rates secured from personal interview subjects. This study will involve 300 people who will be assigned to one of the following conditions: (1) no compensation, (2) $1 compensation, and (3) $3 compensation. A number of sensitive issues will be explored concerning various social problems, and the 300 people will be drawn from the adult population. Describe your design. You may find Appendix 11a valuable for this question.

10 What type of experimental design would you recommend in each of the following cases? Suggest in some detail how you would design each study:

a A test of three methods of compensation of factory workers. The methods are hourly wage, incentive pay, and weekly salary. The dependent variable is direct labor cost per unit of output.

b A study of the effects of various levels of advertising effort and price reduction on the sale of specific branded grocery products by a retail grocery chain.

c A study to determine whether it is true that the use of fast-paced music played over a store's public address system will speed the shopping rate of customers without an adverse effect on the amount spent per customer.

Bringing Research to Life

11 Design an experiment for the opening vignette.

From Concept to Practice

12 Using Exhibit 11-4, diagram an experiment described in one of the Snapshots in this chapter using research design symbols.

>wwwexercises

1 For experiments and surveys on the Web, visit **http://www.psych.upenn.edu/~baron/qs.html#webexpts** and participate in an online experiment. Prepare a short paper describing your experience, and make suggestions for improving the experimental design.

2 Use a search engine to find an experiment described on the Web. Remember that experiments sometimes go by other names, like *taste test* in consumer food products or *beta test* in software products. Also, use terms introduced in this chapter. What experiment could you do that would use the same methodology as the one you discovered?

>cases*

McDonald's Tests Catfish Sandwich

NetConversions Influences Kelley Blue Book

Retailers Unhappy with Displays by Manufacturers

* All cases appear on the text CD; you will find abstracts of these cases in the Case Abstracts section of this text.

>appendix11a

Complex Experimental Designs

Earlier in the chapter, we discussed true experimental designs in their most frequently used forms, but researchers often require an extension of the basic design for sophisticated experiments and market tests. Extensions differ from the traditional designs in (1) the number of different experimental stimuli that are considered simultaneously by the experimenter and (2) the extent to which assignment procedures are used to increase precision.

Before we consider the types of variations, there are some commonly used terms that should be defined. *Factor* is widely used to denote an independent variable. Factors are divided into treatment levels, which represent various subgroups. A factor may have two or more levels, such as (1) male and female; (2) large, medium, and small; or (3) no training, brief training, and extended training. These levels should be defined operationally.

Factors also may be classified by whether the experimenter can manipulate the levels associated with the participant. *Active factors* are those the researcher can manipulate by causing a participant to receive one level or another. Treatment is used to denote the different levels of active factors. With the second type, the *blocking factor,* the experimenter can only identify and classify the participant on an existing level. Gender, age group, customer status, and ethnicity are examples of blocking factors, because the participant comes to the experiment with a preexisting level of each.

Up to this point, the assumption is that experimental participants are people, but this is often not so. A broader term is *test unit;* it can refer equally well to an individual, product type, geographic market, medium of information dissemination, and innumerable other entities.*

Completely Randomized Design

The basic form of the true experiment is a completely randomized design. To illustrate its use, and that of more complex designs, consider a decision now facing the pricing manager at the Top Cannery. He would like to know what the ideal difference in price is between Top's private brand of canned vegetables and national brands such as Del Monte and Stokely's.

It is possible to set up an experiment on price differentials for canned green beans. Eighteen company stores and three price spreads (treatment levels) of 7 cents, 12 cents, and 17 cents between the company brand and national brands are used for the study. Six of the stores are assigned randomly to each of the treatment groups. The price differentials are maintained for a period, and then a tally is made of the sales volumes and gross profits of the canned green beans for each group of stores.

This design can be diagrammed as follows:

$$\begin{array}{cccc} R & O_1 & X_1 & O_2 \\ R & O_3 & X_3 & O_4 \\ R & O_5 & X_5 & O_6 \end{array} \qquad \text{(A1)}$$

Here, O_1, O_3, and O_5 represent the total gross profits for canned green beans in the treatment stores for the month before the test. X_1, X_3, and X_5 represent 7-cent, 12-cent, and 17-cent treatments, while O_2, O_4, and O_6 are the gross profits for the month after the test started.

We assume that the randomization of stores to the three treatment groups was sufficient to make the three store groups equivalent. When there is reason to believe this is not so, we must use a more complex design.

Randomized Block Design

If there is a single major extraneous variable, the randomized block design is used. Random assignment is still the basic way to produce equivalence among treatment groups, but the researcher may need additional assurances. First, if the sample being studied is very small, it is risky to depend on random assignment alone to guarantee equivalence. Small samples, such as the 18 company stores, are typical in field experiments because of high costs or because few test units are available. Another reason for blocking is to learn whether treatments bring different results among various groups of participants.

Consider again the canned green beans pricing experiment. Assume there is reason to believe that

*Check this Web site for examples of industrial experiments: **http://www.statsoft.com/.**

lower-income families are more sensitive to price differentials than are higher-income families. This factor could seriously distort our results unless we stratify the stores by customer income. Therefore, each of the 18 stores is assigned to one of three income blocks and randomly assigned, within blocks, to the price difference treatments. The design is shown in the following table.

		Blocking Factor: Customer Income		
Active Factor: Price Difference		High	Medium	Low
7 cents	R	X_1	X_1	X_1
12 cents	R	X_2	X_2	X_2
17 cents	R	X_3	X_3	X_3

Note: The O's have been omitted. The horizontal rows no longer indicate a time sequence, but various levels of the blocking factor. However, before-and-after measurements are associated with each of the treatments.

(A2)

In this design, one can measure both main effects and interaction effects. The *main effect* is the average direct influence that a particular treatment of the independent variable (IV) has on the dependent variable (DV), independent of other factors. The *interaction effect* is the influence of one factor or variable on the effect of another. The main effect of each price difference is discovered by calculating the impact of each of the three treatments averaged over the different blocks. Interaction effects occur if you find that different customer income levels have a pronounced influence on customer reactions to the price differentials. (See Chapter 18, "Hypothesis Testing.")

Whether the randomized block design improves the precision of the experimental measurement depends on how successfully the design minimizes the variation within blocks and maximizes the variation between blocks. If the response patterns are about the same in each block, there is little value to the more complex design. Blocking may be counterproductive.

Latin Square Design

The Latin square design may be used when there are two major extraneous factors. To continue with the pricing example, assume we decide to block on the size of store and on customer income. It is convenient to consider these two blocking factors as forming the rows and columns of a table. We divide each factor into three levels to provide nine groups of stores, each representing a unique combination of the two blocking variables. Treatments are then randomly assigned to these cells so that a given treatment appears only once in each row and column. Because of this restriction, a Latin Square must have the same number of rows, columns, and treatments.

The design looks like the following table.

	Customer Income		
Store Size	High	Medium	Low
Large	X_3	X_1	X_2
Medium	X_2	X_3	X_1
Small	X_1	X_2	X_3

(A3)

Treatments can be assigned by using a table of random numbers to set the order of treatment in the first row. For example. the pattern may be 3, 1, 2 as shown above. Following this, the other two cells of the first column are filled similarly, and the remaining treatments are assigned to meet the restriction that there can be no more than one treatment type in each row and column.

The experiment takes place, sales results are gathered, and the average treatment effect is calculated. From this, we can determine the main effect of the various price spreads on the sales of company and national brands. The cost information allows us to discover which price differential produces the greatest margin.

A limitation of the Latin square is that we must assume there is no interaction between treatments and blocking factors. Therefore, we cannot determine the interrelationships among store size, customer income, and price spreads. This limitation exists because there is not an exposure of all combinations of treatments, store sizes, and customer income groups. Such an exposure would require a table of 27 cells, while this one has only 9. If one is not especially interested in interaction, the Latin square is much more economical.

	Price Spread		
Unit Price Information?	7 Cents	12 Cents	17 Cents
Yes	X_1Y_1	X_1Y_2	X_1Y_3
No	X_2Y_1	X_2Y_2	X_2Y_3

(A4)

Factorial Design

One commonly held misconception about experiments is that the researcher can manipulate only one variable at a time. This is not true; with factorial designs, you can deal with more than one treatment simultaneously. Consider again the pricing experiment. The president of the chain might also be interested in finding the effect of posting unit prices on the shelf to aid shopper decision making. The following table can be used to design an experiment that includes both the price differentials and the unit pricing.

This is known as a 2 × 3 factorial design in which we use two factors: one with two levels and one with three levels of intensity.* The version shown here is completely randomized, with the stores being randomly assigned to one of six treatment combinations. With such a design, it is possible to estimate the main effects of each of the two independent variables and the interactions between them. The results can help to answer the following questions:

1. What are the sales effects of the different price spreads between company and national brands?
2. What are the sales effects of using unit-price marking on the shelves?
3. What are the sales effect interrelations between price spread and the presence of unit-price information?

Covariance Analysis

We have discussed direct control of extraneous variables through blocking. It is also possible to apply some degree of indirect statistical control on one or more variables through analysis of covariance. Even with randomization, one may find that the before-measurement shows an average knowledge-level difference between experimental and control groups. With covariance analysis, one can adjust statistically for this before-difference. Another application might occur if the canned green beans pricing experiment were carried out with a completely randomized design, only to reveal a contamination effect from differences in average customer income levels. With covariance analysis, one can still do some statistical blocking on average customer income after the experiment has been run.†

* We describe factorial designs used with conjoint analysis in Chapter 20.

† We discuss the statistical aspects of covariance analysis with analysis of variance (ANOVA) in Chapter 18.

>part III

Chapter 12 **Measurement**

Chapter 13 **Measurement Scales**

Chapter 14 **Questionnaires and Instruments**

Chapter 15 **Sampling**

The Sources and Collection of Data

>chapter 12

Measurement

“If you can’t measure it, you can’t manage it.”

Bob Donath, consultant,
Bob Donath and Co., Inc.

>**learning**objectives

After reading this chapter, you should understand . . .

1 The distinction between measuring objects, properties, and indicants of properties.

2 The similarities and differences between the four scale types used in measurement and when each is used.

3 The four major sources of measurement error.

4 The criteria for evaluating good measurement.

>**bringing**research**to**life

The executive director of Glacier Symphony gestures broadly at the still snowcapped Canadian Rockies. "It has been three very happy years for me here, though not easy ones since I let corporate America intrude on our idyllic existence."

"You mean the MindWriter people?" prompts Jason Henry. "The ones who flew me up here? My clients and your benefactor?"

"Please, don't misunderstand," says the executive director as she propels Jason across a manicured lawn toward the refreshment tent. "When I rented them a part of our compound for use in corporate education, they quite generously insisted that I avail myself of some of their training for midlevel managers."

"They said you were having trouble with attendance?" ventures Jason. "Tell me what you do here."

"We offer one of the most outstanding summer music festivals in the country—maybe the continent. We present several concerts each week, all summer long, with evening performances on both Friday and Saturday. During the week, rehearsals are open to music patrons and students. And, of course, our skilled musicians enhance their own skills by networking with each other.

"During the winter my artistic directors prepare the next summer's program and hire the musicians, coordinating closely with me on the budget. This is quite complicated, as most of our musicians spend only two weeks with us. Fully 600 performing artists from many parts of the continent are part of this orchestra over the course of a summer festival.

"Colleges in British Columbia send me their music scholarship students for summer employment as dishwashers, waiters, cleaners, and the like. It is a special opportunity for them, rubbing shoulders with their idols and learning to enhance their own performance skills in the process."

"So your problem is . . . ?" urges Jason again.

"My problem is patronage, specifically the lack of commitment of the local residents of Glacier to consistently support their Glacier Symphony Festival. Do you realize how rare it is for a town this size to have more than 600 performing musicians in a summer? You would think the residents would be as ecstatic as our dishwashers!"

"Do you know why they are less than supportive?" inquires Jason, glad they have finally arrived at the reason MindWriter had asked him to divert his homebound flight from San Francisco to British Columbia.

"Well, some of the residents have communicated with us informally," comments the director, somewhat hesitantly.

"And they said . . . ?" urges Jason, more than a little impatiently, remembering why he so values his partner for usually handling this phase of exploratory research.

"One commented: 'I've never heard this music before—why can't the performers play something I'll recognize.' Another, 'Where were the video screens? And the special visual effects?' And another: 'Why would I want to spend more than an hour watching a stage full of people sitting in chairs?'"

"Hold on," says Jason, making a note in his PalmPilot. "I can see your orchestra is striking a sour note." Jason smiles, chuckling at his own wit, while the director remains stoic. "MindWriter uses an extensive program for measuring customer satisfaction, and . . ."

"Ah, yes, measuring customer satisfaction," interrupts the director, "second only to cash flow for the MindWriter folks. The care and frequency with which they measure customer satisfaction in the MindWriter seminars here dumbfounds me. Throughout each seminar they host here, morning, afternoon, or evening, everyone breaks for coffee and is required to fill out a critique of the speaker. The results are tabulated by the time the last coffee cup has been collected, and the seminar leader has been given feedback. Is he or she presenting material too slowly or too quickly? Are there too many jokes or not enough? Are concrete examples being used often enough? Do the participants want a hard copy of the slides? They measure attitudinal data six times a day and even query you about the meals, including taste, appearance, cleanliness and speed, friendliness, and accuracy of service."

"Understandable," observes Jason. "Your scholarship students have frequent contact with the residents, both here and in town, right? We might use them to collect some more formal data," brainstorms Jason to himself.

"Jason," interjects the director, "were you ever a musician?"

"No," explains Jason, "my interests ran more toward statistics than Schubert."

"Then you wouldn't realize that while musicians could talk about music—and the intricacies of performing music—for hours with each other, once a resident exclaims little or no interest, our scholarship students would likely tune them out."

"It is just as well," comments Jason, now resigned to getting more involved in Glacier Symphony's problem than he had first assumed would be necessary. "Untrained interviewers and observers can be highly unreliable and inaccurate in measuring and reporting behavior," says Jason. "Have you tried a suggestion box?"

"No, but I do send reminder postcards for each concert."

"Not quite the same thing," murmurs Jason as he hands the director his business card. "As a devotee of the MindWriter way, I'm sure you have a current satisfaction survey for concert goers in your files." At her nod, Jason continues, "Send it to me. At MindWriter's request and at its expense, I'll revise it for you. I'll work out the collection and analysis details on my flight home and be in touch next week."

The director, smiling and shaking Jason's hand, responds, "I'll ask one of our scholarship students to drive you back to the community airport, then. You're bound to have a lot in common."

> The Nature of Measurement

In everyday usage, measurement occurs when an established index verifies the height, weight, or other feature of a physical object. How well you like a song, a painting, or the personality of a friend is also a measurement. To measure is to discover the extent, dimensions, quantity, or capacity of something, especially by comparison with a standard. We measure casually in daily life, but in research the requirements are rigorous.

Measurement in research consists of assigning numbers to empirical events, objects or properties, or activities in compliance with a set of rules. This definition implies that measurement is a three-part process:

1. Selecting observable empirical events.
2. Developing a set of **mapping rules:** a scheme for assigning numbers or symbols to represent aspects of the event being measured.
3. Applying the mapping rule(s) to each observation of that event.[1]

You recall the term *empirical.* Researchers use an empirical approach to describe, explain, and make pre-dictions by relying on information gained through observation.

Assume you are studying people who attend an auto show where prototypes for new models are on display. You are interested in learning the male-to-female ratio among attendees. You observe those who enter the show area. If a person is female, you record an F; if male, an M. Any other symbols such as 0 and 1 or # and % also may be used if you know what group the symbol identifies. Exhibit 12-1 uses this example to illustrate the above components.

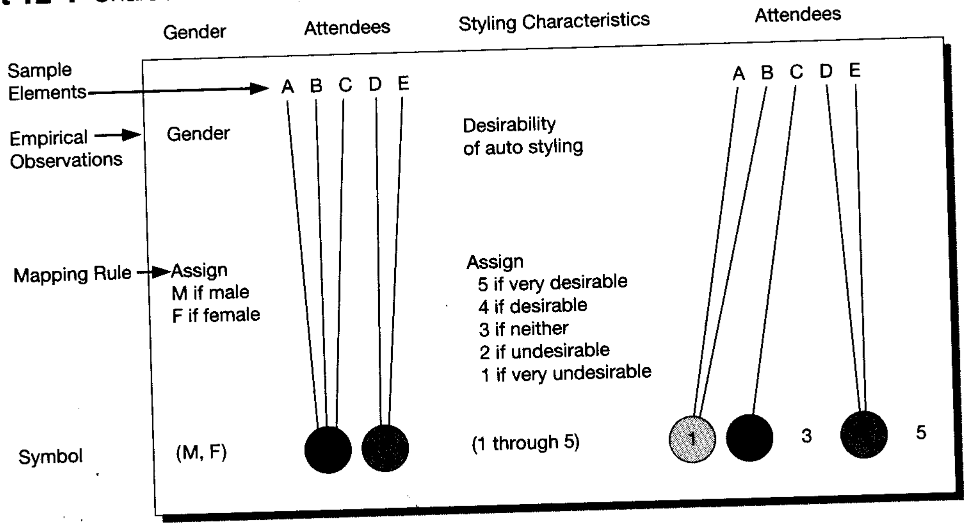
Researchers might also want to measure the styling desirability of a new concept car at this show. They interview a sample of visitors and assign, with a different mapping rule, their opinions to the following scale:

What is your opinion of the styling of the Speedbird?

Very desirable 5 4 3 2 1 Very undesirable

All measurement theorists would call the rating scale in Exhibit 12-1 a form of measurement, but some would challenge whether classifying males and females is a form of measurement. Their argument is that

> Exhibit 12-1 Characteristics of Measurement


Gender
Attendees
Styling Characteristics
Attendees
Sample Elements
A B C D E
A B C D E
Empirical Observations
Gender
Desirability of auto styling
Mapping Rule
Assign M if male F if female
Assign 5 if very desirable 4 if desirable 3 if neither 2 if undesirable 1 if very undesirable
Symbol
(M, F)
(1 through 5)
1
3
5

Attendees A, B, and C are male, and find the auto's styling to be undesirable.
Attendees D and E are female and find the auto's styling desirable.

measurement must involve quantification—that is, "the assignment of numbers to objects to represent amounts or degrees of a property possessed by all of the objects."[2] This condition was met when measuring opinions of car styling. Our approach endorses the more general view that "numbers as symbols within a mapping rule" can reflect both qualitative and quantitative concepts.

The goal of measurement—indeed, the goal of "assigning numbers to empirical events in compliance with a set of rules"—is to provide the highest-quality, lowest-error data for testing hypotheses, estimation or prediction, or description. Researchers deduce from a hypothesis that certain conditions should exist. Then they measure for these conditions in the real world. If found, the data lend support to the hypothesis; if not, researchers conclude the hypothesis is faulty. An important question at this point is, "Just what does one measure?"

The object of measurement is a *concept*, the symbols we attach to bundles of meaning that we hold and share with others. We invent higher-level concepts—*constructs*—for specialized scientific explanatory purposes that are not directly observable and for thinking about and communicating abstractions. Concepts and constructs are used at theoretical levels; variables are used at the empirical level. *Variables* accept numerals or values for the purpose of testing and measurement. Concepts, constructs, and variables may be defined descriptively or operationally. An *operational definition* defines a variable in terms of specific measurement and testing criteria. It must specify adequately the empirical information needed and how it will be collected. In addition, it must have the proper scope or fit for the research problem at hand. We review these terms with examples in Exhibit 12-2.

< You may want to revisit Chapter 2 for a thorough discussion of these research terms.

> Exhibit 12-2 Review of Key Terms

***Concept:* a bundle of meanings or characteristics associated with certain events, objects, conditions, situations, or behaviors.**

Classifying and categorizing objects or events that have common characteristics beyond any single observation creates concepts. When you think of a spreadsheet or a warranty card, what comes to mind is not a single example but your collected memories of all spreadsheets and warranty cards from which you abstract a set of specific and definable characteristics.

[illegible]

***Variable:* an event, act, characteristic, trait, or attribute that can be measured and to which we assign numerals or values; a synonym for the construct or the property being studied.**

The numerical value assigned to a variable is based on the variable's properties. For example, some variables, said to be *dichotomous*, have only two values, reflecting the presence or absence of a property: employed-unemployed or male-female have two values, generally 0 and 1. Variables also take on values representing added categories, such as the demographic variables of race and religion. All such variables that produce data that fit into categories are *discrete* variables, since only certain values are possible. An automotive variable, for example, where "Chevrolet" is assigned a 5 and "Honda" is assigned a 6 provides no option for a 5.5. Income, temperature, age, and a test score are examples of *continuous* variables. These variables may take on values within a given range or, in some cases, an infinite set. Your test score may range from 0 to 100, your age may be 23.5, and your present income could be $35,000.

[illegible]

What Is Measured?

Variables being studied in research may be classified as objects or as properties. **Objects** include the concepts of ordinary experience, such as tangible items like furniture, laundry detergent, people, or automobiles. Objects also include things that are not as concrete, such as genes, attitudes, and peer-group pressures. **Properties** are the characteristics of the object. A person's *physical properties* may be stated in terms of weight, height, and posture, among others. *Psychological properties* include attitudes and intelligence. *Social properties* include leadership ability, class affiliation, and status. These and many other properties of an individual can be measured in a research study.

In a literal sense, researchers do not measure either objects or properties. They measure indicants of the properties or indicants of the properties of objects. It is easy to observe that A is taller than B and that C participates more than D in a group process. Or suppose you are analyzing members of a sales force of several hundred people to learn what personal properties contribute to sales success. The properties are age, years of experience, and number of calls made per week. The indicants in these cases are so accepted that one considers the properties to be observed directly.

In contrast, it is not easy to measure properties of constructs like "lifestyles," "opinion leadership," "distribution channel structure," and "persuasiveness." Since each property cannot be measured directly, one must infer its presence or absence by observing some indicant or pointer measurement. When you begin to make such inferences, there is often disagreement about how to develop an operational definition for each indicant.

Not only is it a challenge to measure such constructs, but a study's quality depends on what measures are selected or developed and how they fit the circumstances. The nature of measurement scales, sources of error, and characteristics of sound measurement are considered next.

> Measurement Scales

In measuring, one devises some mapping rule and then translates the observation of property indicants using this rule. For each concept or construct, several types of measurement are possible; the appropriate choice depends on what you assume about the mapping rules. Each one has its own set of underlying assumptions about how the numerical symbols correspond to real-world observations.

Mapping rules have four characteristics:

1. *Classification.* Numbers are used to group or sort responses. No order exists.
2. *Order.* Numbers are ordered. One number is greater than, less than, or equal to an-other number.
3. *Distance.* Differences between numbers are ordered. The difference between any pair of numbers is greater than, less than, or equal to the difference between any other pair of numbers.
4. *Origin.* The number series has a unique origin indicated by the number zero. This is an absolute and meaningful zero point.

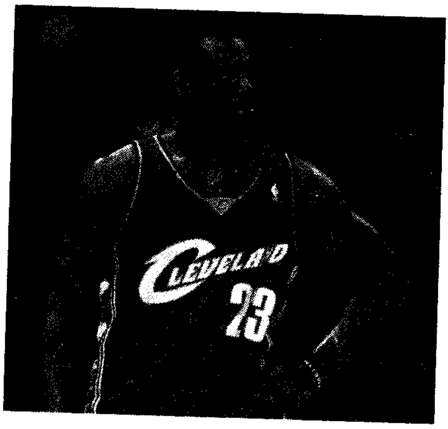
Combinations of these characteristics of classification, order, distance, and origin provide four widely used classifications of measurement scales:[3] (1) nominal, (2) ordinal, (3) interval, and (4) ratio. Let's preview these measurement scales before we discuss their technical details. Suppose your professor asks a student volunteer to taste-test six candy bars. The student begins by evaluating each on a chocolate–not chocolate scale; this is a nominal measurement. Then the student ranks the candy bars from best to worst; this is an ordinal measurement. Next, the student uses a 7-point scale that has equal distance between points to rate the candy bars with regard

> **Exhibit 12-3** Measurement Scales

Type of Scale	Characteristics of Data	Basic Empirical Operation	Example
Nominal	Classification (mutually exclusive and collectively exhaustive categories), but no order, distance, or natural origin	Determination of equality	Gender (male, female)
Ordinal	Classification and order, but no distance or natural origin	Determination of greater or lesser value	Doneness of meat (well, medium well, medium rare, rare)
Interval	Classification, order, and distance, but no natural origin	Determination of equality of intervals or differences	Temperature in degrees
Ratio	Classification, order, distance, and natural origin	Determination of equality of ratios	Age in years

to some taste criterion (e.g., crunchiness); this is an interval measurement. Finally, the student, considers another taste dimension and assigns 100 points among the six candy bars; this is a ratio measurement.

The characteristics of these measurement scales are summarized in Exhibit 12-3. Deciding which type of scale is appropriate for your research needs should be seen as a part of the research process, as shown in Exhibit 12-4.



Nominal Scales

In business research, nominal data are widely used. With **nominal scales,** you are collecting information on a variable that naturally or by design can be grouped into two or more categories that are mutually exclusive and collectively exhaustive. If data were collected from the symphony patrons at the Glacier compound, patrons could be classified by whether they had attended prior symphony performances or this was their first time. Every patron would fit into one of the two groups within the variable *attendance.*

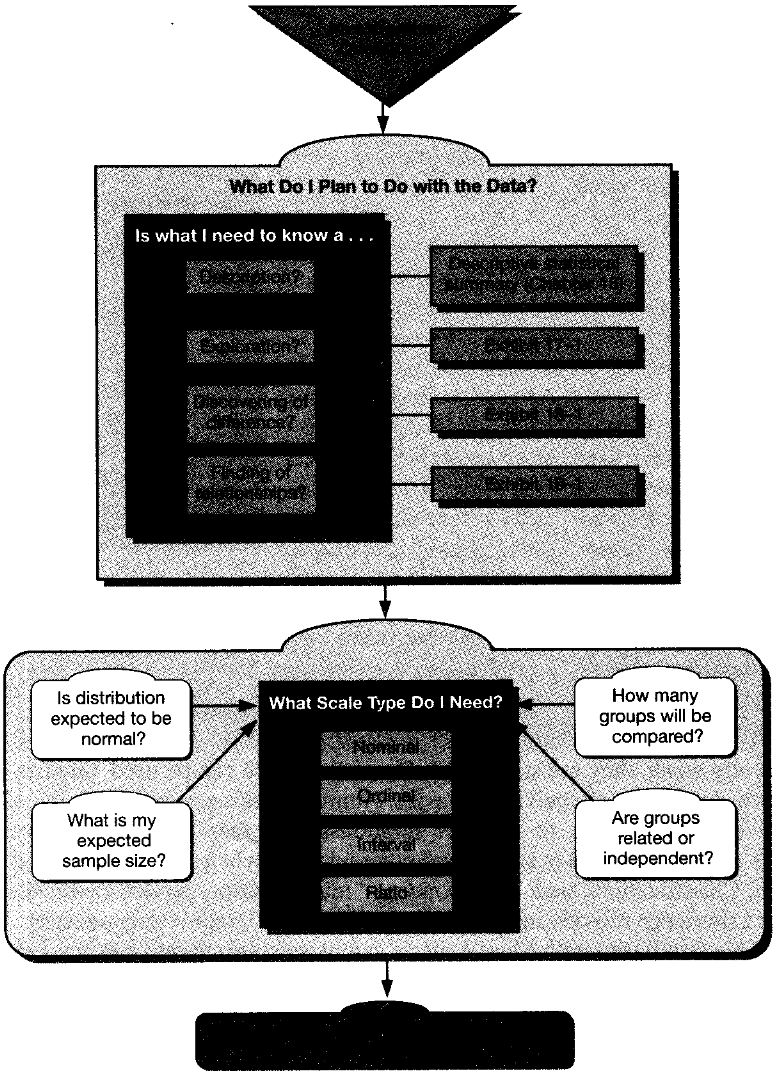
The counting of members in each group is the only possible arithmetic operation when a nominal scale is employed. If we use numerical symbols within our mapping rule to identify categories, these numbers are recognized as labels only and have no quantitative value. The number 23, we know, does not imply a sequential count of players or a skill level; it is only a means of identification. Of course, you might want to argue about a jersey number representing a skill level if it is LeBron James wearing jersey 23.

Nominal classifications may consist of any number of separate groups if the groups are mutually exclusive and collectively exhaustive. Thus, one might classify the students in a course according to their

Religious Preferences

Mapping Rule A	Mapping Rule B
1 = Baptist	1 = Christian
2 = Catholic	2 = Muslim
3 = Protestant	3 = Hindu
4 = Scientology	4 = Buddhist
5 = Unitarian-Universalist	5 = Jewish
6 = Jewish	6 = Other
7 = Secular/nonreligious/agnostic/atheist	

> Exhibit 12-4 Moving from Investigative to Measurement Questions


What Do I Plan to Do with the Data?
Is what I need to know a . . .
Description?
Exploration?
Discovering of difference?
Finding of relationships?
Descriptive statistical summary (Chapter 18)
Exhibit 17-1
Exhibit 18-1
Exhibit 19-1
What Scale Type Do I Need?
Nominal
Ordinal
Interval
Ratio
Is distribution expected to be normal?
What is my expected sample size?
How many groups will be compared?
Are groups related or independent?

expressed religious preferences. Mapping rule A given in the table is not a sound nominal scale because its categories are not mutually exclusive or collectively exhaustive. Mapping rule B meets the minimum requirements; it covers all the major religions and offers an "other" option. Nominal scales are the least powerful of the four data types. They suggest no order or distance relationship and have no arithmetic origin. The scale wastes any information a sample element might share about varying degrees of the property being measured.

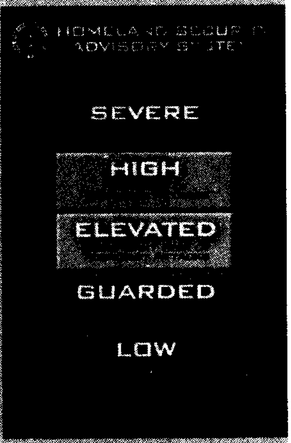
Since the only quantification is the number count of cases in each category (the frequency distribution), the researcher is restricted to the use of the mode as the measure of central tendency.[4] The *mode* is the most frequently occurring value. You can conclude which category has the most members, but that is all. There is no generally used measure of *dispersion* for nominal scales. Dispersion describes how scores cluster or scatter in a distribution. By cross-tabulating nominal variables with other variables, you can begin to discern patterns in data.

> We discuss significance tests and measures of association in Chapters 18 and 19. Several tests for statistical significance may be used with nominal data; the most common is the chi-square test.

>snapshot

In March 2002, the Homeland Security Advisory System was announced as a threat-level system to publicize information about the risk of terrorist attacks to federal, state, and local authorities, and to the American people. The five-rank, color-coded system begins at the lowest level, green, and moves upward through blue, yellow, orange, and red ranks. Each color represents a threat level (low, guarded, elevated, high, severe) and is accompanied by security criteria in each category. Officials compare the alert levels to the color-coded system used during World War II.

The system was designed to be a comprehensive and effective communications structure based on a terror-measurement system. As a public relations communication initiative it has been less successful. It has caused confusion and consternation when government spokespeople explain the reasons for raising and lowering threat levels—which are essentially based on changing threat assessments by intelligence analysts. Then Homeland Security secretary Tom Ridge contended that the five-level system gives government and industry a "common vocabulary" and concrete suggestions for preparing against terrorist attacks. Private citizens find that as a measurement system it provides little concrete evidence or advice for preparation, as evidenced by nearly three weeks in February 2003 when an orange alert caused anxious Americans to stock up on water, food, duct tape, and plastic sheeting. The frustration with this measurement system was summed up by one editorial writer: "If it's normal to be 'elevated,' being higher inevitably will be alarming. And if the color coordinators can't say anything about the nature or location of the threat that's changing the color, they will frighten people everywhere."



www.whitehouse.gov; www.gopbi.com

While nominal data are statistically weak, they are still useful. If no other scale can be used, one can almost always classify a set of properties into a set of equivalent classes. Nominal measures are especially valuable in exploratory work where the objective is to uncover relationships rather than secure precise measurements. This type of scale is also widely used in survey and other research when data are classified by major subgroups of the population. Classifications such as respondents' marital status, gender, political orientation, and exposure to a certain experience provide insight into important demographic data patterns.

Jason visited Glacier because of his familiarity with MindWriter's extensive research into customer satisfaction. His visit revealed Glacier's need for some exploratory nominal data on symphony patrons. Patrons could be divided into groups—based on their appreciation of the conductor (favorable, unfavorable), on their attitude toward facilities (suitable, not suitable), on their perception of the program (clichéd, virtuoso), on their level of symphony support (financial support, no financial support)—and then analyzed.

Ordinal Scales

Correlational analysis of ordinal data is restricted to various ordinal techniques. Measures of statistical significance are technically confined to a body of statistics known as **nonparametric methods,** ***synonymous with*** **distribution-free statistics.**[6]

Ordinal scales include the characteristics of the nominal scale plus an indication of order. Ordinal data require conformity to a logical postulate, which states: If a is greater than b and b is greater than c, then a is greater than c.[5] The use of an ordinal scale implies a statement of "greater than" or "less than" (an equality statement is also acceptable) without stating how much greater or less. While ordinal measurement speaks of greater-than and less-than measurements, other descriptors may be used—"superior to," "happier than," "poorer than," or "important than." Like a rubber yardstick, an ordinal scale can stretch varying amounts at different places along its length. Thus, the real difference between ranks 1 and 2 on a satisfaction scale may be more or less than the difference between ranks 2 and 3. An ordinal concept can be extended beyond the three cases used in the simple illustration of $a > b > c$. Any number of cases can be ranked.

Another extension of the ordinal concept occurs when there is more than one property of interest. We may ask a taster to rank varieties of carbonated soft drinks by flavor, color, carbonation, and a combination of these characteristics. We can secure the combined ranking either by asking the respondent to base his or her ranking on the combination of properties or by constructing a combination ranking of the individual rankings on each property.

Examples of ordinal data include attitude and preference scales. (In the next chapter, we provide detailed examples of attitude scales.) Because the numbers used with ordinal scales have only a rank meaning, the appropriate measure of central tendency is the median. The *median* is the midpoint of a distribution. A percentile or quartile reveals the dispersion.

Researchers differ about whether more powerful tests are appropriate for analyzing ordinal measures. Because nonparametric tests are abundant, simple to calculate, have good statistical power,[7] and do not require that the researcher accept the assumptions of parametric testing, we advise their use with nominal and ordinal data. It is understandable, however, that because parametric tests (such as the *t*-test or analysis of variance) are versatile, accepted, and understood, they will continue to be used with ordinal data when those data approach the characteristics required for interval measurement.

Interval Scales

Interval scales have the power of nominal and ordinal data plus one additional strength: They incorporate the concept of equality of interval (the scaled distance between 1 and 2 equals the distance between 2 and 3). Calendar time is such a scale. For example, the elapsed time between 3 and 6 a.m. equals the time between 4 and 7 a.m. One cannot say, however, that 6 a.m. is twice as late as 3 a.m., because "zero time" is an arbitrary zero point. Centigrade and Fahrenheit temperature scales are other examples of classical interval scales. Both have an arbitrarily determined zero point, not a unique origin.

The product-moment correlation, t-tests, F-tests, and other parametric tests are the statistical procedures of choice for interval data.[8]

Researchers treat many attitude scales as interval, as we illustrate in the next chapter. When a scale is interval and the data are relatively symmetric with one mode, you use the arithmetic mean as the measure of central tendency. You can compute the average time of a TV promotional message or the average attitude value for different age groups in an insurance benefits study. The standard deviation is the measure of dispersion.

When the distribution of scores computed from interval data lean in one direction or the other (skewed right or left), we use the median as the measure of central tendency and the interquartile range as the measure of dispersion. The reasons for this are discussed in Chapter 16, Appendix 16a.

Ratio Scales

Ratio scales incorporate all of the powers of the previous scales plus the provision for absolute zero or origin. Ratio data represent the actual amounts of a variable. Measures of physical dimensions such as weight, height, distance, and area are examples. In the behavioral sciences, few situations satisfy the requirements of the ratio scale—the area of psychophysics offering some exceptions. In business research, we find ratio scales in many areas. There are money values, population counts, distances, return rates, productivity rates, and amounts of time (e.g., elapsed time in seconds before a customer service representative answers a phone inquiry).

Swatch's *BeatTime*—a proposed standard global time introduced at the 2000 Olympics that may gain favor as more of us participate in cross-time-zone chats (Internet or otherwise)—is a ratio scale. It offers a standard time with its origin at 0 beats (12 midnight in Biel, Switzerland, at the new Biel Meridian timeline). A day is composed of 1,000 beats, with a "beat" worth 1 minute, 26.4 seconds.[9]

With the Glacier project, Jason could measure a customer's age, the number of years he or she has attended, and the number of times a selection has been performed in the Glacier summer festival. These measures all generate ratio data. For practical purposes, however, the analyst would use the same statistical techniques as with interval data.

All statistical techniques mentioned up to this point are usable with ratio scales. Other manipulations carried out with real numbers may be done with ratio-scale values. Thus, multiplication and division can be used with this scale but not with the others mentioned. Geometric and harmonic means are measures of central tendency, and coefficients of variation may also be calculated for describing variability.

Researchers often encounter the problem of evaluating variables that have been measured on different scales. For example, the choice to purchase a product by a consumer is a nominal variable, and cost is a ratio variable. Certain statistical techniques require that the measurement levels be the same. Since the nominal variable does not have the characteristics of order, distance, or point of origin, we cannot create them artificially after the fact. The ratio-based salary variable, on the other hand, can be reduced. Rescaling product cost into categories (e.g., high, medium, low) simplifies the comparison. This example may be extended to other measurement situations—that is, converting or rescaling a variable involves reducing the measure from the more powerful and robust level to a lesser one.[10] The loss of measurement power with this decision means that lesser-powered statistics are then used in data analysis, but fewer assumptions for their proper use are required.

In summary, higher levels of measurement generally yield more information. Because of the measurement precision at higher levels, more powerful and sensitive statistical procedures can be used. As we saw with the candy bar example, when moving from a higher measurement level to a lower one, there is always a loss of information. Finally, when we collect information at higher levels, we can always convert, rescale, or reduce the data to arrive at a lower level.

> Sources of Measurement Differences

The ideal study should be designed and controlled for precise and unambiguous measurement of the variables. Since complete control is unattainable, error does occur. Much error is systematic (results from a bias), while the remainder is random (occurs erratically). One authority has pointed out several sources from which measured differences can come.[11]

The Prince Corporation image study starts here and is used throughout this chapter.

Assume you are conducting an ex post facto study of corporate citizenship of a multi-national manufacturer. The company produces family, personal, and household care products. The participants are residents of a major city. The study concerns the Prince Corporation, a large manufacturer with its headquarters and several major facilities located in the city. The objective of the study is to discover the public's opinions about the company's approach to health, social welfare, and the environment. You also want to know the origin of any generally held adverse opinions.

Ideally, any variation of scores among the respondents would reflect true differences in their opinions about the company. Attitudes toward the firm as an employer, as an ecologically sensitive organization, or as a progressive corporate citizen would be accurately expressed. However, four major error sources may contaminate the results: (1) the respondent, (2) the situation, (3) the measurer, and (4) the data collection instrument.

Error Sources

The Respondent

Opinion differences that affect measurement come from relatively stable characteristics of the respondent. Typical of these are employee status, ethnic group membership, social class, and nearness to manufacturing facilities. The skilled researcher will anticipate many of these dimensions, adjusting the design to eliminate,

>**snap**shot

In connection with an issue centering on privacy issues, the editors of *American Demographics* hired TNS Intersearch to conduct a study of adults regarding their behavior and attitudes relating to copyright infringement. The survey instrument for the telephone study asked 1,051 adult respondents several questions about activities that might or might not be considered copyright infringement. The lead question asked about specific copyright-related activities:

Do you know someone who has done or tried to do any of the following?

1. Copying software not licensed for personal use.
2. Copying a prerecorded videocassette such as a rental or purchased video.
3. Copying a prerecorded audiocassette or compact disc.
4. Downloading music free of charge from the Internet.
5. Photocopying pages from a book or magazine.

A subsequent question asked respondents, "In the future, do you think that the amount of (ACTIVITY) will increase, decrease, or stay the same?" where "(ACTIVITY)" relates to one of the five numbered elements. Also, each respondent was asked to select a phrase from a list of four phrases "that best describes how you feel about (ACTIVITY)" and to select a phrase from a list of four phrases that "best describes what you think may happen as a result of (ACTIVITY)." The last content question asked the degree to which respondents would feel favorably toward a company that provided "some type of media content for free": more favor-able, less favorable, or "it wouldn't impact your impression of the company." As you might expect, younger adults had different behaviors and attitudes compared to older adults on some indicants. What measurement issues were involved in this study?

www.intersearch.tnsofres.com

neutralize, or otherwise deal with them. However, even the skilled researcher may not be as aware of less obvious dimensions. The latter variety might be a traumatic experience a given participant had with the Prince Corporation, its programs, or its employees. Respondents may be reluctant to express strong negative (or positive) feelings, may purposefully express attitudes that they perceive as different from those of others, or may have little knowledge about Prince but be reluctant to admit ignorance. This reluctance to admit ignorance of a topic can lead to an interview consisting of "guesses" or assumptions, which, in turn, create erroneous data.

Respondents may also suffer from temporary factors like fatigue, boredom, anxiety, hunger, impatience, or general variations in mood or other distractions; these limit the ability to respond accurately and fully. Designing measurement scales that engage the participant for the duration of the measurement is crucial.

Situational Factors

Any condition that places a strain on the interview or measurement session can have serious effects on the interviewer-respondent rapport. If another person is present, that person can distort responses by joining in, by distracting, or by merely being there. If the respondents believe anonymity is not ensured, they may be reluctant to express certain feelings. Curbside or intercept interviews are unlikely to elicit elaborate responses, while in-home interviews more often do.

The Measurer

The interviewer can distort responses by rewording, paraphrasing, or reordering questions. Stereotypes in appearance and action introduce bias. Inflections of voice and conscious or unconscious prompting with smiles, nods, and so forth, may encourage or discourage certain replies. Careless mechanical processing—checking of the wrong response or failure to record full replies—will obviously distort findings. In the data analysis stage, incorrect coding, careless tabulation, and faulty statistical calculation may introduce further errors.

The Instrument

A defective instrument can cause distortion in two major ways. First, it can be too confusing and ambiguous. The use of complex words and syntax beyond participant comprehension is typical. Leading questions, ambiguous meanings, mechanical defects (inadequate space for replies, response-choice omissions, and poor printing), and multiple questions suggest the range of problems. Many of these problems are the direct result of operational definitions that are insufficient, resulting in an inappropriate scale being chosen or developed.

A more elusive type of instrument deficiency is poor selection from the universe of content items. Seldom does the instrument explore all the potentially important issues. The Prince Corporation study might treat company image in areas of employment and ecology but omit the company management's civic leadership, its support of local education programs, its philanthropy, or its position on minority issues. Even if the general issues are studied, the questions may not cover enough aspects of each area of concern. While we might study the Prince Corporation's image as an employer in terms of salary and wage scales, promotion opportunities, and work stability, perhaps such topics as working conditions, company management relations with organized labor, and retirement and other benefit programs should also be included.

> The Characteristics of Good Measurement

What are the characteristics of a good measurement tool? An intuitive answer to this question is that the tool should be an accurate counter or indicator of what we are interested in measuring. In addition, it should be easy and efficient to use. There are three major criteria for evaluating a measurement tool: validity, reliability, and practicality.

- *Validity* is the extent to which a test measures what we actually wish to measure.
- *Reliability* has to do with the accuracy and precision of a measurement procedure.
- *Practicality* is concerned with a wide range of factors of economy, convenience, and interpretability.[12]

In the following sections, we discuss the nature of these qualities and how researchers can achieve them in their measurement procedures.

Validity

Many forms of **validity** are mentioned in the research literature, and the number grows as we expand the concern for more scientific measurement. This text features two major forms: external and internal validity.[13] The *external validity* of research findings is the data's ability to be generalized across persons, settings, and times; we discussed this in reference to experimentation in Chapter 11, and more will be said in Chapter 15 on sampling.[14] In this chapter, we discuss only internal validity. **Internal validity** is further limited in this discussion to the ability of a research instrument to measure what it is purported to measure. Does the instrument really measure what its designer claims it does?

One widely accepted classification of validity consists of three major forms: (1) content validity, (2) criterion-related validity, and (3) construct validity (see Exhibit 12-5).[15]

The management-research question hierarchy discussed in Chapter 3 helps to reduce research questions into specific investigative and measurement questions that have content validity.

Content Validity

The **content validity** of a measuring instrument is the extent to which it provides adequate coverage of the investigative questions guiding the study. If the instrument contains a representative sample of the universe of subject matter of interest, then content validity is good. To evaluate the content validity of an instrument, one must first agree on what elements constitute adequate coverage. In the Prince Corporation study, we must decide what knowledge and attitudes

> **Exhibit 12-5** Summary of Validity Estimates

Type	What Is Measured	Methods
Content	Degree to which the content of the items adequately represents the universe of all relevant items under study.	• Judgmental • Panel evaluation with content validity ratio
[redacted]	[redacted]	[redacted]
Construct	Answers the question, "What accounts for the variance in the measure?"; attempts to identify the underlying construct(s) being measured and determine how well the test represents it (them).	• Judgmental • Correlation of proposed test with established one • Convergent-discriminant techniques • Factor analysis • Multitrait-multimethod analysis

are relevant to the measurement of corporate public image and then decide which forms of these opinions are relevant positions on these topics. In the Glacier study, Jason must first determine what factors are influencing customer satisfaction before determining if published indexes can be of value. If the data collection instrument adequately covers the topics that have been defined as the relevant dimensions, we conclude the instrument has good content validity.

A determination of content validity involves judgment. First, the designer may determine it through a careful definition of the topic, the items to be scaled, and the scales to be used. This logical process is often intuitive and unique to each research designer.

A second way is to use a panel of persons to judge how well the instrument meets the standards. The panel independently assesses the test items for an instrument as essential, useful but not essential, or not necessary. "Essential" responses on each item from each panelist are evaluated by a content validity ratio, and those meeting a statistical significance value are retained. In both informal judgments and this systematic process, "content validity is primarily concerned with inferences about test construction rather than inferences about test scores."[16]

It is important not to define content too narrowly. If you were to secure only superficial expressions of opinion in the Prince Corporation attitude survey, it would probably not have adequate content coverage. The research should delve into the processes by which these attitudes came about. How did the respondents come to feel as they do, and what is the intensity of feeling? The same would be true of MindWriter's evaluation of service quality and satisfaction. It is not enough to know a customer is dissatisfied. The manager charged with enhancing or correcting the program needs to know what processes, employees, parts, and time sequences within the CompleteCare program have led to that dissatisfaction.

Criterion-Related Validity

Criterion-related validity reflects the success of measures used for prediction or estimation. You may want to predict an outcome or estimate the existence of a current behavior or time perspective. An attitude scale that correctly forecasts the outcome of a purchase decision has predictive validity. An observational method that correctly categorizes families by current income class has concurrent validity. While these examples appear to have simple and unambiguous validity criteria, there are difficulties in estimating validity. Consider the problem of estimating family income. There is a knowable true income for every family, but we may find the figure difficult to secure. Thus, while the criterion is conceptually clear, it may be unavailable.

A researcher may want to develop a preemployment test that will predict sales success. There may be several possible criteria, none of which individually tells the full story. Total sales per salesperson may not adequately reflect territory market potential, competitive conditions, or the different profitability rates of various products. One might rely on the sales manager's overall evaluation, but how unbiased and accurate are such impressions? The researcher must ensure that the validity criterion used is itself "valid." Any criterion measure must be judged in terms of four qualities: (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability.[17]

A criterion is *relevant* if it is defined and scored in the terms we judge to be the proper measures of salesperson success. If you believe sales success is adequately measured by dollar sales volume achieved per year, then it is the relevant criterion. If you believe success should include a high level of penetration of large accounts, then sales volume alone is not fully relevant. In making this decision, you must rely on your judgment in deciding what partial criteria are appropriate indicants of salesperson success.

Freedom from bias is attained when the criterion gives each salesperson an equal opportunity to score well. The sales criterion would be biased if it did not show adjustments for differences in territory potential and competitive conditions.

A *reliable* criterion is stable or reproducible. An erratic criterion (using monthly sales, which are highly variable from month to month) can hardly be considered a reliable standard by which to judge performance on a sales employment test. Finally, the information specified by the criterion must be *available*. If it is not available, how much will it cost and how difficult will it be to secure? The amount of money and effort that should be spent on development of a criterion depends on the importance of the problem for which the test is used.

> Chapter 19 describes statistical techniques used to find correlation between variables.

Once there are test and criterion scores, they must be compared in some way. The usual approach is to correlate them. For example, you might correlate test scores of 40 new salespeople with first-year sales achievements adjusted to reflect differences in territorial selling conditions.

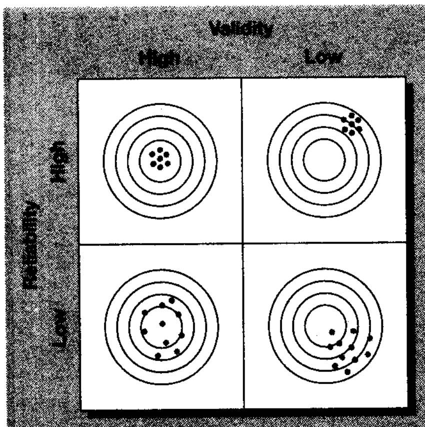
> An example of factor analysis is described in Chapter 20.

Construct Validity

In attempting to evaluate **construct validity,** we consider both the theory and the measuring instrument being used. If we were interested in measuring the effect of trust in cross-functional teams, the way in which "trust" was operationally defined would have to correspond to an empirically grounded theory. If a known measure of trust was available, we might correlate the results obtained using this measure with those derived from our new instrument. Such an approach would provide us with preliminary indications of *convergent* validity (the degree to which scores on one scale correlate with scores on other scales designed to assess the same construct). If Jason were to develop a customer satisfaction index for Glacier and, when compared, the results revealed the same indications as a predeveloped, established index, Jason's instrument would have convergent validity. Similarly, if Jason developed an instrument to measure satisfaction with the CompleteCare program and the derived measure could be confirmed with a standardized customer satisfaction measure, convergent validity would exist.

Returning to our example above, another method of validating the trust construct would be to separate it from other constructs in the theory or related theories. To the extent that trust could be separated from bonding, reciprocity, and empathy, we would have completed the first steps toward *discriminant* validity (the degree to which scores on a scale *do not* correlate with scores from scales designed to measure different constructs).

> **Exhibit 12-6** Understanding Validity and Reliability



Let's use an archer's bow and target as an analogy.

High reliability means that repeated arrows shot from the same bow would hit the target in essentially the same place—although not necessarily the intended place (first row of the graphic). If we had a bow with high validity as well, then every arrow would hit the bull's-eye (upper left panel). If reliability is low or decreases for some reason, arrows would be more scattered (lacking similarity or closeness, like those shown in the second row).

High validity means that the bow would shoot true every time. It would not pull to the right or send an arrow careening into the woods. Arrows shot from a high-validity bow will be clustered around a central point (the bull's eye), even when they are dispersed by reduced reliability (lower left graphic). Low-validity shots are distorted, rather like a flawed bow pulling our arrows to the right (second column of the graphic). We wouldn't hit the bull's-eye we were aiming at because the low-validity bow—like the flawed data collection instrument—would not perform as planned. When low validity is compounded by low reliability, the pattern of arrows is not only off the bull's-eye but also dispersed (lower right graphic).

We discuss the three forms of validity separately, but they are interrelated, both theoretically and operationally. Predictive validity is important for a test designed to predict product success. In developing such a test, you would probably first list the factors (constructs) that provide the basis for useful prediction. For example, you would advance a theory about the variables in product success—an area for construct validity. Finally, in developing the specific items for inclusion in the success prediction test, you would be concerned with how well the specific items sample the full range of each construct (a matter of content validity). Looking at Exhibit 12-6, we can better understand the concepts of validity and reliability by using an archer's bow and target as an analogy.

Reliability

A measure is reliable to the degree that it supplies consistent results. **Reliability** is a necessary contributor to validity but is not a sufficient condition for validity. The relationship between reliability and validity can be simply illustrated with the use of a bathroom scale. If the scale measures your weight correctly (using a concurrent criterion such as a scale known to be accurate), then it is both reliable and valid. If it consistently overweighs you by 6 pounds, then the scale is reliable but not valid. If the scale measures erratically from time to time, then it is not reliable and therefore cannot be valid. So if a measurement is not valid, it hardly matters if it is reliable—because it does not measure what the designer needs to measure in order to solve the research problem. In this context, reliability is not as valuable as validity, but it is much easier to assess.

Reliability is concerned with estimates of the degree to which a measurement is free of random or unstable error. Reliable instruments can be used with confidence that transient and situational factors are not interfering. Reliable instruments are robust; they work well at different times under different conditions. This distinction of time and condition is the basis for frequently used perspectives on reliability—stability, equivalence, and internal consistency (see Exhibit 12-7).

Stability

A measure is said to possess **stability** if you can secure consistent results with repeated measurements of the same person with the same instrument. An observation procedure is stable if it gives the same reading on a particular person when repeated one or more times. It is often possible to repeat observations on a subject and

> **Exhibit 12-7** Summary of Reliability Estimates

Type	Coefficient	What Is Measured	Methods
Test-Retest	Stability	Reliability of a test or instrument inferred from examinee scores; same test is administered twice to same subjects over an interval of less than six months.	Correlation
[illegible]	[illegible]	[illegible]	[illegible]
Split-Half, KR20, Cronbach's Alpha	Internal consistency	Degree to which instrument items are homogeneous and reflect the same underlying construct(s).	Specialized correlational formulas.

to compare them for consistency. When there is much time between measurements, there is a chance for situational factors to change, thereby affecting the observations. The change would appear incorrectly as a drop in the reliability of the measurement process.

Stability measurement in survey situations is more difficult and less easily executed than in observational studies. While you can observe a certain action repeatedly, you usually can resurvey only once. This leads to a test-retest arrangement—with comparisons between the two tests to learn how reliable they are. Some of the difficulties that can occur in the test-retest methodology and cause a downward bias in stability include:

- *Time delay between measurements*—leads to situational factor changes (also a problem in observation studies).
- *Insufficient time between measurements*—permits the respondent to remember previous answers and repeat them, resulting in biased reliability indicators.
- *Respondent's discernment of a study's disguised purpose*—may introduce bias if the respondent holds opinions related to the purpose but not assessed with current measurement questions.
- *Topic sensitivity*—occurs when the respondent seeks to learn more about the topic or form new and different opinions before the retest.

A suggested remedy is to extend the interval between test and retest (from two weeks to a month). While this may help, the researcher must be alert to the chance that an outside factor will contaminate the measurement and distort the stability score. Consequently, stability measurement through the test-retest approach has limited applications. More interest has centered on equivalence.

Equivalence

A second perspective on reliability considers how much error may be introduced by different investigators (in observation) or different samples of items being studied (in questioning or scales). Thus, while stability is concerned with personal and situational fluctuations from one time to another, **equivalence** is concerned with variations at one point in time among observers and samples of items. A good way to test for the equivalence of measurements by different observers is to compare their scoring of the same event. An example of this is the scoring of Olympic figure skaters by a panel of judges.

In studies where a consensus among experts or observers is required, the similarity of the judges' perceptions is sometimes questioned. How does a panel of supervisors render a judgment on merit raises, a new product's packaging, or future business trends? *Interrater reliability* may be used in these cases to correlate the observations or scores of the judges and render an index of how consistent their ratings are. In Olympic figure skating, a judge's relative positioning of skaters (determined by establishing a rank order for each judge and comparing each judge's ordering for all skaters) is a means of measuring equivalence.

The major interest with equivalence is typically not how respondents differ from item to item but how well a given set of items will categorize individuals. There may be many differences in response between two samples of items, but if a person is classified the same way by each test, then the tests have good equivalence.

One tests for item sample equivalence by using alternative or *parallel forms* of the same test administered to the same persons simultaneously. The results of the two tests are then correlated. Under this condition, the length of the testing process is likely to affect the subjects' responses through fatigue, and the inferred reliability of the parallel form will be reduced accordingly. Some measurement theorists recommend an interval between the two tests to compensate for this problem. This approach, called *delayed equivalent forms*, is a composite of test-retest and the equivalence method. As in test-retest, one would administer form X followed by form Y to half the examinees and form Y followed by form X to the other half to prevent "order-of-presentation" effects.[18]

The researcher can include only a limited number of measurement questions in an instrument. This limitation implies that a sample of measurement questions from a content domain has been chosen and another sample producing a similar number will need to be drawn for the second instrument. It is frequently difficult to create this second set. Yet if the pool is initially large enough, the items may be randomly selected for each instrument. Even with more sophisticated procedures used by publishers of standardized tests, it is rare to find fully equivalent and interchangeable questions.[19]

Internal Consistency

A third approach to reliability uses only one administration of an instrument or test to assess the **internal consistency** or homogeneity among the items. The *split-half* technique can be used when the measuring tool has many similar questions or statements to which the participant can respond. The instrument is administered and the results are separated by item into even and odd numbers or into randomly selected halves. When the two halves are correlated, if the results of the correlation are high, the instrument is said to have high reliability in an internal consistency sense. The high correlation tells us there is similarity (or homogeneity) among the items. The potential for incorrect inferences about high internal consistency exists when the test contains many items—which inflates the correlation index.

The Spearman-Brown correction formula is used to adjust for the effect of test length and to estimate reliability of the whole test.[20]

Practicality

The scientific requirements of a project call for the measurement process to be reliable and valid, while the operational requirements call for it to be practical. **Practicality** has been defined as *economy, convenience, and interpretability*.[21] While this definition refers to the development of educational and psychological tests, it is meaningful for business measurements as well.

Economy

Some trade-off usually occurs between the ideal research project and the budget. Data are not free, and instrument length is one area where economic pressures dominate. More items give more reliability, but in the interest of limiting the interview or observation time (and therefore costs), we hold down the number of measurement questions. The choice of data collection method is also often dictated by economic factors. The rising cost of personal interviewing first led to an increased use of telephone surveys and subsequently to the cur-rent rise in Internet surveys. In standardized tests, the cost of test materials alone can be such a significant expense that it encourages multiple reuse. Add to this the need for fast and economical scoring, and we see why computer scoring and scanning are attractive.

Convenience

A measuring device passes the convenience test if it is easy to administer. A questionnaire or a measurement scale with a set of detailed but clear instructions, with examples, is easier to complete correctly than one that lacks these features. In a well-prepared study, it is not uncommon for the interviewer instructions to be several times longer than the interview questions. Naturally, the more complex the concepts and constructs, the greater is the need for clear and complete instructions. We can also make the instrument easier to administer by giving close attention to its design and layout. While reliability and validity dominate our choices in design of scales here and later in Chapter 13, administrative difficulty should play some role. A long completion time, complex instructions, participant's perceived difficulty with the survey, and their rated enjoyment of the process also influence design. Layout issues include crowding of material, poor reproductions of illustrations, and the carryover of items from one page to the next or the need to scroll the screen when taking a Web survey. Both design and layout issues make completion of the instrument more difficult.

Interpretability

This aspect of practicality is relevant when persons other than the test designers must interpret the results. It is usually, but not exclusively, an issue with standardized tests. In such cases, the designer of the data collection instrument provides several key pieces of information to make interpretation possible:

- A statement of the functions the test was designed to measure and the procedures by which it was developed.
- Detailed instructions for administration.
- Scoring keys and instructions.
- Norms for appropriate reference groups.
- Evidence about reliability.
- Evidence regarding the intercorrelations of subscores.
- Evidence regarding the relationship of the test to other measures.
- Guides for test use.

>summary

1 While people measure things casually in daily life, research measurement is more precise and controlled. In measurement, one settles for measuring properties of the objects rather than the objects themselves. An event is measured in terms of its duration. What happened during it, who was involved, where it occurred, and so forth, are all properties of the event. To be more precise, what are measured are indicants of the properties. Thus, for duration, one measures the number of hours and minutes recorded. For what happened, one uses some system to classify types of activities that occurred. Measurement typically uses some sort of scale to classify or quantify the data collected.

2 There are four scale types. In increasing order of power, they are nominal, ordinal, interval, and ratio. Nominal scales classify without indicating order, distance, or unique origin. Ordinal data show magnitude relationships of more than and less than but have no distance

or unique origin. Interval scales have both order and distance but no unique origin. Ratio scales possess classification, order, distance, and unique origin.

3 Instruments may yield incorrect readings of an indicant for many reasons. These may be classified according to error sources: (a) the respondent or participant, (b) situational factors, (c) the measurer, and (d) the instrument.

4 Sound measurement must meet the tests of validity, reliability, and practicality. Validity reveals the degree to which an instrument measures what it is supposed to measure to assist the researcher in solving the research problem. Three forms of validity are used to evaluate measurement scales. Content validity exists to the degree that a measure provides an adequate reflection of the topic under study. Its determination is primarily judgmental and intuitive. Criterion-related validity relates to our ability to predict some outcome or estimate the existence of some current condition. Construct validity is the most complex and abstract. A measure has construct validity to the degree that it conforms to predicted correlations of other theoretical propositions.

A measure is reliable if it provides consistent results. Reliability is a partial contributor to validity, but a measurement tool may be reliable without being valid. Three forms of reliability are stability, equivalence, and internal consistency. A measure has practical value for the research if it is economical, convenient, and interpretable.

>keyterms

internal validity 318
interval scale 315
mapping rules 309
measurement 309
nominal scale 312
objects 311
ordinal scale 314
practicality 323
properties 311
ratio scale 315
reliability 321
  equivalence 322
  internal consistency 323
  stability 321
validity 318
  construct 320
  content 318
  criterion-related 319

>discussionquestions

Terms in Review

1 What can we measure about the four objects listed below? Be as specific as possible.
  a Laundry detergent
  b Employees
  c Factory output
  d Job satisfaction

2 What are the essential differences among nominal, ordinal, interval, and ratio scales? How do these differences affect the statistical analysis techniques we can use?

3 What are the four major sources of measurement error? Illustrate by example how each of these might affect measurement results in a face-to-face interview situation.

4 Do you agree or disagree with the following statements? Explain.
  a Validity is more critical to measurement than reliability.
  b Content validity is the most difficult type of validity to determine.
  c A valid measurement is reliable, but a reliable measurement may not be valid.
  d Stability and equivalence are essentially the same thing.

Making Research Decisions

5 You have data from a corporation on the annual salary of each of its 200 employees.
  a Illustrate how the data can be presented as ratio, interval, ordinal, and nominal data.
  b Describe the successive loss of information as the presentation changes from ratio to nominal.

6 Below are listed some objects of varying degrees of abstraction. Suggest properties of each of these objects that can be measured by each of the four basic types of scales.
  a Store customers.
  b Voter attitudes.
  c Hardness of steel alloys.
  d Preference for a particular common stock.
  e Profitability of various divisions in a company.

7 You have been asked by the head of marketing to design an instrument by which your private, for-profit school can evaluate the quality and value of its various curricula and courses. How might you try to ensure that your instrument has:
  a Stability?
  b Equivalence?

- c Internal consistency?
- d Content validity?
- e Predictive validity?
- f Construct validity?

8 A new hire at Mobil Oil, you are asked to assume the management of the *Mobil Restaurant Guide.* Each restaurant striving to be included in the guide needs to be evaluated. Only a select few restaurants may earn the five-star status. What dimensions would you choose to measure to apply the one to five stars in the *Mobil Restaurant Guide?*

9 You have been asked to develop an index of student morale in your department.
- a What constructs or concepts might you employ?
- b Choose several of the major concepts, and specify their dimensions.
- c Select observable indicators that you might use to measure these dimensions.
- d How would you compile these various dimensions into a single index?
- e How would you judge the reliability and/or validity of these measurements?

Bringing Research to Life

10 Given that Glacier Symphony has previously measured its customer satisfaction by survey, how might Jason assess the internal validity of the Glacier questionnaire?

From Concept to Practice

11 Using Exhibit 12-3 and one of the case questionnaires on your text CD, match each question to its appropriate scale type. For each scale type not represented, develop a measurement question that would be of that scale type.

>wwwexercise

Visit sites like those of The Gallup Organization, Harris Interactive, and Kaiser Family Foundation. Select a study and identify the measurement scale types and the measurement decisions made in the study.

>cases*

Campbell-Ewald: R-E-S-P-E-C-T Spells Loyalty

Donatos: Finding the New Pizza

NCRCC: Teeing Up and New Strategic Direction

NetConversions Influences Kelley Blue Book

Ramada Demonstrates Its *Personal Best*™

USTA: Come Out Swinging

Yahoo!: *Consumer Direct* Marries Purchase Metrics to Banner Ads

* All cases appear on the text CD; you will find abstracts of these cases in the Case Abstracts section of this text. Video cases are indicated with a video icon.

>chapter 13

Measurement Scales

“All survey questions have to be actionable if you want results.”

Frank Schmidt, senior scientist,
The Gallup Organization

>**learning**objectives

After reading this chapter, you should understand . . .

1 The nature of attitudes and their relationship to behavior.

2 The critical decisions involved in selecting an appropriate measurement scale.

3 The characteristics and use of rating, ranking, sorting, and other preference scales.

>bringingresearchtolife

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

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

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

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

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

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

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

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

"It's our company credo, 'Underpromise 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 scribbles his signature on the contract. "You'll receive this contract in three days, after it has wended its way past the

>cont'd

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.

"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," says Myra, sympathetically.

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

This chapter covers procedures that will help you understand measurement scales so that you might select or design measures that are appropriate for your research. We concentrate here on the problems of measuring more complex constructs, like attitudes. Conceptually, we start this process by revisiting the research process (see Exhibit 13-1) to understand where the act of scaling fits in the process.

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

> The Nature of Attitudes

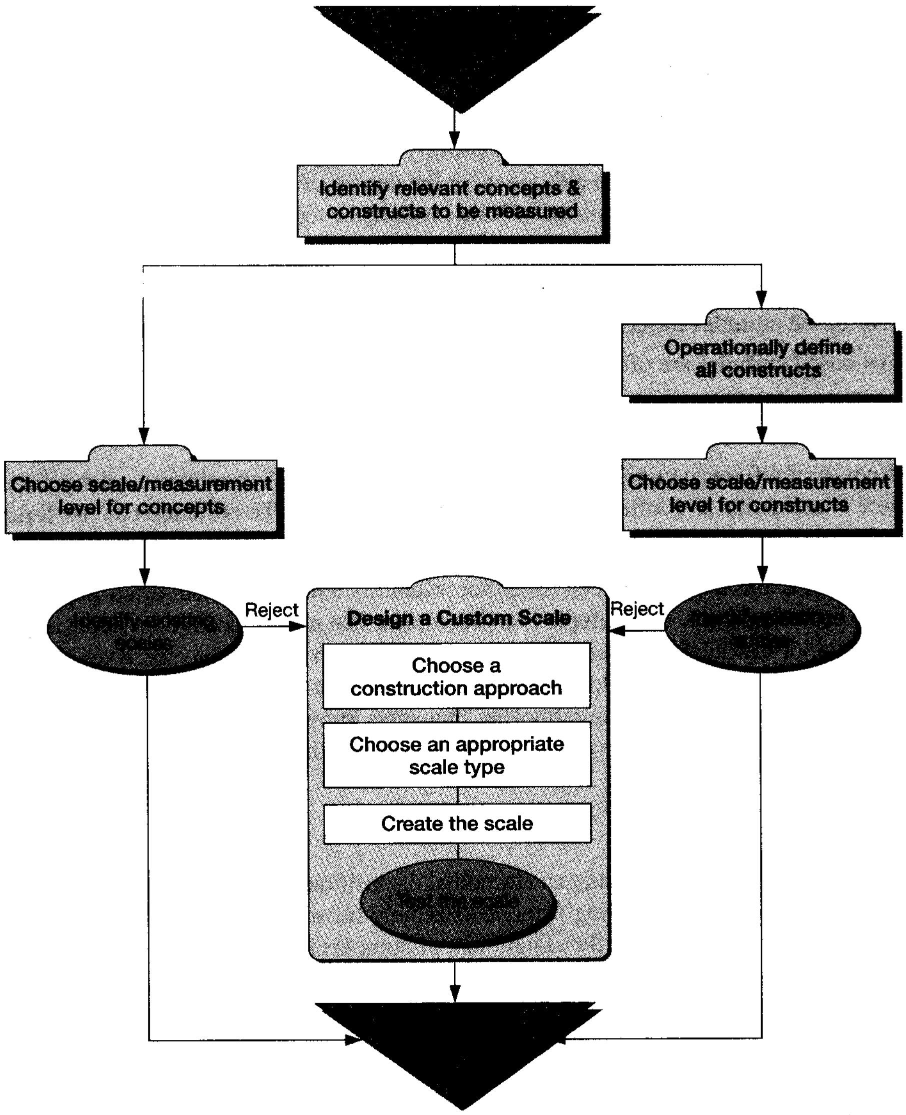
Jason is properly concerned about attitude measurement for the MindWriter study. But what is an attitude? There are numerous definitions, but one seems to capture the essence: An **attitude** is a learned, stable predisposition to respond to oneself, other persons, objects, or issues in a consistently favorable or unfavorable way.[1] Important aspects of this definition include the learned nature of attitudes, their relative permanence, and their association with socially significant events and objects. Because an attitude is a *predisposition,* it would seem that the more favorable one's attitude is toward a product or service, the more likely that the product or service will be purchased. But, as we will see, that is not always the case.

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

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

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

> **Exhibit 13-1** The Scaling Process


Identify relevant concepts & constructs to be measured
Operationally define all constructs
Choose scale/measurement level for concepts
Choose scale/measurement level for constructs
Identify existing scales
Reject
Design a Custom Scale
Reject
Choose a construction approach
Choose an appropriate scale type
Create the scale
Test the scale

The Relationship between Attitudes and Behavior

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

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

- Specific attitudes are better predictors of behavior than general ones.
- Strong attitudes (strength is affected by *accessibility* or how well the object is remembered and brought to consciousness, how extreme the attitude is, or the degree of confidence in it) are better predictors of behavior than weak attitudes composed of little intensity or topical interest.
- Direct experiences with the attitude object (when the attitude is formed, during repeated exposure, or through reminders) produce behavior more reliably.
- Cognitive-based attitudes influence behaviors better than affective-based attitudes.
- Affective-based attitudes are often better predictors of consumption behaviors.
- Using multiple measurements of attitude or several behavioral assessments across time and environments improves prediction.
- The influence of reference groups (interpersonal support, urges of compliance, peer pressure) and the individual's inclination to conform to these influences improves the attitude-behavior linkage.[2]

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

Attitude Scaling

Attitude scaling is the process of assessing an attitudinal disposition using a number that represents a person's score on an attitudinal continuum ranging from an extremely favorable disposition to an extremely unfavorable one. **Scaling** is the "procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some of the characteristics of numbers to the properties in question."[3] Procedurally, we assign numbers to indicants of the properties of objects. Thus, one assigns a number scale to the various levels of heat and cold and calls it a thermometer. To measure the temperature of the air, you know that a property of temperature is that its variation leads to an expansion or contraction of mercury. A glass tube with mercury provides an indicant of temperature change by the rise or fall of the mercury in the tube. Similarly, your attitude toward your university could be measured on numerous scales that capture indicators of the different dimensions of your awareness, feelings, or behavioral intentions toward the school.

> Selecting a Measurement Scale

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

- Research objectives
- Response types

- Data properties
- Number of dimensions
- Balanced or unbalanced
- Forced or unforced choices
- Number of scale points
- Rater errors

Research Objectives

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

- To measure characteristics of the participants who participate in the study.
- To use participants as judges of the objects or indicants presented to them.

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

Response Types

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

Data Properties

Decisions about the choice of measurement scales are often made with regard to the data properties generated by each scale. In Chapter 12, we said that we classify scales in increasing order of power; scales are nominal, ordinal, interval, or ratio. Nominal scales classify data into categories without indicating order, distance, or unique origin. Ordinal data show relationships of *more than* and *less than* but have no distance or unique origin. Interval scales have both order and distance but no unique origin. Ratio scales possess all four properties' features. The assumptions underlying each level of scale determine how a particular measurement scale's data will be analyzed statistically.

Number of Dimensions

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

Balanced or Unbalanced

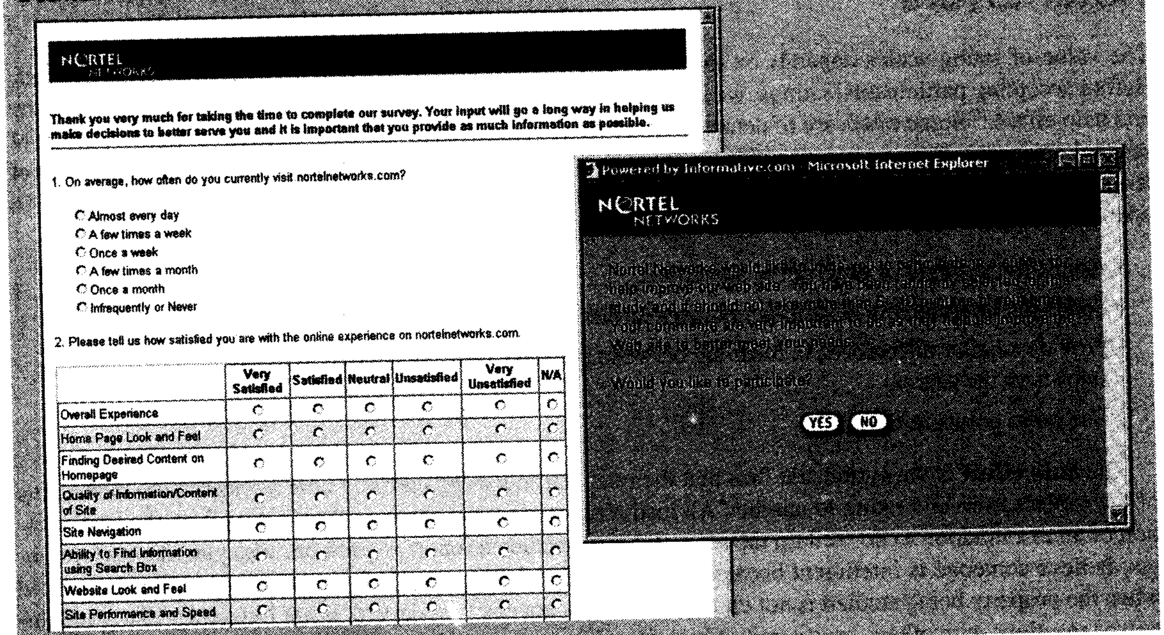
A **balanced rating scale** has an equal number of categories above and below the midpoint. Generally, rating scales should be balanced, with an equal number of favorable and unfavorable response choices. However, scales may be balanced with or without an indifference or midpoint option. A balanced scale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

The use of an unbalanced rating scale can be justified in studies where researchers know in advance that nearly all participants' scores will lean in one direction or the other. Raters are inclined to score attitude objects higher if the objects are very familiar and if they are ego-involved.[4] Brand-loyal customers are also expected to respond favorably. When researchers know that one side of the scale is not likely to be used, they try to achieve precision on the side that will most often receive the participant's attention. Unbalanced scales are also considered when participants are known to be either "easy raters" or "hard raters." An unbalanced scale can help compensate for the error of *leniency* created by such raters.

Forced or Unforced Choices

An **unforced-choice rating scale** provides participants with an opportunity to express no opinion when they are unable to make a choice among the alternatives offered. A **forcedchoice scale** requires that participants select one of the offered alternatives. Researchers often exclude the response choice "no opinion," "undecided," "don't know," "uncertain," or "neutral" when they know that most participants have an attitude on the topic. It is reasonable in this circumstance to constrain participants so that they focus on alternatives carefully and do not idly choose the middle position. However, when many participants are clearly undecided and the scale does not allow them to express their uncertainty, the forced-choice scale biases results. Researchers discover such bias when a larger percentage of participants express an attitude than did so in previous studies on the same issue. Some of this bias is attributable to participants providing meaningless responses or reacting to questions about which they have no attitudes (see Chapter 14). This affects the statistical measures of the mean and median, which shift toward the scale's midpoint, making it difficult to discern attitudinal differences throughout the instrument.[5] Understanding neutral answers is a challenge for researchers. In a customer satisfaction study

Online surveys are increasingly common due in large part to their speed in data collection. They also offer versatility for use with various types of measurement scales; flexibility in containing not only verbal but graphical, photographic, video, and digital elements; access to difficult-to-contact or inaccessible participants; and lower cost of large-sample completion. The visual appearance of the measurement scale is very important in getting the participant to click through to completion. This invitation from Nortel Networks and the opening screen of the questionnaire are designed to encourage participation. Informative, Inc., fielded this survey for Nortel Networks (designed to evaluate Nortel's Web site). The first screen of the questionnaire indicates two strategies: a multiple-choice, single-response strategy incorporating forced choice, and a multi-item rating grid which does not force choice (notice the NA column). If you look closely, you can also see a scroll bar on the first screen. Some designers will put only one question to a screen in Web questionnaires believing that participants who have to scroll may not fully complete the survey. This survey was designed for a technical audience, so that was not as much a concern. **www.nortelnetworks.com; www.informative.com**



Thank you very much for taking the time to complete our survey. Your input will go a long way in helping us make decisions to better serve you and it is important that you provide as much information as possible.

1. On average, how often do you currently visit nortelnetworks.com?

○ Almost every day
○ A few times a week
○ Once a week
○ A few times a month
○ Once a month
○ Infrequently or Never

2. Please tell us how satisfied you are with the online experience on nortelnetworks.com.

	Very Satisfied	Satisfied	Neutral	Unsatisfied	Very Unsatisfied	N/A
Overall Experience	○	○	○	○	○	○
Home Page Look and Feel	○	○	○	○	○	○
Finding Desired Content on Homepage	○	○	○	○	○	○
Quality of Information/Content of Site	○	○	○	○	○	○
Site Navigation	○	○	○	○	○	○
Ability to Find Information using Search Box	○	○	○	○	○	○
Website Look and Feel	○	○	○	○	○	○
Site Performance and Speed	○	○	○	○	○	○

Nortel Networks would like to [illegible] help improve our Web site. [illegible] study and it should not take more than [illegible] Your comments are very important to [illegible] Web site to better meet your needs.

Would you like to participate?

that focused on the overall satisfaction question with a company in the electronics industry, an unforced scale was used. Study results, however, revealed that 75 percent of those in the "neutral" participant group could be converted to brand loyalists if the company excelled (received highly favorable ratings) on only 2 of the 26 other scaled questions in the study.[6] Thus, the participants in the neutral group weren't truly neutral, and a forced-choice scale would have revealed the desired information.

Number of Scale Points

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

As we noted in Chapter 12, the characteristics of reliability and validity are important factors affecting measurement decisions. First, as the number of scale points increases, the *reliability* of the measure increases.[7] Second, in some studies, scales with 11 points may produce more *valid* results than 3-, 5-, or 7-point scales.[8] Third, some constructs require greater measurement sensitivity and the opportunity to extract more variance, which additional scale points provide. Fourth, a larger number of scale points are needed to produce accuracy when using single-dimension versus multiple-dimension scales.[9] Finally, in cross-cultural measurement, the cultural practices may condition participants to a standard metric—a 10-point scale in Italy, for example.

Rater Errors

The value of rating scales depends on the assumption that a person can and will make good judgments. Before accepting participants' ratings, we should consider their tendencies to make errors of central tendency and halo effect.[10] Some raters are reluctant to give extreme judgments, and this fact accounts for the **error of central tendency.** Participants may also be "easy raters" or "hard raters," making what is called an **error of leniency.** These errors most often occur when the rater does not know the object or property being rated. To address these tendencies, researchers can:

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

The **halo effect** is the systematic bias that the rater introduces by carrying over a generalized impression of the subject from one rating to another. An instructor expects the student who does well on the first question of an examination to do well on the second. You conclude a report is good because you like its form, or you believe someone is intelligent because you agree with him or her. Halo is especially difficult to avoid when the property being studied is not clearly defined, is not easily observed, is not frequently discussed, involves reactions with others, or is a trait of high moral importance.[11] Ways of counteracting the halo effect include having the participant rate one trait at a time, revealing one trait per page (as in an Internet survey, where the participant cannot return to change his or her answer), or periodically reversing the terms that anchor the endpoints of the scale, so positive attributes are not always on the same end of each scale.

> Rating Scales

In Chapter 12, we said that questioning is a widely used stimulus for measuring concepts and constructs. For example, a researcher asks questions about participant's attitudes toward the taste of a soft drink. The responses are "thirst quenching," "sour," "strong bubbly," "orange taste," and "syrupy." These answers alone do not provide a means of discerning the degree of favorability and thus would be of limited value to the researcher. However, with a properly constructed scale, the researcher could develop a taste profile for the target brand. We use rating scales to judge properties of objects without reference to other similar objects. These ratings may be in such forms as "like—dislike," "approve—indifferent—disapprove," or other classifications using even more categories.

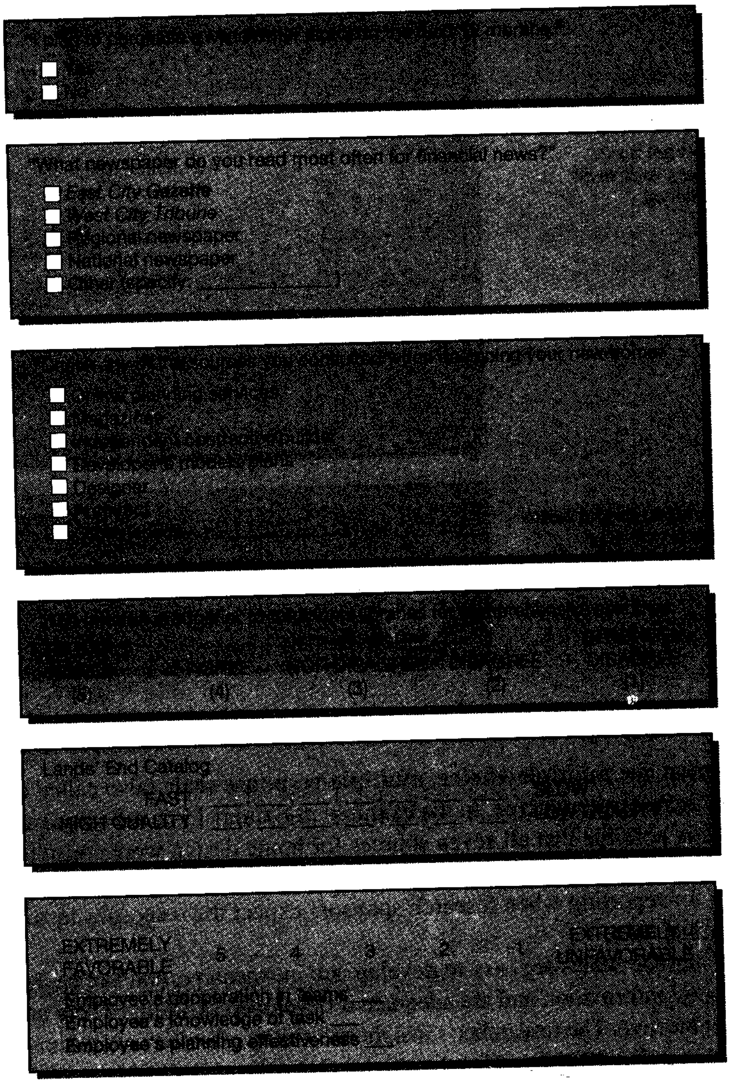
Examples of rating scales we discuss in this section are shown in Exhibit 13-2. Since this exhibit amplifies the overview presented in this section, we will refer you to the exhibit frequently.[12]

Simple Attitude Scales

The **simple category scale** (also called a *dichotomous scale*) offers two mutually exclusive response choices. In Exhibit 13-2 they are "yes" and "no," but they could just as easily be "important" and "unimportant," "agree" and "disagree," or another set of discrete categories if the question were different. This response strategy is particularly useful for demographic questions or where a dichotomous response is adequate.

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

> Exhibit 13-2 Sample Rating Scales



Simple Category Scale (dichotomous)
data: nominal

[illegible]
□ [illegible]
□ [illegible]

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

[illegible]
□ [illegible]
□ [illegible]
□ [illegible]
□ [illegible]
□ [illegible]
□ [illegible]
□ [illegible]

Likert Scale Summated Rating
data: interval

[illegible]
(5) (4) (3) (2) (1)

Semantic Differential Scale
data: interval

Lands' End Catalog
FAST [illegible]
HIGH QUALITY [illegible]

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 ____

> Exhibit 13-2 Cont'd

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

Constant-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):"

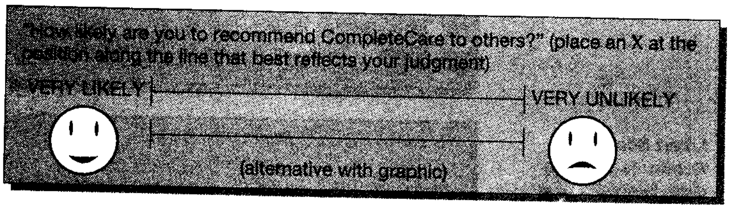
Being one of the lowest-cost suppliers	□
All other aspects of supplier performance	□
Sum	100

Stapel Scale
data: ordinal or* interval

(Company Name)

Technology Leader	Exciting Products	World-Class Reputation
+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
+1	+1	+1
−1	−1	−1
−2	−2	−2
−3	−3	−3
−4	−4	−4
−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)

* In chapter 12 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.

A variation, the **multiple-choice, multiple-response scale** (also called a *checklist*), allows the rater to select one or several alternatives. In the example in Exhibit 13-2, we are measuring seven items with one question, and it is possible that all seven sources for home design were consulted. The cumulative feature of this scale can be beneficial when a complete picture of the participant's choice is desired, but it may also present a problem for reporting when research sponsors expect the responses to sum to 100 percent. This scale generates nominal data.

Simple attitude scales are easy to develop, are inexpensive, and can be designed to be highly specific. They provide useful information and are adequate if developed skillfully. There are also weaknesses. The design approach is subjective. The researcher's insight and ability offer the only assurance that the items chosen are a representative sample of the universe of attitudes about the attitude object. We have no evidence that each person

>snapshot

Campbell-Ewald, an award-winning integrated communications agency headquartered in Detroit, believes it is good business to treat customers with respect—and the agency can prove it. As part of a major research initiative to discover why customer relationship management (CRM) solutions were falling short of expectations, Campbell-Ewald mailed more than 5,000 surveys to adults 18 or older who were customers in each of three business sectors: insurance, automotive, and retail. The goal? To answer the question: "Does respect influence customer loyalty and, thereby, purchasing?" With partner research company, Synovate and Campbell-Ewald clients, three surveys were developed. Each included 27 to 29 attitudinal statements that queried the adults on how they defined respect and the importance of respect to purchase behavior in each sector. Customers responded to the statements using a 5-point scale (strongly agree to strongly disagree). Using analysis of the results, Campbell-Ewald validated the relevance of its five "People Principles," which, in turn, have helped clients like General Motors, Continental Airlines, and Farmers Insurance incorporate respectful behavior into their business practices. The five "People Principles" of respect are:

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

How would you operationalize the construct of respect? To learn more about this research, read the case: "Campbell-Ewald: R-E-S-P-E-C-T Spells Loyalty".

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

will view all items with the same frame of reference as will other people. While such scales are frequently used, there has been a great effort to develop construction techniques that overcome some of their deficiencies.

Likert Scales

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

In Exhibit 13-2, the participant 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. Likert scales also use 7 and 9 scale points. The values for each choice are normally not printed on the instrument, but they are shown in Exhibit 13-2 to illustrate the scoring system.

The Likert scale has many advantages that account for its popularity. It is easy and quick to construct.[13] Conscientious researchers are careful that each item meets an empirical test for discriminating ability between favorable and unfavorable attitudes. Likert scales are probably more reliable and provide a greater volume of data than many other scales. The scale produces interval data.

Originally, creating a Likert scale involved a procedure known as *item analysis.* In the first step, a large number of statements were collected that met two criteria: (1) Each statement was relevant to the attitude being studied; (2) each was believed to reflect a favorable or unfavorable position on that attitude. People similar to those who are going to be studied were asked to read each statement and to state the level of their agreement with it, using a 5-point scale. A scale value of 1 indicated a strongly unfavorable attitude (strongly disagree). The other intensities were 2 (disagree), 3 (neither agree nor disagree), 4 (agree), and 5 (strongly agree), a strongly favorable attitude (see Exhibit 13-2). To ensure consistent results, the assigned numerical values are reversed if the statement is worded negatively (1 is always strongly unfavorable and 5 is always strongly favorable). Each person's responses are then added to secure a total score. The next step is to array these total scores and select some portion representing the highest and lowest total scores (generally defined as the top and bottom 10 to 25 percent of the distribution). The middle group (50 to 80 percent of participants) are excluded from the subsequent analysis.

The two extreme groups represent people with the most favorable and least favorable attitudes toward the attitude being studied. These extremes are the two criterion groups by which individual items are evaluated. **Item analysis** assesses each item based on how well it discriminates between those persons whose total score is high and those whose total score is low. It involves calculating the mean scores for each scale item among the low scorers and high scorers. The mean scores for the high-score and low-score groups are then tested for statistical significance by computing *t* values. (In evaluating response patterns of the high and low groups to the statement "My digital camera's features are exciting," we secure the results shown in Exhibit 13-3.) After finding the *t* values for each statement, they are rank-ordered, and those statements with the highest *t* values are selected. The 20 to 25 items that have the highest *t* values (statistically significant differences between mean scores) are selected for inclusion in the final scale.[14] Researchers have found that a larger number of items for each attitude object improve the reliability of the scale. As an approximate indicator of a statement's discrimination power, one authority also suggests using only those statements whose *t* value is 1.75 or greater, provided there are 25 or more subjects in each group.[15]

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

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

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

Semantic Differential Scales

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

> Exhibit 13-3 Evaluating a Scale Statement by Item Analysis

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

	Low Total Score Group				High Total Score Group			
Response Categories	X	f	fX	X(fX)	X	f	fX	X(fX)
① (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	②	73	285	1,183
		n_L	ΣX_L	$\Sigma X(fX)_L$		n_H	ΣX_H	$\Sigma X(fX)_H$

$$\bar{X}_L = \frac{177}{73} = 2.42 \longleftarrow ③ \longrightarrow \bar{X}_H = \frac{285}{73} = 3.90$$

$$\Sigma(X_L - \bar{X}_L)^2 = 503 - \frac{(177)^2}{73} \longleftarrow ④ \longrightarrow \Sigma(X_H - \bar{X}_H)^2 = 1,183 - \frac{(285)^2}{73}$$

$$= 73.84 \qquad\qquad = 70.33$$

$$t = \frac{\bar{X}_H - \bar{X}_L}{\sqrt{\frac{\Sigma(X_H - \bar{X}_H)^2 + \Sigma(X_L - \bar{X}_L)^2}{n(n-1)}}} \longleftarrow ⑤$$

$$= \frac{3.90 - 2.42}{\sqrt{\frac{70.33 + 73.84}{73(73-1)}}} = 8.92 \qquad ⑥$$

Legend:

① For each of the response categories, the scale's value (X) is multiplied by the frequency or number of participants (f) who chose that value. These values produce the product (fX). This number is then multiplied by X. For example, there are 3 participants in the low-score group who scored a 5 (strongly agreed with the statement): (fX) = (3) (5) = 15; (X)(fX) = (5) (15) = 75.

② The frequencies and products are summed.

③ A mean score for each group is computed.

④ Deviation scores are computed, squared, and summed as required by the formula.

⑤ The data are tested in a modified t-test that compares high- and low-scoring groups for the item. Notice the mean scores in the numerator of the formula.

⑥ The calculated value is compared with a criterion, 1.75. If the calculated value (in this case, 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 measurement item and repeat the process.

signers have long known that they must use words, shapes, associations, and images to activate a person's connotative meanings.

Results of the thesaurus study are shown in Exhibit 13-4.

Osgood and his associates developed the semantic differential method to measure the psychological meanings of an object to an individual.[16] They produced a list of 289 bipolar adjective pairs, which were reduced to 76 pairs and formed into rating scales for attitude research. Their analysis allowed them to conclude that semantic space is multidimensional rather than

>snapshot

The American Heart Association (AHA), the premier research association studying cardiovascular diseases and stroke, has exceptionally low awareness: Less than 16 percent of adults could identify the AHA from unaided-recall measure. Using unaided recall, selected participants from the target market are asked to bring to mind advertisements they have seen or heard recently. For a nonprofit that depends on donations to support its mission, low awareness and unclear brand identity translate to lower revenue. As a result, AHA took a very important step in 2004: It decided to venture into its first-ever paid advertising campaign. The $36 million effort spans three years.

Campbell-Ewald (C-E) was chosen to develop the campaign. Extensive secondary data analysis, expert interviews on both heart disease and how people identify and deal with risk, and focus groups were used to understand the problem. C-E found that to some degree the AHA's success had created a misconception—that heart disease is not a major health threat. AHA research has led to significant advances in treatment for heart disease. Open-heart surgery that now often keeps a person in the hospital less than a week, arterial surgery (stent implants and angioplasty) that has patients returning to everyday activities in less than a week, and drugs that unclot arteries merely reinforce the erroneous belief that heart disease should be lower on people's risk scale. "What we found is that people didn't have the correct facts about heart disease," concluded C-E's executive vice president and director of strategic planning, Arthur Mitchell. "We knew any campaign would need to help people understand the true risk of heart disease and what the American Heart Association really does." This understanding led C-E to depart from the approach used in past public service announcements for the American Heart Association that typically focused on getting the audience to adopt behaviors (like exercising, stopping smoking, etc.) to lower their risk.

The campaign developed from the research features real stories about people who have experienced heart disease on a very personal level. And people are encouraged to visit the AHA Web site to take a quiz on heart disease facts. Web site visits will be an important measure of the campaign's success, as well as subsequent unaided-recall measures and attitude shifts. What measurement scales would you use to test the effectiveness of this campaign during the next three years and after its completion?

www.americanheart.org; www.campbell-ewald.com

You can learn more about this research by reading the case "Campbell-Ewald Campaign Pumps Awareness into American Heart Association."

unidimensional. Three factors contributed most to meaningful judgments by participants: (1) evaluation, (2) potency, and (3) activity. These concepts from the historical thesaurus study (Exhibit 13-4) illustrate the wide applicability of the technique to persons, abstract concepts, events, institutions, and physical objects.[17]

Researchers have followed a somewhat different approach to SD scales than did the original study advocates. They have developed their own adjectives or phrases and have focused on the evaluative dimension more often (which might help explain the popularity of the Likert scale). The positive benefit is that the scales created have been adapted to specific management questions. One study explored a retail store image using 35 pairs of words or phrases classified into eight groups. These word pairs were especially created for the study. Excerpts from this scale are presented in Exhibit 13-5. 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.

The semantic differential has several advantages. It is an efficient and easy way to secure attitudes from a large sample. These attitudes may be measured in both direction and intensity. The total set of responses provides a comprehensive picture of the meaning of an object and a measure of the person doing the rating. It is a standardized technique that is easily repeated but escapes many problems of response distortion found with more direct methods. It produces interval data. Basic instructions for constructing an SD scale are found in Exhibit 13-6.

In Exhibit 13-7 we see a scale being used by a panel of corporate leaders evaluating candidates for a high-level position in their industry's lobbying association. The selection of the concepts is driven by the

> Exhibit 13-4 Results of the Thesaurus Study

Evaluation (E)	Potency (P)	Activity (A)
Good–bad	Hard–soft	Active–passive
Positive–negative	Strong–weak	Fast–slow
Optimistic–pessimistic	Heavy–light	Hot–cold
Complete–incomplete	Masculine–feminine	Excitable–calm
Timely–untimely	Severe–lenient	
	Tenacious–yielding	

Subcategories of Evaluation			
Meek Goodness	Dynamic Goodness	Dependable Goodness	Hedonistic Goodness
Clean–dirty	Successful–unsuccessful	True–false	Pleasurable–painful
Kind–cruel	High–low	Reputable–disreputable	Beautiful–ugly
Sociable–unsociable	Meaningful–meaningless	Believing–skeptical	Sociable–unsociable
Light–dark	Important–unimportant	Wise–foolish	Meaningful–meaningless
Altruistic–egotistical	Progressive–regressive	Healthy–sick	
Grateful–ungrateful	Clean–dirty		
Beautiful–ugly			
Harmonious–dissonant			

Source: Adapted from Charles E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning* (Urbana: University of Illinois Press, 1957), table 5, pp. 52–61.

> Exhibit 13-5 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.

> Exhibit 13-6 Steps in Constructing an SD Scale

1. Select the concepts: nouns, noun phrases, or nonverbal stimuli such as visual sketches. Concepts are chosen by judgment and reflect the nature of the investigative question. In the MindWriter study, one concept might be "Call Center accessibility."
2. Select bipolar word pairs or phrase pairs appropriate to your needs. If the traditional Osgood adjectives are used, several criteria guide your selection:
   - Three bipolar pairs are required when using evaluation, potency, and activity. Scores on these individual items can be averaged, by factor, to improve their reliability.
   - The scale should be relevant to the concepts being judged. Choose adjectives that allow connotative perceptions to be expressed. Irrelevant concept-scale pairings yield neutral midpoint values that convey little information.
   - Scales should be stable across raters and concepts. A pair such as "large–small" may be interpreted as denotative when judging a physical object such as "automobile" but may be used connotatively with abstract concepts such as "product quality."
   - Scales should be linear between polar opposites and pass through the origin. A pair that fails this test is "rugged–delicate," which is nonlinear on the evaluation dimension. When used separately, both adjectives have favorable meanings.*
3. Create the scoring system and assign a weight to each point on the scale. The negative signs in the original scoring procedure (−3, −2, −1, 0, +1, +2, +3) were found to produce coding errors, and the 0 point is arbitrary. Most SD scales have 7 points: 7, 6, 5, 4, 3, 2, and 1.
4. As with Likert scales, about half of the adjective pairs are randomly reversed to minimize the halo effect.

* Charles E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning* (Urbana: University of Illinois Press, 1957).

> Exhibit 13-7 SD Scale for Analyzing Industry Association Candidates

[illegible]

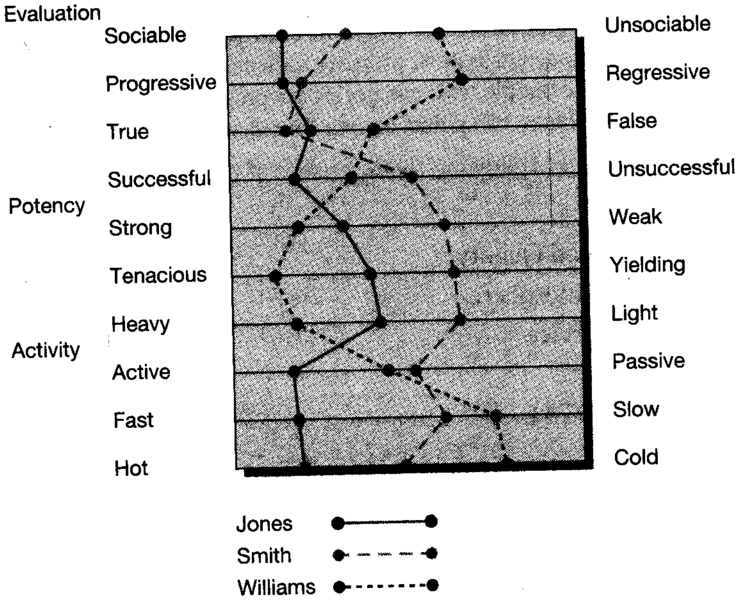
(E) [illegible]
(P) [illegible]
(A) [illegible]
(E) [illegible]
(P) [illegible]
(A) [illegible]
(E) [illegible]
(P) [illegible]
(A) [illegible]
(E) [illegible]

characteristics they believe the candidate must possess to be successful in advancing their agenda. There are three candidates.

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

The data are plotted in a "snake diagram" in Exhibit 13-8. Here the adjective pairs are reordered so that evaluation, potency, and activity descriptors are grouped together, with the ideal factor reflected by the left side of the scale.

> Exhibit 13-8 Graphic Representation of SD Analysis



Numerical/Multiple Rating List Scales

Numerical scales have equal intervals that separate their numeric scale points, as shown in Exhibit 13-2. The verbal anchors serve as the labels for the extreme points. Numerical scales are often 5-point scales but may have 7 or 10 points. The participants write a number from the scale next to each item. If numerous questions about a product's performance were included in the example, the scale would provide both an absolute measure of importance and a relative measure (ranking) of the various items rated. The scale's linearity, simplicity, and production of ordinal or interval data make it popular for managers and researchers. When evaluating a new product concept, purchase intent is frequently measured with a 5- to 7-point numerical scale, with the anchors being "definitely would buy" and "definitely would not buy."

A **multiple rating list scale** (Exhibit 13-2) is similar to the numerical scale but differs in two ways: (1) It accepts a circled response from the rater, and (2) the layout facilitates visualization of the results. The advantage is that a mental map of the participant's evaluations is evident to both the rater and the researcher. This scale produces interval data.

Stapel Scales

The **Stapel scale** is used as an alternative to the semantic differential, especially when it is difficult to find bipolar adjectives that match the investigative question. In the example in Exhibit 13-2 there are three attributes of corporate image. The scale is composed of the word (or phrase) identifying the image dimension and a set of 10 response categories for each of the three attributes.

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

Constant-Sum Scales

A scale that helps the researcher discover proportions is the **constant-sum scale.** With a constant-sum scale, the participant allocates points to more than one attribute or property indicant, such that they total a constant sum, usually 100 or 10. In the Exhibit 13-2 example, two categories are presented that must sum to 100. In the restaurant example, the participant distributes 100 points among four categories:

> You have 100 points to distribute among the following characteristics of the Dallas Steakhouse. Indicate the relative importance of each attribute:
>
> _____ Food Quality
> _____ Atmosphere
> _____ Service
> _____ Price
> 100 TOTAL

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

Graphic Rating Scales

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

Never X__________________________ Always

Other graphic rating scales (see Exhibit 13-2) use pictures, icons, or other visuals to communicate with the rater and represent a variety of data types. Graphic scales are often used with children, whose more limited vocabulary prevents the use of scales anchored with words.

> Ranking Scales

In ranking scales, the participant directly compares two or more objects and makes choices among them. Frequently, the participant is asked to select one as the "best" or the "most preferred." When there are only two choices, this approach is satisfactory, but it often results in ties when more than two choices are found. For example, assume participants are asked to select the most preferred among three or more models of a product. In response, 40 percent choose model A, 30 percent choose model B, and 30 percent choose model C. Which is the preferred model? The analyst would be taking a risk to suggest that A is most preferred. Perhaps that interpretation is correct, but 60 percent of the participants chose some model other than A.

> Exhibit 13-9 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."

___ BMW Z4
___ Porsche Boxster

___ Chevrolet Corvette
___ Porsche Boxster

___ Chevrolet Corvette
___ BMW Z4

___ Porsche Boxster
___ Dodge Viper

___ Chevrolet Corvette
___ Dodge Viper

___ Dodge Viper
___ BMW Z4

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 hair dryer's performance, the new one is":

SUPERIOR ... ABOUT THE SAME ... INFERIOR

___ 1 ___ 2 ___ 3 ___ 4 ___ 5

Perhaps all B and C voters would place A last, preferring either B or C to A. This ambiguity can be avoided by using some of the techniques described in this section.

Using the **paired-comparison scale,** the participant can express attitudes unambiguously by choosing between two objects. Typical of paired comparisons would be the sports car preference example in Exhibit 13-9. The number of judgments required in a paired comparison is $[(n)(n - 1)/2]$, where n is the number of stimuli or objects to be judged. When four cars are evaluated, the participant evaluates six paired comparisons $[(4)(3)/2 = 6]$.

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?

> **Exhibit 13-10** Response Patterns of 200 Heavy Users' Paired Comparisons on Five Alternative Package Designs

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

	Designs				
	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
[illegible]	3	1	2	[illegible]	[illegible]

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

In another example we might compare packaging design proposals considered by a brand manager (see Exhibit 13-10). Generally, there are more than two stimuli to judge, resulting in a potentially tedious task for participants. If 15 suggestions for design proposals are available, 105 paired comparisons would be made.

Reducing the number of comparisons per participant without reducing the number of objects can lighten this burden. You can present each participant with only a sample of the stimuli. In this way, each pair of objects must be compared an equal number of times. Another procedure is to choose a few objects that are believed to cover the range of attractiveness at 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 participants will tire to the point that they give ill-considered answers or refuse to continue. Opinions differ about the upper limit, but five or six stimuli are not unreasonable when the participant has other questions to answer. If the data collection consists only of paired comparisons, as many as 10 stimuli are reasonable. A paired comparison provides ordinal data.

The **forced ranking scale,** shown in Exhibit 13-9, lists attributes that are ranked relative to each other. This method is faster than paired comparisons and is usually easier and more motivating to the participant. With five items, it takes 10 paired comparisons to complete the task, and the simple forced ranking of five is easier. Also, ranking has no transitivity problem where A is preferred to B, and B to C, but C is preferred to A—although it also forces a false unidimensionality.

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

>snapshot

Should Northwest Airlines, Marriott, or Alaskan Airlines attempt to attract the business of Americans with disabilities? If so, what would it take to capture the segment? Eric Lipp, executive director of the Open Doors Organization (ODO), an advocacy organization for those with disabilities, sponsored a study to find out. High on his agenda was providing an incentive to the travel industry to make accommodations to attract the 22 million adults with disabilities who have traveled in the last two years on 63 million trips—and who may want to travel more. "We now estimate that Americans with disabilities currently spent $13.2 billion in travel expenditures and that amount would at least double [to $27.2 billion] if travel businesses were more attuned to the needs of those with disabilities."

ODO hired Harris Interactive, a global market research and consulting firm best known for The Harris Poll and for pioneering the Internet method to conduct scientifically accurate market research. Harris Interactive conducted a hybrid study via both online and phone surveys to determine the magnitude of the disability travel segment, its purchasing power, and the accommodations the segment needed to increase travel. "Those with disabilities can't all be reached with one method," explained Laura Light, project director with Harris Interactive. "The nature of their physical limitation might preclude one method or the other." And how did the firm evaluate all the possible accommodations—from Braille safety cards on airplanes to a designated person to handle problems in a hotel? Harris Interactive used its proprietary COMPASS™ methodology, which uses paired comparisons as a measurement tool. "COMPASS™ saves the participant time and energy," explained Light. "Even with a long list, COMPASS™ can be done quickly." In the ODO study, COMPASS™ was used twice: once to measure 17 possible airline accommodations and once to measure 23 possible hotel accommodations. By having each participant evaluate only a portion of the large number of accommodation pairs rather than the full list (136 for airline accommodations and 253 for hotel accommodations), each question was answered in under four minutes. By using this process with all members of the sample, Harris Interactive is able to rank order the items and measure the magnitude of difference between items. This makes it easier for Delta, Marriott, or Alaskan Airlines to make the right choices about accommodations for those with disabilities.

www.opendoorsnfp.org; www.harrisinteractive.com

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

Often the manager is interested in benchmarking. This calls for a standard by which other programs, processes, brands, point-of-sale promotions, or people can be compared. The **comparative scale** is ideal for such comparisons if the participants are familiar with the standard. In the Exhibit 13-9 example, the standard is the participant's previous hair dryer. The new dryer is being assessed relative to it. The provision to compare yet other dryers to the standard is not shown in the example but is nonetheless available to the researcher.

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

> Sorting

Q-sorts require sorting of a deck of cards into piles that represent points along a continuum. The participant—or judge—groups the cards based on his or her response to the concept written on the card. Researchers using Q-sort resolve three special problems: item selection, structured or unstructured choices in sorting, and data analysis. The basic Q-sort procedure involves the selection of a set of verbal statements, phrases, single words, or photos related to the concept being studied. For statistical stability, the number of cards should not be less than 60; and for convenience, not be more than 120. After the cards are created, they are shuffled, and the participant is instructed to sort the cards into a set of piles (usually 7 to 11), each pile

>closeup

Jason has been working on scaling for MindWriter's CompleteCare project for a week when the request comes to Myra Wines to report her progress to MindWriter's general manager. He has narrowed the choice to the three scales in Exhibit 13-11: a Likert scale, a numerical rating scale with two verbal anchors, and a hybrid expectation scale. All are 5-point scales that are presumed to measure at the interval level.

He needs a statement that can accompany the scale for preliminary evaluation. Returning to their list of investigative questions, he finds a question that seems to capture the essence of the repair process: "Are customers' problems

>Exhibit 13-11 Alternative Scales Considered for MindWriter

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

Numerical 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

representing a point on the judgment continuum. The left-most pile represents the concept statements, which are "most valuable," "favorable," "agreeable," and so forth. The right-most pile contains the least favorable cards. The researcher asks the participant to fill the center, or neutral, pile with the cards about which the participant is indecisive. In the case of a *structured* sort, the distribution of cards allowed in each pile is predetermined. With an *unstructured* sort, only the number of piles will be determined. Although the distribution of cards in most structured sorts resembles a normal distribution, there is some controversy about analyzing the data as ranking (ordinal data) versus interval data.

The purpose of sorting is to get a conceptual representation of the sorter's attitude toward the attitude object and to compare the relationships between people. The relative ranking of concepts allows researchers to derive clusters of individuals possessing similar preferences. By varying the instructions, the technique can be used to describe products, services, behavioral intentions, and a host of other applications. In the example below, participants are asked to complete a structured sort of cards containing the names of magazines. The scale values and the number of cards in each pile are predetermined, although the distribution in this case represents a normal statistical distribution.