic Differential: An Information Source for Designing Retail Patronage Appeals," *Journal of Marketing* 31 (October 1967), p. 45.

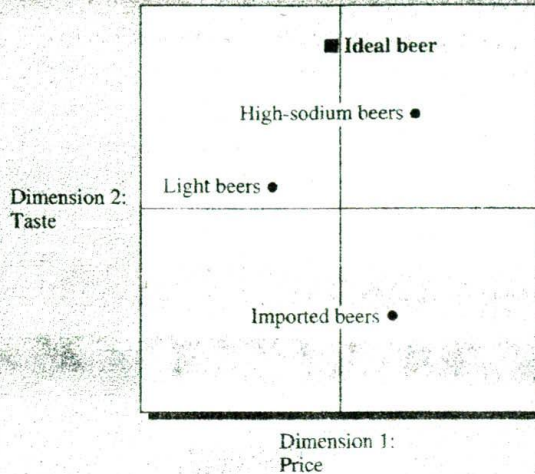
Advanced Scaling Techniques

New construction approaches have removed many of the deficiencies of traditional scales. Some have evolved to handle specific management research applications. Most techniques mentioned in this section rely on complex computer algorithms and require an understanding of multivariate statistics. Students interested in further information on these topics should refer to the statistical examples in Chapter 19 and the references.

Multidimensional scaling (MDS) describes a collection of techniques that deal with property space in a more general manner than the semantic differential. With MDS, one can scale objects, people, or both in ways that provide a visual impression of the relationships among variables. The data-handling characteristics of MDS provide several options: ordinal input (with interval output), and fully metric (interval) and non-metric modes. The various techniques use proximities as input data. A **proximity** is an index of perceived similarity or dissimilarity between objects. The objects might be 20 nations (or 10 primary exports) that respondents are asked to judge in pairs of possible combinations as to their similarity. By means of a computer program, the ranked or rated relationships are then represented as points on a map in multidimensional space.²⁰

We may think of three types of attribute space, each representing a multidimensional map. First, in *objective* space a product can be positioned in terms of, say, its price, taste, and brand image. Second, a person's perceptions also may be positioned in *subjective* space using similar dimensions. These maps do not always coincide, but they do provide information about perceptual disparities. Since the subjective maps vary over time, they also provide important trend data. Third, we can describe our preferences for the object's *ideal* attributes. All objects close to the ideal are more preferred than those farther away. These various configurations are said to reflect the "hidden structure" of the

EXHIBIT 9-10 Multidimensional Map of Beer Preferences



data and make complicated problems much easier to understand. In Exhibit 9-10 two dimensions are plotted: price and taste. The high-sodium beers are closest to the ideal beer on the price dimension while the imported beers are farthest away.

Another approach, representing a collection of techniques, is **conjoint analysis**. Conjoint analysis is used to measure complex decision making that requires multiattribute judgments. Its primary focus has been the explanation of consumer behavior with numerous applications in product development and marketing.²¹

When discovering and learning about products, consumers define a set of attributes or characteristics they use to compare competing brands or models in a product class. Using these attributes, they evaluate the product range and eliminate some brands. Then a final set of alternatives (including a nonpurchase or delayed purchase decision) is developed. These evaluations can change if there is new information about additional competitors, corrections to attribute knowledge, or further thoughts about the attribute. Algebraic theory can be used to model these cognitive processes and develop statistical approximations that reveal the rules the consumer follows in decision making.²²

For example, a consumer might be considering the purchase of a personal computer. MindWriter has a fast processing speed and a high price. Brand X has a low price and a slower processor. The consumer's choice will be evidence of the utility of the processing-speed attribute. Simultaneously, other attributes are being evaluated—such as memory, portability, graphics support, and user friendliness.

Conjoint analysis can produce a scaled value for each attribute as well as a utility value for attributes that have levels (e.g., memory may have a range of 128 to more than 512 megabytes). Both ranking and rating inputs may be used to evaluate product attributes. Conjoint analysis is not restricted to marketing applications, nor should it be considered a single generalized technique (see Chapter 19).

Finally, advanced students who are interested in the above techniques may also wish to investigate *magnitude estimation scaling*.²³ Magnitude scales provide access to ratio measurement and open new alternatives to management problems previously addressed through ordinal scales alone. *Rasch models* also offer alternative approaches to a range of traditional measures from dichotomous responses to Likert-type response formats.²⁴



Close-Up

Myra and Jason had been working on scaling for the CompleteCare project for a week when the call came to Myra to report her progress to MindWriter's general manager, Jean-Claude. They had narrowed the choice to three scales: a Likert scale, a conventional rating scale with two verbal anchors, and their hybrid expectation scale. All were five-point scales that were presumed to measure at the interval level.

They needed a statement that could accompany the scale for preliminary evaluation. Returning to their list of

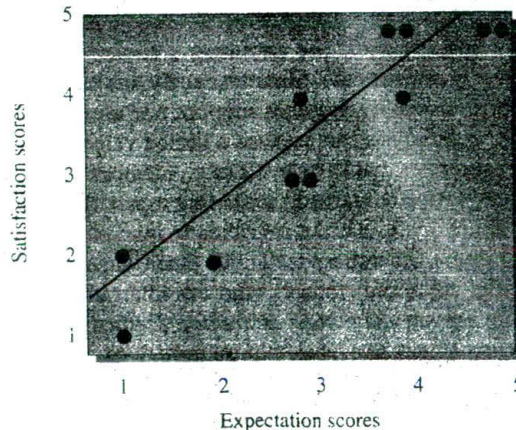
investigative questions, they found a question that seemed to capture the essence of the repair process: "Are customers' problems resolved?" Translated into an assertion for the scale, the statement became, "Resolution of problems that prompted service/repair." They continued to labor over the wording of the verbal anchors after their meeting with Jean-Claude. It was important for the distance between the numbers to resemble the psychological distance implied by the words. Appropriate versions of the investigative question were constructed and then the scales were added:

Likert Scale				
The problem that prompted service/repair was resolved.				
Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
1	2	3	4	5

Conventional Likert Rating Scale (MindWriter's favorite)				
To what extent are you satisfied that the problem that prompted service/repair was resolved?				
Very Dissatisfied				Very Satisfied
1	2	3	4	5

Hybrid Expectation Scale				
Resolution of the problem that prompted service/repair.				
Met few expectations	Met some expectations	Met most expectations	Met all expectations	Exceeded expectations
1	2	3	4	5

EXHIBIT 9-11 Plot of MindWriter Scale Evaluation



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

They also considered a seven-point scale to remedy this but in the end thought the term *exceeded* on the expectation scale could compensate for scores that clustered on the positive end, making the end point less susceptible to leniency.

They were ready for a pilot test. They decided to compare their hybrid expectation scale with MindWriter's conventional Likert rating scale. The Likert scale required that they create more potential items than they had room for on the postcard. Using the CompleteCare database, names,

addresses, and phone numbers were selected. Thirty customers were selected at random from those who had recent service. They chose the delayed equivalent forms method for reliability testing (see Chapter 8). Myra administered the expectation scale followed by the satisfaction scale to half of the respondents and the satisfaction scale followed by the expectation scale to the other half. Each half sample experienced a time delay. No "order-of-presentation" effects were found. Subsequently, they correlated the satisfaction scores with the expectation scores and plotted the results, shown in Exhibit 9-11.

Satisfaction and expectation were positively correlated for "problem resolution" ($r = .90$); reliability based on equivalence was supported. An assessment of test-retest reliability ($r = .93$) showed that the expectation scale had a higher degree of stability over the one-week interval than did the satisfaction scale ($r = .75$). Their scale also produced linear results (as evidenced by the plot).

The decision was made. Myra and Jason would use the new hybrid expectation scale for the CompleteCare project.

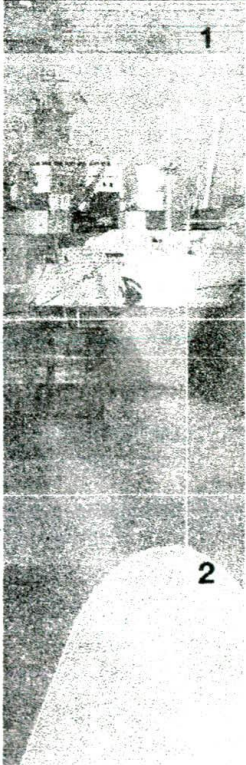
SUMMARY

Scaling describes the procedures by which we assign numbers to measurements of opinions, attitudes, and other concepts. Selection of a measurement scale to best meet our needs involves six decisions:

- **Study objective:** Do we measure the characteristics of the respondent or the stimulus object?
- **Response form:** Do we measure with a rating scale or a ranking scale?
- **Degree of preference:** Do we measure our preferences or make nonpreference judgments?
- **Data properties:** Do we measure with nominal, ordinal, interval, or ratio data?
- **Number of dimensions:** Do we measure using a unidimensional or multidimensional scale?
- **Scale construction:** Do we develop scales by arbitrary decision, consensus, item analysis, cumulative scaling, or factor analysis?

In this chapter, two classifications—the response form and scale construction techniques—were emphasized.

When using rating scales, one judges an object in absolute terms against certain specified criteria. Several scales were proposed: simple category; multiple choice, single-response; multiple choice, multiple-response; Likert scales; semantic differential; numerical scales; multiple rating lists; fixed sum scales; stapel scales; and graphic rating scales. When you use ranking methods, you make relative comparisons against other similar objects. Three well-known methods are the paired-comparison, forced ranking, and the comparative scale.



3

Scaled measurement strategies are classified by the techniques used to construct them. Of the five techniques, three are used frequently: the arbitrary approach, item analysis, and factoring. Consensus and cumulative methods receive less attention because they are time-consuming or have fewer business applications. Arbitrary scales are designed by the researcher's own subjective selection of items. These scales are simple to construct and have content validity only.

In the consensus method, a panel is used to judge the relevance, ambiguity, and attitude level of scale items. Those items that are judged best are then included in the final instrument. The Thurstone method of equal-appearing intervals is a historic consensus method that has given impetus for many current scales.

With the item analysis approach, one develops many items believed to express either a favorable or an unfavorable attitude toward some general object. These items are then pretested to decide which ones discriminate between persons with high total scores and those with low total scores on the test. Those items that meet this discrimination test are included in the final instrument. The most successful Likert scales are developed using this approach.

With the cumulative approach scales, it is possible to estimate how a respondent has answered individual items by knowing the total score. The items are related to each other on a particular attitude dimension, so that if one agrees with a more extreme item, one will also agree with items representing less extreme views. The scalogram is the classic example.

Factoring develops measurement questions through factor analysis or similar correlation techniques. It is particularly useful in uncovering latent attitude dimensions, and it approaches scaling through the concept of multidimensional attribute space. The semantic differential scale is an example.

Other developments in scaling include multidimensional scaling and conjoint analysis. Each represents a family of related techniques with a variety of applications for handling complex judgments. Magnitude estimation and Rasch models provide an avenue for reconceptualizing traditional scaling techniques for greater efficiency and freedom from error.

KEY TERMS

- | | | | | | |
|--------------------------------|-----|------------------------------------|-----|-----------------------------|-----|
| arbitrary scales | 260 | halo effect (error) | 257 | proximity | 268 |
| categorization | 251 | item analysis scaling | 262 | ranking scales | 251 |
| central tendency (error) | 257 | leniency (error) | 256 | rating scales | 251 |
| comparative scale | 260 | Likert scale | 253 | scaling | 250 |
| conjoint analysis | 269 | multidimensional scaling | 251 | scalogram | 263 |
| consensus scaling | 261 | multiple choice, multiple-response | | semantic differential scale | 253 |
| cumulative scales | 263 | scale | 253 | simple category scale | 253 |
| equal-appearing interval scale | 261 | multiple choice, single-response | | stapel scale | 256 |
| factor scales | 264 | scale | 253 | successive intervals | 260 |
| fixed sum scale | 256 | multiple rating list scale | 256 | unidimensional scale | 251 |
| forced ranking scale | 260 | numerical scale | 256 | | |
| graphic rating scale | 256 | paired-comparison scale | 257 | | |

EXAMPLES

Company	Scenario	Page
Airsole*	Constructing agreement items for a scale.	263
Galaxy Department Stores	Seeking to assess teen shopping preferences prior to constructing intrastore teen boutiques.	261
Henry J. Kaiser Family Foundation (KFF)	One-year tracking study to assess physicians' knowledge and attitudes regarding mifepristone (RU486).	264
MindWriter*	Evaluating the CompleteCare program for servicing laptops.	BRIL, Close-Up, 269
Miro Beach City Government*	Evaluating voters' approval or disapproval of a regulatory program.	251
Princeton Survey Research Associates	Conducted the phone survey of physicians in KFF's one-year tracking study of physicians' knowledge and attitudes regarding mifepristone (RU486).	264
Wittenberg University Department of Education	Determining demand for a new program.	258

*Due to the confidential and proprietary nature of most research, the names of some companies have been changed.

DISCUSSION QUESTIONS

Terms in Review

Making Research Decisions

- Discuss the relative merits of and problems with:
 - Rating and ranking scales.
 - Likert and differential scales.
 - Unidimensional and multidimensional scales.
- Suppose your firm had planned a major research study for November 2001. Given the incidents of September 11, your superior decides to add a question to the study. The question must measure consumers' confidence that the U.S. economic system will be able to rebound following the terrorist attacks of September and the subsequent effects of those incidents (increased layoffs, higher unemployment, numerous firms failing to meet their sales and profit projections, lower holiday retail sales, war on terrorism). Draft a scale of each of the following types to measure that confidence level.
 - Fixed sum scale.
 - Likert-type summated scale.
 - Semantic differential scale.
 - Stapel scale.
 - Forced ranking scale.
- An investigative question in your employee satisfaction study seeks to assess employee "job involvement." Create a measurement question that uses the following scales:
 - A graphic rating scale.
 - A multiple rating list.
 - Which do you recommend and why?

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

	Koak	Zip	Pabze	Mr. Peepers
Koak	—	50*	115	35
Zip	150	—	160	70
Pabze	85	40	—	45
Mr. Peepers	165	130	155	—

*Read as 50 persons preferred Zip to Koak.

- How do these brands rank in overall preference in this sample?
 - Develop an interval scale for these four brands.
5. One of the problems in developing rating scales is the choice of response terms to use. Below are samples of some widely used scaling codes. Do you see any problems with them?
- Yes _____ Depends _____ No _____
 - Excellent _____ Good _____ Fair _____ Poor _____
 - Excellent _____ Good _____ Average _____ Fair _____ Poor _____
 - Strongly Approve _____ Approve _____ Uncertain _____ Disapprove _____
Strongly Disapprove _____
6. You are working on a consumer perception study of four brands of bicycles. You will need to develop measurement questions and scales to accomplish the following tasks. Also be sure to explain which data levels (nominal, ordinal, interval, ratio) are appropriate and which quantitative techniques you will use.
- Prepare an overall assessment of all the brands.
 - Provide a comparison of the brands for each of the following dimensions:
 - Styling
 - Durability
 - Gear quality
 - Brand image
7. Below is a Likert-type scale that might be used to evaluate your opinion of the educational program you are in. There are five response categories: Strongly Agree through Neither Agree nor Disagree to Strongly Disagree. If 5 represents the most positive attitude, how would the different items be valued?
- This program is not very challenging.
SA A N D SD
 - The general level of teaching is good.
SA A N D SD
 - I really think I am learning a lot from this program.
SA A N D SD
 - Students' suggestions are given little attention here.
SA A N D SD
 - This program does a good job of preparing one for a career.
SA A N D SD
 - This program is below my expectations.
SA A N D SD

Record your answers to the above items. In what two different ways could such responses be used? What would be the purpose of each?

Bringing Research to Life**From Concept to Practice****WWW Exercises**

8. What is the basis of Jason and Myra's argument for the need of an arbitrary scale to address customer expectations?
9. Using the response strategies within Exhibit 9-1 or 9-2, which would be appropriate and add insight to understanding the various indicants of student demand for the academic program in which they are enrolled?

Visit our website for Internet exercises related to this chapter at www.mhhe.com/business/cooper8

CASES**A GEM OF A STUDY****BBQ PRODUCT CROSSES OVER THE LINES OF VARIED TASTES****CALLING UP ATTENDANCE****CUMMINS ENGINES****INQUIRING MINDS WANT TO KNOW—NOW!****MASTERING TEACHER LEADERSHIP****NCR: TEEING UP A NEW STRATEGIC DIRECTION****PEBBLE BEACH CO.****RAMADA DEMONSTRATES ITS PERSONAL BEST****THE CATALYST FOR WOMEN IN FINANCIAL SERVICES****VOLKSWAGEN'S BEETLE**

*All cases indicating a video icon are located on the Instructor's Videotape Supplement. All nonvideo cases are in the case section of the textbook. All cases indicating a CD icon offer a data set, which is located on the accompanying CD.

REFERENCE NOTES

1. Bernard S. Phillips, *Social Research Strategy and Tactics*, 2nd ed. (New York: Macmillan, 1971), p. 205.
2. For a discussion of various scale classifications, see W. S. Torgerson, *Theory and Methods of Scaling* (New York: Wiley, 1958), Chapter 3.
3. E. A. Suchman and R. G. Francis, "Scaling Techniques in Social Research," in *An Introduction to Social Research*, ed. J. T. Doby (Harrisburg, PA: Stackpole, 1954), pp. 126-29.
4. A study of the historic research literature found that more than three-fourths of the attitude scales used were of the five-point type. An examination of more recent literature suggests that the five-point scale is still common but there is a growing use of longer scales. For the historic study, see Daniel D. Day, "Methods in Attitude Research," *American Sociological Review* 5 (1940), pp. 395-410. Single versus multiple-item scaling requirements are discussed in Jum C. Nunnally, *Psychometric Theory* (New York: McGraw-Hill, 1967), Chapter 14.
5. This section is adapted from Pamela L. Alreck and Robert B. Settle, *The Survey Research Handbook* (Burr Ridge, IL: Irwin, 1995), Chapter 5.
6. J. P. Guilford, *Psychometric Methods* (New York: McGraw-Hill, 1954), pp. 278-79.
7. P. M. Synonds, "Notes on Rating," *Journal of Applied Psychology* 9 (1925), pp. 188-95.
8. See L. L. Thurstone, "A Law of Comparative Judgment," *Psychological Review* 34 (1927), pp. 273-86.
9. Guilford, *Psychometric Methods*.
10. See Milton A. Saffir, "A Comparative Study of Scales Constructed by Three Psychophysical Methods," *Psychometrika* 11, no. 3 (September 1937), pp. 179-98.
11. Allen I. Edwards, *Techniques of Attitude Scale Construction* (New York: Appleton-Century-Crofts, 1957), pp. 152-54.
12. One study reported that the construction of a Likert scale took only half the time required to construct a Thurstone scale. See

- L. L. Thurstone and K. K. Kenney, "A Comparison of the Thurstone and Likert Techniques of Attitude Scale Construction," *Journal of Applied Psychology* 30 (1946), pp. 72-83.
13. Edwards, *Techniques*, p. 153.
 14. Louis Guttman, "A Basis for Scaling Qualitative Data," *American Sociological Review* 9 (1944), pp. 139-50.
 15. John P. Robinson, "Toward a More Appropriate Use of Guttman Scaling," *Public Opinion Quarterly* 37 (Summer 1973), pp. 260-67.
 16. For more on the process of factor analysis, see Chapter 19.
 17. Charles E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning* (Urbana, IL: University of Illinois Press, 1957).
 18. *Ibid.*, p. 49. See also James G. Snider and Charles E. Osgood, eds., *Semantic Differential Technique* (Chicago: Aldine, 1969).
 19. *Ibid.*, p. 79.
 20. See, for example, Joseph B. Kruskal and Myron Wish, *Multidimensional Scaling* (Beverly Hills, CA: Sage Publications, 1978); Paul Green and V. R. Rao, *Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms* (New York: Holt, Rinehart & Winston, 1972); and Paul E. Green and F. J. Carmone, *Multidimensional Scaling in Marketing Analysis* (Boston: Allyn & Bacon, 1970).
 21. See P. Cattin and D. R. Wittink, "Commercial Use of Conjoint Analysis: A Survey," *Journal of Marketing* 46 (1982), pp. 44-53; and Cattin and Wittink, "Commercial Use of Conjoint Analysis: An Update" (paper presented at the ORSA/TIMS Marketing Science Meetings, Richardson, TX, March 12-15, 1986).
 22. Jordan J. Louviere, *Analyzing Decision Making: Metric Conjoint Analysis* (Beverly Hills, CA: Sage Publications, 1988), pp. 9-11.
 23. See, for example, Milton Lodge, *Magnitude Scaling: Quantitative Measurement of Opinions* (Beverly Hills, CA: Sage Publications, 1981); and Donald R. Cooper and Donald A. Clare, "A Magnitude Estimation Scale for Human Values," *Psychological Reports* 49 (1981).
 24. David Andrich, *Rasch Models for Measurement* (Beverly Hills, CA: Sage Publications, 1988).

REFERENCES FOR SNAPSHOTS AND CAPTIONS

RU486

- "One Year Later: Medical Abortion After FDA Approval," The Henry J. Kaiser Family Foundation, September 24, 2001 (<http://www.kff.org/content/2001/3170/>).
- "National Survey of Women's Health Care Providers on Reproductive Health: Medical Abortion Results: Selected Findings from the Kaiser/Harvard Health News Index (August 2001): Public Knowledge and Awareness of Mifepristone," The Henry J. Kaiser

Family Foundation, September 24, 2001 (<http://www.kff.org/content/2001/3170/SurveyToplinesNew.pdf>).

Wittenberg University

- Interview with Dr. Robert Welker, project director, Master of Education, Wittenberg University, December 16, 1999.
- Survey instrument used in Master of Education market test.

CLASSIC AND CONTEMPORARY READINGS

- Edwards, Allen L. *Techniques of Attitude Scale Construction*. New York: Irvington, 1979. Thorough discussion of basic unidimensional scaling techniques.
- Kerlinger, Fred N., and Howard B. Lee. *Foundations of Behavioral Research*. 4th ed. New York: HBJ College & School Division, 1999.
- Krebs, Dagmar, and Peter Schmidt, ed. *New Directions in Attitude Measurement*. Chicago: Walter De Gruyter, 1993.
- Miller, Delbert C. *Handbook of Research Design and Social Measurement*. 5th ed. Thousand Oaks, CA: Sage Publications, 1991. Presents a large number of existing sociometric scales and indexes as well as information on their characteristics, validity, and sources.
- Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. Urbana, IL: University of Illinois Press, 1957. The basic reference on SD scaling.

The Sources and Collection of Data

CHAPTER 10 Exploring Secondary Data

CHAPTER 11 Survey Methods: Communicating with Participants

CHAPTER 12 Instruments for Participant Communication

CHAPTER 13 Observational Studies

CHAPTER 14 Experimentation