

MEASUREMENT AND SCALING CONCEPTS

What Do I Measure?

The decision statement, corresponding research questions, and research hypotheses can be used to decide what concepts need to be measured in a given project. **Measurement** is the process of describing some property of a phenomenon of interest, usually by assigning numbers in a reliable and valid way. The numbers convey information about the property being measured. When numbers are used, the researcher must have a rule for assigning a number to an observation in a way that provides an accurate description.

Measurement can be illustrated by thinking about the way instructors assign students' grades. A grade represents a student's performance in a class. Students with higher performance should receive a different grade than do students with lower performance. Even the apparently simple concept of student performance is measured in many different ways. Consider the following options:

1. A student can be assigned a letter corresponding to his/her performance.
 - a. A — Represents excellent performance
 - b. B — Represents good performance
 - c. C — Represents average performance
 - d. D — Represents poor performance
 - e. F — Represents failing performance
2. A student can be assigned a number from 1 to 20.
 - a. 20 — Represents outstanding performance
 - b. 11—20 — Represents differing degrees of passing performance
 - c. Below 11 — Represents failing performance
3. A student can be assigned a number corresponding to a percentage performance scale.
 - a. 100 percent — Represents a perfect score. All assignments are performed correctly.
 - b. 60—99 percent — Represents differing degrees of passing performance, each number representing the proportion of correct work.
 - c. 0—59 percent — Represents failing performance but still captures proportion of correct work.
4. A student can be assigned one of two letters corresponding to performance.
 - a. P — Represents a passing mark
 - b. F — Represents a failing mark

Actually, this is not terribly different than a manager who must assign performance scores to employees. In each case, students with different marks are distinguished in some way. However, some scales may better distinguish students. Each scale also has the potential of producing error or some lack of validity. Exhibit 13.2 illustrates a common measurement application.

Often, instructors may use a percentage scale all semester long and then be required to assign a letter grade for a student's overall performance. Does this produce any measurement problems? Consider two students who have percentage scores of 79.4 and 70.0, respectively. The most likely outcome when these scores are translated into "letter grades" is that each receives a C (the common 10-point spread would yield a 70—80 percent range for a C). Consider a third student who finishes with a 69.0 percent average and a fourth student who finishes with a 79.9 percent average.

Which students are happiest with this arrangement? The first two students receive the same grade, even though their scores are 9.4 percent apart. The third student gets a grade lower (D) performance than the second student, even though their percentage scores are only 1.0 percentage point different. The fourth student, who has a score only 0.5 percent higher than the first student, would receive a B. Thus, the measuring system (final grade) suggests that the fourth student outperformed the first (assuming that 79.9 is rounded up to 80) student (B versus C), but the first student did not outperform

the second (each gets a C), even though the first and second students have the greatest difference in percentage scores.





	Student	Percentage Grade	Difference from Next Highest Student	Letter Grade
	1	79.4%	0.5%	C
	2	70.0%	9.4%	C
	3	69.0%	1.0%	D
	4	79.9%	NA	B

EXHIBIT 13.2
Are There Any Validity Issues with This Measurement?

A strong case can be made that error exists in this measurement system. All measurement, particularly in the social sciences, contains error. Researchers, if we are to represent concepts truthfully, must make sure that the measures used, if not perfect, are accurate enough to yield correct conclusions. Ultimately, research and measurement are tied closely together.

Concepts

A researcher has to know what to measure before knowing how to measure something. The problem definition process should suggest the concepts that must be measured. A **concept** can be thought of as a generalized idea that represents something of meaning. Concepts such as *age*, *sex*, *education*, and *number of children* are relatively concrete properties. They present few problems in either definition or measurement. Other concepts are more abstract. Concepts such as *loyalty*, *personality*, *channel power*, *trust*, *corporate culture*, *customer satisfaction*, *value*, and so on are more difficult to both define and measure. For example, *loyalty* has been measured as a combination of *customer share* (the relative proportion of a person's purchases going to one competing brand/store) and *commitment* (the degree to which a customer will sacrifice to do business with a brand/store).¹ Thus, we can see that loyalty consists of two components, the first is behavioral and the second is attitudinal.

Operational Definitions

Researchers measure concepts through a process known as **operationalization**. This process involves identifying scales that correspond to variance in the concept. **Scales**, just as a scale you may use to check your weight, provide a range of values that correspond to different values in the concept being measured. In other words, scales provide **correspondence rules** that indicate that a certain value on a scale corresponds to some true value of a concept. Hopefully, they do this in a truthful way.

Here is an example of a correspondence rule: "Assign the numbers 1 through 7 according to how much trust that you have in your sales representative. If the sales representative is perceived as completely untrustworthy, assign the numeral 1, if the sales rep is completely trustworthy, assign a 7."

■ VARIABLES

Researchers use variance in concepts to make diagnoses. Therefore, when we defined variables in an earlier chapter, we really were suggesting that variables capture different concept values. Scales capture variance in concepts and, as such, the scales provide the researcher's variables. Thus, for practical purposes, once a research project is underway, there is little difference between a concept and a variable. Consider the following hypothesis:

H1: Experience is positively related to job performance.

The hypothesis implies a relationship between two variables, experience and job performance. The variables capture variance in the experience and performance concepts. One employee may have 15 years of experience and be a top performer. A second may have 10 years experience and be a good performer. The scale used to measure experience is quite straightforward in this case and would involve simply providing the number of years an employee has been with the company. Job performance, on the other hand, can be quite complex, as described in the opening vignette.

■ CONSTRUCTS

Sometimes, a single variable cannot capture a concept alone. Using multiple variables to measure one concept can often provide a more complete account of some concept than could any single variable. Even in the physical sciences, multiple measurements are often used to make sure an accurate representation is obtained. In social science, many concepts are measured with multiple measurements.

A **construct** is a term used for concepts that are measured with multiple variables. For instance, when a business researcher wishes to measure the customer orientation of a salesperson, several variables like these may be used, each captured on a 1—5 scale:

1. I offer the product that is best suited to a customer's problem.
2. A good employee has to have the customer's best interests in mind.
3. I try to find out what kind of products will be most helpful to a customer.

Constructs can be very helpful in operationalizing a concept.

An operational definition is like a manual of instructions or a recipe: even the truth of a statement like "Gaston Gourmet likes key lime pie" depends on the recipe. Different instructions lead to different results. In other words, how we define the construct will affect the way we measure it.

Levels of Scale Measurement

Business researchers use many scales or number systems. Not all scales capture the same richness in a measure. Not all concepts require a rich measure. Traditionally, the level of scale measurement is seen as important because it determines the mathematical comparisons that are allowable. The four levels or types of scale measurement are **nominal, ordinal, interval, and ratio level scales**.

1-Nominal Scale

Nominal scales represent the most elementary level of measurement. A nominal scale assigns a value to an object for identification or classification purposes only. The value can be, but does not have to be, a number because no quantities are being represented. In this sense, a nominal scale is truly a qualitative scale. Nominal scales are extremely useful, and are sometimes the only appropriate measure, even though they can be considered elementary.

Business researchers use nominal scales quite often. Suppose Barq's Root Beer was experimenting with three different types of sweeteners (cane sugar, corn syrup, or fruit extract). The researchers would like the experiment to be blind, so when subjects are asked to taste one of the three root beers, the drinks are labeled A, B, or C, not cane sugar, corn syrup, or fruit extract. Or, a researcher interested in examining the production efficiency of a company's different plants might refer to them

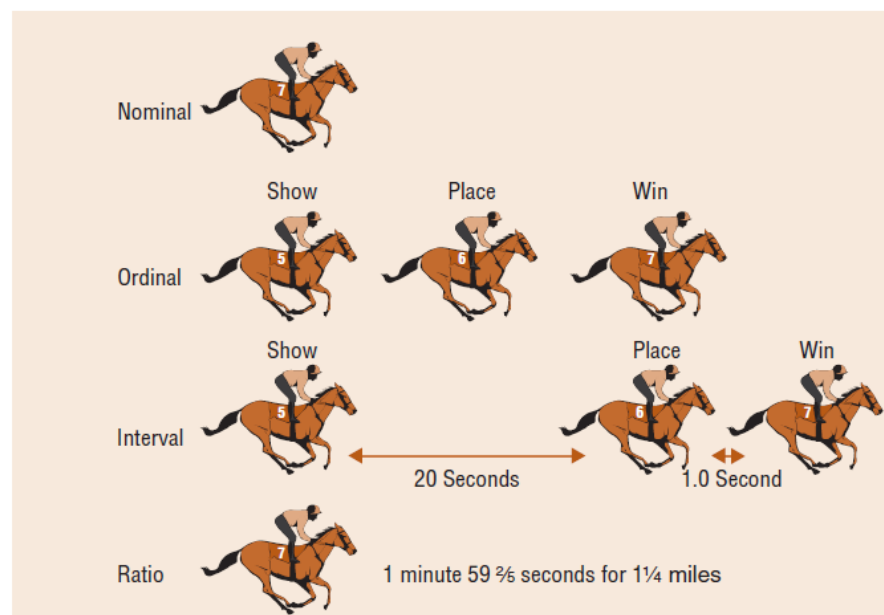
as “Plant 1,” “Plant 2,” and so forth.

Level	Examples	Numerical Operations	Descriptive Statistics
Nominal	Student ID number Yes – No Male – Female Buy – Did Not Buy East region Central region West region	Counting	<ul style="list-style-type: none"> • Frequencies • Mode
Ordinal	Student class rank Please rank your three favorite movies. Choose from the following: <ul style="list-style-type: none"> • Dissatisfied • Satisfied • Very satisfied • Delighted Indicate your level of education: <ul style="list-style-type: none"> • Some high school • High school diploma • Some college • College degree • Graduate degree 	Counting Ordering	<ul style="list-style-type: none"> • Frequencies • Mode • Median • Range
Interval	Student grade point average (GPA) Temperature (Celsius and Fahrenheit) Points given on an essay question 100-point job performance rating provided by supervisor	Common arithmetic operations	<ul style="list-style-type: none"> • Frequencies • Mode • Median • Range • Mean • Variance • Standard deviation
Ratio	Amount spent on last purchase Salesperson sales volume Number of stores visited on a shopping trip Annual family income Time spent viewing a Web page	All arithmetic operations	<ul style="list-style-type: none"> • Frequencies • Mode • Median • Range • Mean • Variance • Standard deviation

EXHIBIT 13.5 Facts About the Four Levels of Scales

Nominal scaling is arbitrary. What we mean is that each label can be assigned to any of the categories without introducing error. For instance, in the root beer example above, the researcher can assign the letter C to any of the three options without damaging scale validity. The researcher could just as easily use numbers instead of letters, as in the plant efficiency example, and vice versa. If so, cane sugar, corn syrup, and fruit extract might be identified with the numbers 1, 2, and 3, respectively, or even 543, 26, and 2010, respectively. The important thing to note is the numbers are not representing different quantities or the value of the object. Thus any set of numbers, letters, or any other identification is equally valid.

EXHIBIT 13.4 Nominal, Ordinal, Interval, and Ratio Scales Provide Different Information



We encounter nominal numbering systems all the time. Sports uniform numbers are nominal numbers. Ben Roethlisberger is identified on the football field by his jersey number. School bus numbers are nominal in that they simply identify a bus. Elementary school buses sometimes use both a number and an animal designation to help small children get on the right bus. So, bus number “8” may also be the “tiger” bus, but it could just as easily be the “horse” bus or the “cardinal” bus.

The first drawing in Exhibit 13.4 depicts the number 7 on a horse’s colors. This is merely a label to allow bettors and racing enthusiasts to identify the horse. The assignment of a 7 to this horse does not mean that it is the seventh fastest horse or that it is the seventh biggest, or anything else meaningful. But the 7 does let you know when you have won or lost your bet.

2-Ordinal Scale

Ordinal scale allows things to be arranged in order based on how much of some concept they possess. In other words, an ordinal scale is a order based on how much of some concept they ranking scale. In fact, we often use the term *rank order* to describe an ordinal scale. When class rank for high school students is determined, we have used an ordinal scale. We know that the student ranked seventh finished ahead of the student ranked eighth, who finished ahead of the ninth ranked student. However, we do not really know what the actual GPA was or how close these three students are to each other in overall grade point average.

Research participants often are asked to *rank* things based on preference. So, preference is the concept, and the ordinal scale lists the options from most to least preferred, or vice versa. Five objects can be ranked from 1—5 (least preferred to most preferred) or 1—5 (most preferred to least preferred) with no loss of meaning. In this sense, ordinal scales are somewhat arbitrary, but not nearly as arbitrary as a nominal scale.

When business professors take some time off and go to the race track, even they know that a horse finishing in the “show” position has finished after the “win” and “place” horses. The order of finish can be accurately represented by an ordinal scale using an ordered number rule:

- Assign 1 to the “win” position
- Assign 2 to the “place” position
- Assign 3 to the “show” position

Perhaps the winning horse defeated the place horse by a nose, but the place horse defeated the show horse by 20 seconds. The ordinal scale does not tell how far apart the horses were, but it is good enough to let someone know the result of a wager. Typical ordinal scales in business research ask respondents to rank their three favorite brands, have personnel managers rank potential employees after job interviews, or judge investments as “buy,” “hold,” or “sell.” Researchers know how each item, person, or stock is judged relative to others, but they do not know by how much.

3-Interval Scale

Interval scales have both nominal and ordinal properties, but they also capture information about differences in quantities of a concept. So, not only would a sales manager know that a particular salesperson outperformed a colleague, information that would be available with an ordinal measure, but the manager would know by how much. If a professor assigns grades to term papers using a numbering system ranging from 1.0—20.0, not only does the scale represent the fact that a student with a 16.0 outperformed a student with 12.0, but the scale would show by how much (4.0).

The third drawing in Exhibit 13.4 depicts a horse race in which the win horse is one second ahead of the place horse, which is 20 seconds ahead of the show horse. Not only are the horses identified by the order of finish, but the difference between each horse’s performance is known. So, horse number 7 and horse number 6 performed similarly (1 second apart), but horse number 5 performed not nearly as well (20 seconds slower).

The classic example of an interval scale is temperature. Consider the following weather:

- June 6 was 80° F
- December 7 was 40° F

The interval Fahrenheit scale lets us know that December 7 was 40° F colder than June 6. But, we cannot conclude that December 7 was twice as cold as June 6. Although the actual numeral 80 is indeed twice as great as 40, remember that this is a scaling system. In this case, the scale is not iconic, meaning that it does not exactly represent some phenomenon. In other words, there is no naturally occurring zero point—a temperature of 0° does not mean an absence of heat (or cold for that matter).

Since temperature scales are interval, the gap between the numbers remains constant (i.e., the difference between 20° and 30° is 10°, just as the difference between 68° and 78° is 10°). This is an important element of interval scales and allows us to convert one scale to another. In this case, we can convert Fahrenheit temperatures to Celsius scale. Then, the following would result:

- June 6 was 26.7° C
- December 7 was 4.4° C

Obviously, now we can see that December 7 was not twice as cold as June 6. December 7 was 40° F or 22.3° C cooler, depending upon your thermometer. Interval scales are very useful because they capture relative quantities in the form of distances between observations. No matter what thermometer is used, December 7 was colder than June 6.

4- Ratio Scale

Ratio scales represent the highest form of measurement in that they have all the properties of interval scales with the additional attribute of representing absolute quantities. Interval scales possess only relative meaning, whereas ratio scales represent absolute meaning. In other words, ratio scales provide iconic measurement.

Zero, therefore, has meaning in that it represents an absence of some concept. An absolute zero is the defining characteristic differentiating between ratio and interval scales. For example, money is a way to measure economic value. Consider the following items offered for sale in an online auction:

- “Antique” 1970s digital watch—did not sell and there were no takers for free
- Gold-filled Elgin wristwatch circa 1950—sold for \$100
- Vintage stainless-steel Omega wristwatch—sold for \$1,000
- Antique rose gold Patek Philippe “Top Hat” wristwatch—sold for \$9,000

We can make the ordinal conclusions that the Patek was worth more than the Omega, and the Omega was worth more than the Elgin. All three of these were worth more than the 1970s digital watch. We can make interval conclusions such as that the Omega was worth \$900 more than the Elgin. We can also conclude that the Patek was worth nine times as much as the Omega and that the 1970s watch was worthless (selling price = \$0.00). The latter two conclusions are possible because price represents a ratio scale.

The fourth drawing in Exhibit 13.4 shows the time it took horse 7 to complete the race. If we know that horse 7 took 1 minute 59 $\frac{2}{5}$ seconds to finish the race, and we know the time it took for all the other horses, we can determine the time difference between horses 7, 6, and 5. In other words, if we knew the ratio information regarding the performance of each horse—the time to complete the race—we could determine the interval level information and the ordinal level information. However, if we only knew the ordinal level information, we could not create the interval or ratio information. Similarly, with only the interval level data we cannot create the ratio level information.

Mathematical and Statistical Analysis of Scales

While it is true that mathematical operations can be performed with numbers from nominal scales, the result may not have a great deal of meaning. For instance, a school district may perform mathematical operations on the nominal school bus numbers. With this, they may find that the average school bus number is 77.7 with a standard deviation of 20.5. Will this help them use the buses more efficiently or better assign bus routes? Probably not. Can a professor judge the quality of her classes by the average ID number? While it could be calculated, the result is meaningless. Thus, although you can put numbers into formulas and perform calculations with almost any numbers, the researcher has to know the meaning behind the numbers before meaningful conclusions can be drawn.

1- DISCRETE MEASURES

Discrete measures are those that take on only one of a finite number of values. A discrete scale is most often used to represent a classification variable. Therefore, discrete scales do not represent intensity of measures, only membership. Common discrete scales include any yes-or-no response, matching, color choices, or practically any scale that involves selecting from among a small number of categories. Thus, when someone is asked to choose from the following responses

- Disagree
- Neutral
- Agree

the result is a discrete value that can be coded 1, 2, or 3, respectively. This is also an ordinal scale to the extent that it represents an ordered arrangement of agreement. Nominal and ordinal scales are discrete measures.

Certain statistics are most appropriate for discrete measures. Exhibit 13.5 shows statistics for each scale level. The largest distinction is between statistics used for discrete versus continuous measures. For instance, the central tendency of discrete measures is best captured by the mode. When a student wants to know what the most likely grade is for MGT 341, the mode will be very useful. Observe the results below from the previous semester

A	3 Students	D	3 Students
B	9 Students	F	1 Student
C	6 Students		

The mode is a “B” since more students obtained that value than any other value. Therefore, the “average” student would expect a B in MGT 341.

2-CONTINUOUS MEASURES

Continuous measures are those assigning values anywhere along some scale range in a place that corresponds to the intensity of some concept. Ratio measures are continuous measures. Thus, when Griff measures sales for each salesperson using the dollar amount sold, he is assigning a continuous measure. A number line could be constructed ranging from the least amount sold to the most, and a spot on the line would correspond exactly to a salesperson’s performance.

Strictly speaking, interval scales are not necessarily continuous. Consider the following common type of survey question:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
I enjoy participating in online auctions	1	2	3	4	5

This is a discrete scale because only the values 1, 2, 3, 4, or 5 can be assigned. Furthermore, it is an ordinal scale because it only orders based on agreement. We really have no way of knowing that the difference in agreement of somebody marking a 5 instead of a 4 is the same as the difference in agreement of somebody marking a 2 instead of a 1. Therefore, the mean is not an appropriate way of stating central tendency and, technically, we really shouldn't use many common statistics on these responses.

The researcher should keep in mind, however, the distinction between ratio and interval measures. Errors in judgment can be made when interval measures are treated as ratio. For example, attitude is usually measured with an interval scale. An attitude of zero means nothing. In fact, attitude would only have meaning in a relative sense. In other words, attitude takes on meaning when one person's response is compared to another or through some other comparison. A single attitude score alone contains little useful information.

The mean and standard deviation may be calculated from continuous data. Using the actual quantities for arithmetic operations is permissible with ratio scales. Thus, the ratios of scale values are meaningful. A ratio scale has all the properties of nominal, ordinal, and interval scales. However, the same cannot be said in reverse. An interval scale, for example, has ordinal and nominal properties, but it does not have ratio properties.

Index Measures

Earlier, we distinguished constructs as concepts that require multiple variables to measure them adequately. Likewise, a consumer's attitude toward some product is usually a function of multiple attributes. An **attribute** is a single characteristic or fundamental feature of an object, person, situation, or issue.

Indexes and Composites

Multi-item instruments for measuring a construct are called *index measures*, or *composite measures*. An **index measure** assigns a value based on how much of the concept being measured is associated with an observation. Indexes often are formed by putting several variables together. For example, a social class index might be based on three weighted variables: occupation, education, and area of residence. Usually, occupation is seen as the single best indicator and would be weighted highest. With an index, the different attributes may not be strongly correlated with each other. A person's education does not always relate strongly to their area of residence. The American Consumer Satisfaction Index shows how satisfied American consumers are based on an index of satisfaction scores. Readers are likely not surprised to know that Americans appear more satisfied with soft drinks than they are with cable TV companies based on this index.

Composite measures also assign a value based on a mathematical derivation of multiple variables. For example, salesperson satisfaction may be measured by combining questions such as "How satisfied are you with your job? How satisfied are you with your territory? How satisfied are you with the opportunity your job offers?" For most practical applications, composite measures and indexes are computed in the same way.⁷

Computing Scale Values

Exhibit 13.6 demonstrates how a composite measure can be created from common rating scales. This scale was developed to assess how much a consumer trusts a Web site. This particular composite represents a **summated scale**. A summated scale is created by simply summing the response to each item making up the composite measure. For this scale, a respondent that judged the Web site as extremely trustworthy would choose S.T (value of 5) for each question. Across the five questions, this respondent's score would be 25. Conversely, a respondent that thought the Web site was very untrustworthy would chose SD (value of 1) for each question; a total of 5. Most respondents

would likely be somewhere between these extremes. For the example respondent in Exhibit 13.6, the summated scale score would be 13 based on his responses to the five items ($2 + 3 + 2 + 2 + 4 = 13$). A researcher may sometimes choose to average the scores rather than summing them. The advantage to this is that the composite measure is expressed on the same scale (1—5 rather than 5—25) as the original items. So, instead of a 13, the consumer would have a score of 2.6. While this approach might be more easily understood, the information contained in either situation (13 versus 2.6) is the same.

EXHIBIT 13.6

Computing a Composite Scale

Item	Strongly Disagree (SD) → Strongly Agree (SA)				
This site appears to be more trustworthy than other sites I have visited.	SD	(D)	N	A	SA
My overall trust in this site is very high.	SD	D	(N)	A	SA
My overall impression of the believability of the information on this site is very high.	SD	(D)	N	A	SA
My overall confidence in the recommendations on this site is very high.	SD	(D)	N	A	SA
The company represented in this site delivers on its promises.	SD	D	N	(A)	SA
Computation: Scale Values: SD = 1, D = 2, N = 3, A = 4, SA = 5					
Thus, the Trust score for this consumer is $2 + 3 + 2 + 2 + 4 = 13$					

Sometimes, a response may need to be reverse-coded before computing a summated or averaged scale value. **Reverse coding** means that the value assigned for a response is treated oppositely from the other items. If a sixth item was included on the Web site trust scale that said, “I do not trust this Web site,” reverse coding would be necessary to make sure the composite made sense. For example, the respondent that judged the Web site is extremely trustworthy would choose S.T for the first five items, then SD for the sixth. We can see that we would not want to just add these up, as this score of 21 would not really reflect someone that felt very positive about the trustworthiness of the site. Since the content of the sixth item is the reverse of trust (distrust), so the scale itself should be reversed. Thus, on a 5-point scale, the values are reversed as follows:

- 5 becomes 1
- 4 becomes 2
- 3 stays 3
- 2 becomes 4
- 1 becomes 5

After the reverse coding, our respondent that felt the Web site was trustworthy would have a sum- mated score of 25, which does correctly reflect a very positive attitude. If the respondent described in Exhibit 13.6 responded to this new item with a S.T (5), it would be reverse coded as a 1 before computing the summated scale. Thus, the summated scale value for the six items would become 14. The process of reverse coding is discussed in the Research Snapshot on the next page titled “Recoding Made Easy.”

Three Criteria for Good Measurement

The three major criteria for evaluating measurements are **reliability, validity, and sensitivity**.

1-Reliability

Reliability is an indicator of a measure’s internal consistency. Consistency is the key to understand reliability. A measure is reliable when different attempts at measuring something converge on the same result. For example, consider an exam that has three parts: 25 multiple-choice questions, 2

essay questions, and a short case. If a student gets 20 of the 25 (80 percent) multiple-choice questions correct, we would expect she would also score about 80 percent on the essay and case portions of the exam. Further, if a professor's research tests are reliable, a student should tend toward consistent scores on all tests. In other words, a student who makes an 80 percent on the first test should make scores close to 80 percent on all subsequent tests. Another way to look at this is that the student who makes the best score on one test will exhibit scores close to the best score in the class on the other tests. If it is difficult to predict what students would make on a test by examining their previous test scores, the tests probably lack reliability or the students are not preparing the same each time.

So, the concept of reliability revolves around consistency. Think of a scale to measure weight. You would expect this scale to be consistent from one time to the next. If you stepped on the scale and it read 140 pounds, then got off and back on, you would expect it to again read 140. If it read 110 the second time, while you may be happier, the scale would not be reliable.

2-INTERNAL CONSISTENCY

Internal consistency represents a measure's homogeneity. An attempt to measure trustworthiness may require asking several similar but not identical questions, as shown in Exhibit 13.6. The set of items that make up a measure are referred to as a *battery* of scale items. *Internal consistency* of a multiple-item measure can be measured by correlating scores on subsets of items making up a scale.

The **split-half method** of checking reliability is performed by taking half the items from a scale (for example, odd-numbered items) and checking them against the results from the other half (even-numbered items). The two scale *halves* should produce similar scores and correlate highly. The problem with split-half method is determining the two halves. Should it be even- and odd- numbered questions? Questions 1—3 compared to 4—6? Coefficient alpha provides a solution to this dilemma.

Coefficient alpha (α) is the most commonly applied estimate of a multiple-item scale's reliability. Coefficient α represents internal consistency by computing the average of all possible split-half reliabilities for a multiple-item scale. The coefficient demonstrates whether or not the different items converge. Although coefficient α does not address validity, many researchers use α as the sole indicator of a scale's quality. Coefficient alpha ranges in value from 0, meaning no consistency, to 1, meaning complete consistency (all items yield corresponding values). Generally speaking, scales with a coefficient α between 0.80 and 0.95 are considered to have very good reliability. Scales with a coefficient α between 0.70 and 0.80 are considered to have good reliability, and an α value between 0.60 and 0.70 indicates fair reliability. When the coefficient α is below 0.6, the scale has poor reliability. Most statistical software packages, such as SPSS, will easily compute coefficient α .

3- TEST-RETEST RELIABILITY

The **test-retest method** of determining reliability involves administering the same scale or measure to the same respondents at two separate times to test for stability. If the measure is stable over time, the test, administered under the same conditions each time, should obtain similar results. Test- retest reliability represents a measure's repeatability.

Suppose a researcher at one time attempts to measure buying intentions and finds that 12 percent of the population is willing to purchase a product. If the study is repeated a few weeks later under similar conditions, and the researcher again finds that 12 percent of the population is willing to purchase the product, the measure appears to be reliable. High stability correlation or consistency between two measures at time 1 and time 2 indicates high reliability.

Let's assume that a person does not change his or her attitude about dark beer. Attitude might be measured with an item like the one shown below:

I prefer dark beer to all other types of beer.

If repeated measurements of that individual's attitude toward dark beer are taken with the same scale, a reliable instrument will produce the same results each time the scale is measured. Thus one's attitude in October of 2009 should tend to be the same as one's attitude in May 2010. When a measuring instrument produces unpredictable results from one testing to the next, the results are said to be unreliable because of error in measurement.

Validity

Good measures should be both consistent and accurate. Reliability represents how consistent a measure is, in that the different attempts at measuring the same thing converge on the same point. Accuracy deals more with how a measure assesses the intended concept. **Validity** is the accuracy of a measure or the extent to which a score truthfully represents a concept. In other words, are we accurately measuring what we think we are measuring?

Achieving validity is not a simple matter. The opening vignette describes this point. The job performance measure should truly reflect job performance. If a supervisor's friendship affects the performance measure, then the scale's validity is diminished. Likewise, if the performance scale is defined as effort, the result may well be a reliable scale but not one that actually reflects performance. Effort may well lead to performance but effort probably does not equal performance.

Students should be able to empathize with the following validity problem. Consider the controversy about highway patrol officers using radar guns to clock speeders. A driver is clocked at 83 mph in a 55 mph zone, but the same radar gun aimed at a house registers 28 mph. The error occurred because the radar gun had picked up impulses from the electrical system of the squad car's idling engine. Obviously, the house was not moving, thus how can we be sure the car was speeding? In this case, we would certainly question if the accusation that the car was actually going 83 mph is completely valid.

■ ESTABLISHING VALIDITY

Researchers have attempted to assess validity in many ways. They attempt to provide some evidence of a measure's degree of validity by answering a variety of questions. Is there a consensus among other researchers that my attitude scale measures what it is supposed to measure? Does my measure cover everything that it should? Does my measure correlate with other measures of the same concept? Does the behavior expected from my measure predict actual observed behavior? The four basic approaches to establishing validity are *face validity*, *content validity*, *criterion validity*, and *construct validity*.

Face validity refers to the subjective agreement among professionals that a scale logically reflects the concept being measured. Do the test items look like they make sense given a concept's definition? When an inspection of the test items convinces experts that the items match the definition, the scale is said to have face validity.

Clear, understandable questions such as "How many children do you have?" generally are agreed to have face validity. But it becomes more difficult to assess face validity in regard to more complicated business phenomena. For instance, consider the concept of *customer loyalty*. Does the statement "I prefer to purchase my groceries at Delavan Fine Foods" appear to capture loyalty? How about "I am very satisfied with my purchases from Delavan Fine Foods"? What about "Delavan Fine Foods offers very good value"? While the first statement appears to capture loyalty, it can be argued the second question is not loyalty but rather satisfaction. What does the third statement reflect? Do you think it looks like a loyalty statement?

In scientific studies, face validity might be considered a first hurdle. In comparison to other forms of validity, face validity is relatively easy to assess. However, researchers are generally not satisfied with simply establishing face validity. Because of the elusive nature of attitudes and other business phenomena, additional forms of validity are sought.

Content validity refers to the degree that a measure covers the domain of interest. Do the items capture the entire scope, but not go beyond, the concept we are measuring? If an exam is supposed to cover chapters 1—5, it is fair for students to expect that questions should come from all five chapters, rather than just one or two. It is also fair to assume that the questions will not come from chapter 6. Thus, when students complain about the material on an exam, they are often claiming it lacks content validity. Similarly, an evaluation of an employee's job performance should cover all important aspects of the job, but not something outside of the employee's specified duties.

Criterion validity addresses the question, "How well does my measure work in practice?" Because of this, criterion validity is sometimes referred to as *pragmatic validity*. In other words, is my measure practical? Criterion validity may be classified as either *concurrent validity* or *predictive validity* depending on the time sequence in which the new measurement scale and the criterion measure are correlated. If the new measure is taken at the same time as the criterion measure and is shown to be valid, then it has concurrent validity. Predictive validity is established when a new measure predicts a future event. The two measures differ only on the basis of a time dimension—that is, the criterion measure is separated in time from the predictor measure.

For instance, a home pregnancy test is designed to have concurrent validity—to accurately determine if a person is pregnant at the time of the test. Fertility tests, on the other hand, are designed for predictive validity—to determine if a person can become pregnant in the future. In a business setting, participants in a training seminar might be given a test to assess their knowledge of the concepts covered, establishing concurrent validity. Personnel managers may give potential employees an exam to predict if they will be effective salespeople (predictive validity). While face validity is a subjective evaluation, criterion validity provides a more rigorous empirical test.

Construct validity exists when a measure reliably measures and truthfully represents a unique concept. Construct validity consists of several components, including

- Face validity
- Content validity
- Criterion validity
- Convergent validity
- Discriminant validity

We have discussed face validity, content validity, and criterion validity. Before we move further, we must be sure our measures look like they are measuring what they are intended to measure (face validity) and adequately cover the domain of interest (content validity). If so, we can assess **convergent validity** and **discriminant validity**.

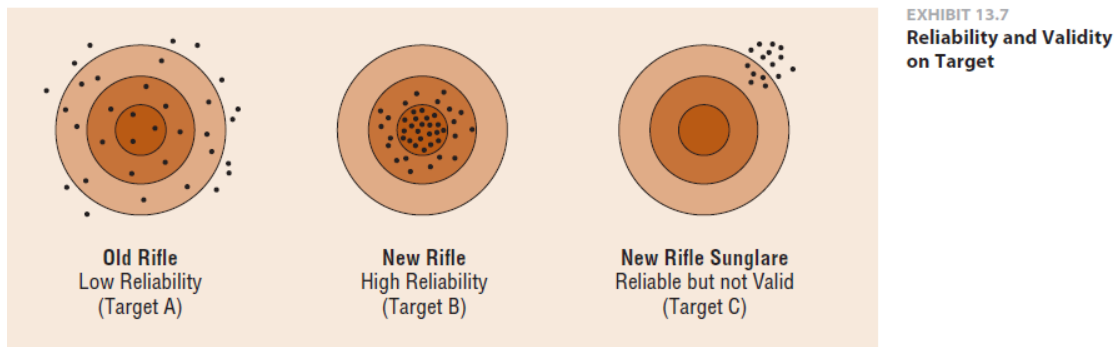
These forms of validity represent how unique or distinct a measure is. Convergent validity requires that concepts that should be related are indeed related.

Reliability versus Validity

Reliability is a necessary but not sufficient condition for validity. A reliable scale may not be valid. For example, a purchase intention measurement technique may consistently indicate that 20 percent of those sampled are willing to purchase a new product. Whether the measure is valid depends on whether 20 percent of the population indeed purchases the product. A reliable but invalid instrument will yield consistently inaccurate results.

The differences between reliability and validity can be illustrated by the rifle targets in Exhibit 13.7. Suppose an expert sharpshooter fires an equal number of rounds with a century-old rifle and a modern rifle.¹⁴ The shots from the older gun are considerably scattered, but those from the newer gun are closely clustered. The variability of the old rifle compared with that of the new one indicates it is less reliable. The target on the right illustrates the concept of a systematic bias influencing validity.

The new rifle is reliable (because it has little variance), but the sharpshooter's vision is hampered by glare. Although shots are consistent, the sharpshooter is unable to hit the bull's-eye.



Sensitivity

The sensitivity of a scale is an important measurement concept, particularly when *changes* in attitudes or other hypothetical constructs are under investigation. **Sensitivity** refers to an instrument's ability to accurately measure variability in a concept. A dichotomous response category, such as "agree or disagree," does not allow the recording of subtle attitude changes. A more sensitive measure with numerous categories on the scale may be needed. For example, adding "strongly agree," "mildly agree," "neither agree nor disagree," "mildly disagree," and "strongly disagree" will increase the scale's sensitivity.

The sensitivity of a scale based on a single question or single item can also be increased by adding questions or items. In other words, because composite measures allow for a greater range of possible scores, they are more sensitive than single-item scales. Thus, sensitivity is generally increased by adding more response points or adding scale items.

ATTITUDE MEASUREMENT

Introduction:

For social scientists, an **attitude** is as an enduring disposition to respond consistently to specific aspects of the world, including actions, people, or objects. One way to understand an attitude is to break it down into its components. Consider this brief statement: "Sally likes shopping at Wal-Mart. She believes the store is *clean*, conveniently *located*, and has low *prices*. She intends to shop there every Thursday." This simple example demonstrates attitude's three components: affective, cognitive, and behavioral. The affective component refers to an individual's general feelings or emotions toward an object. Statements such as "I really like my Corvette," "I enjoy reading new Harry Potter books," and "I hate cranberry juice" reflect an emotional character of attitudes. A person's attitudinal feelings are driven directly by his/her *beliefs* or *cognitions*. This cognitive component represents an individual's knowledge about attributes and their consequences. One person might feel happy about the purchase of an automobile because she believes the car "gets great gas mileage" or knows that the dealer is "the best in New Jersey." The behavioral component of an attitude reflects a predisposition to action by reflecting an individual's intentions.

Attitudes as Hypothetical Constructs

Business researchers often pose questions involving psychological variables that cannot directly be observed. For example, someone may have an attitude toward working on a commission basis. We cannot actually see this attitude. Rather, we can measure an attitude by making an inference

based on the way an individual responds to multiple scale indicators. Because we can't directly see these phenomena, they are known as latent constructs, **hypothetical constructs**, or just simply constructs. Common constructs include job satisfaction, organizational commitment, personal values, feelings, role stress, perceived value, and many more. The Research Snapshot on page 317 talks about measuring love. Is love a hypothetical construct?

Importance of Measuring Attitudes

Most managers hold the intuitive belief that changing consumers' or employees' attitudes toward their company or their company's products or services is a major goal. Because modifying attitudes plays a pervasive role in developing strategies to address these goals, the measurement of attitudes is an important task. For example, after Whiskas cat food had been sold in Europe for decades, the brand faced increased competition from new premium brands, and consumers had difficulty identifying with the brand. The company conducted attitude research to determine how people felt about their cats and their food alternatives. The study revealed that cat owners see their pets both as independent and as dependent fragile beings.¹ Cat owners held the attitude that cats wanted to enjoy their food but needed nutrition. This attitude research was directly channeled into managerial action. Whiskas marketers begin positioning the product as having "Catisfaction," using advertisements that featured a purring silver tabby—a pedigreed cat—which symbolizes premium quality but also presents the image of a sweet cat. The message: "Give cats what they like with the nutrition they need. If you do, they'll be so happy that they'll purr for you." This effort reversed the sales decline the brand had been experiencing.

Techniques for Measuring Attitudes

A remarkable variety of techniques has been devised to measure attitudes. This variety stems in part from lack of consensus about the exact definition of the concept. In addition, the affective, cognitive, and behavioral components of an attitude may be measured by different means. For example, sympathetic nervous system responses may be recorded using physiological measures to quantify affect, but they are not good measures of behavioral intentions. Direct verbal statements concerning affect, belief, or behavior are used to measure behavioral intent. However, attitudes may also be interpreted using qualitative techniques like those discussed in Chapter 7.

Research may assess the affective (emotional) components of attitudes through physiological measures such as galvanic skin response (GSR), blood pressure, and pupil dilation. These measures provide a means of assessing attitudes without verbally questioning the respondent. In general, they can provide a gross measure of likes or dislikes, but they are not extremely sensitive to the different gradients of an attitude.

Obtaining verbal statements from respondents generally requires that the respondents perform a task such as ranking, rating, sorting, or making choices. A **ranking** task requires the respondent to rank order a small number of stores, brands, feelings, or objects on the basis of overall preference or some characteristic of the stimulus. **Rating** asks the respondent to estimate the magnitude or the extent to which some characteristic exists. A quantitative score results. The rating task involves marking a response indicating one's position using one or more attitudinal or cognitive scales. A **sorting** task might present the respondent with several different concepts printed on cards and require the respondent to classify the concepts by placing the cards into groups (stacks of cards).

Another type of attitude measurement is **choice** between two or more alternatives. If a respondent chooses one object over another, the researcher assumes that the respondent prefers the chosen object, at least in this setting. The following sections describe the most popular techniques for measuring attitudes.

Attitude Rating Scales

Perhaps the most common practice in business research is using rating scales to measure attitudes. This section discusses many rating scales designed to enable respondents to report the intensity of their attitudes.

Simple Attitude Scales

In its most basic form, attitude scaling requires that an individual agree or disagree with a statement or respond to a single question. For example, respondents in a political poll may be asked whether they agree or disagree with the statement “The president should run for re-election.” Or, an individual might indicate whether he or she likes or dislikes jalapeno bean dip. This type of self-rating scale merely classifies respondents into one of two categories, thus having only the properties of a nominal scale, and the types of mathematical analysis that may be used with this basic scale are limited.

Despite the disadvantages, simple attitude scaling may be used when questionnaires are extremely long, when respondents have little education, or for other specific reasons. A number of simplified scales are merely checklists: A respondent indicates past experience, preference, and the like merely by checking an item. In many cases the items are adjectives that describe a particular object. In a survey of small-business owners and managers, respondents indicated whether they found working in a small firm more rewarding than working in a large firm, as well as whether they agreed with a series of attitude statements about small businesses. For example, 77 percent said small and mid-sized businesses “have less bureaucracy,” and 76 percent said smaller companies “have more flexibility” than large ones.²

Most attitude theorists believe that attitudes vary along continua. Early attitude researchers pioneered the view that the task of attitude scaling is to measure the distance from “good” to “bad,” “low” to “high,” “like” to “dislike,” and so on. Thus, the purpose of an attitude scale is to find an individual’s position on the continuum. However, simple scales do not allow for fine distinctions between attitudes. Several other scales have been developed for making more precise measurements.

Category Scales

The simplest rating scale contains only two response categories: agree/disagree. Expanding the response categories provides the respondent with more flexibility in the rating task. Even more information is provided if the categories are ordered according to a particular descriptive or evaluative dimension. Consider the following question:

How often do you disagree with your spouse about how much to spend on vacation?				
Never	Rarely	Sometimes	Often	Very often
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

This **category scale** is a more sensitive measure than a scale that has only two response categories. By having more choices for a respondent, the potential exists to provide more information. However, if the researcher tries to represent something that is truly bipolar (yes/no, female/male, member/nonmember, and so on) with more than two categories, error may be introduced. Question wording is an extremely important factor in the usefulness of these scales.

Method of Summated Ratings: The Likert Scale

A method that is simple to administer and therefore extremely popular is business researchers’ adaptation of the method of summated ratings, developed by Rensis Likert.³ With the **Likert scale**, respondents indicate their attitudes by checking how strongly they agree or disagree with carefully constructed statements, ranging from very positive to very negative attitudes toward some object. Individuals generally choose from approximately five response alternatives—strongly agree, agree,

uncertain, disagree, and strongly disagree—although the number of alternatives may range from three to nine. In the following example, from a study of food-shopping behavior, there are five alternatives.

In buying food for my family, price is no object.				
Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
(1)	(2)	(3)	(4)	(5)

Researchers assign scores, or weights, to each possible response. In this example, numerical scores of 1, 2, 3, 4, and 5 are assigned to each level of agreement, respectively. The numerical scores, shown in parentheses, may not be printed on the questionnaire or computer screen. Strong agreement indicates the most favorable attitude on the statement, and a numerical score of 5 is assigned to this response.

➤ REVERSE RECODING

The statement given in this example is positively framed. If a statement is framed negatively (such as “I carefully budget my food expenditures”), the numerical scores would need to be reversed. This is done by **reverse recoding** the negative item so that a strong agreement really indicates an unfavorable response rather than a favorable attitude. In the case of a five-point scale, the recoding is done as follows

Old Value	New Value
1	5
2	4
3	3
4	2
5	1

Recoding in this fashion turns agreement with a negatively worded item into a mirror image, meaning the result is the same as disagreement with a positively worded item. SPSS has a recode function that allows simple recoding to be done by entering “old” and “new” scale values. Alternatively, a simple mathematical formula can be entered. For a typical 1—5 scale, the formula

$$X_{\text{new_value}} = 6 - X_{\text{old_value}}$$

would result in the same recoding.

➤ COMPOSITE SCALES

A Likert scale may include several scale items to form a **composite scale**. Each statement is assumed to represent an aspect of a common attitudinal domain. For example, below figure shows the items in a Likert scale for measuring attitudes toward patients’ interaction with a physician’s service staff. The total score is the summation of the numerical scores assigned to an individual’s responses. Here the maximum possible score for the composite would be 20 if a 5 were assigned to “strongly agree” responses for each of the positively worded statements and a 5 to “strongly disagree” responses for the negative statement. Item 3 is negatively worded and therefore it is reverse coded, prior to being used to create the composite scale.

1. My doctor’s office staff takes a warm and personal interest in me.
2. My doctor’s office staff is friendly and courteous.
3. My doctor’s office staff is more interested in serving the doctor’s needs than in serving my needs.
4. My doctor’s office staff always acts in a professional manner.

In Likert's original procedure, a large number of statements are generated, and an *item analysis* is performed. The purpose of the item analysis is to ensure that final items evoke a wide response and discriminate among those with positive and negative attitudes. Items that are poor because they lack clarity or elicit mixed response patterns are eliminated from the final statement list. Scales that use multiple items can be analyzed for reliability and validity. Only a set of items that demonstrates good reliability and validity should be summed or averaged to form a composite scale representing a hypothetical construct. Unfortunately, not all researchers are willing or able to thoroughly assess reliability and validity. Without this test, the use of Likert scales can be disadvantageous because there is no way of knowing exactly what the items represent or how well they represent anything of interest. Without valid and reliable measures, researchers cannot guarantee they are measuring what they say they are measuring.

Semantic Differential

The **semantic differential** is actually a series of attitude scales. This popular attitude measurement technique consists of getting respondents to react to some concept using a series of seven-point bipolar rating scales. Bipolar adjectives—such as “good” and “bad,” “modern” and “old-fashioned,” or “clean” and “dirty”—anchor the beginning and the end (or poles) of the scale. The subject makes repeated judgments about the concept under investigation on each of the scales. Exhibit 14.3 shows seven of eighteen scales used in a research project that measured attitudes toward supermarkets.

The scoring of the semantic differential can be illustrated using the scale bounded by the anchors “modern” and “old-fashioned.” Respondents are instructed to check the place that indicates the nearest appropriate adjective. From left to right, the scale intervals are interpreted as “extremely modern,” “very modern,” “slightly modern,” “both modern and old-fashioned,” “slightly old-fashioned,” “very old-fashioned,” and “extremely old-fashioned”:

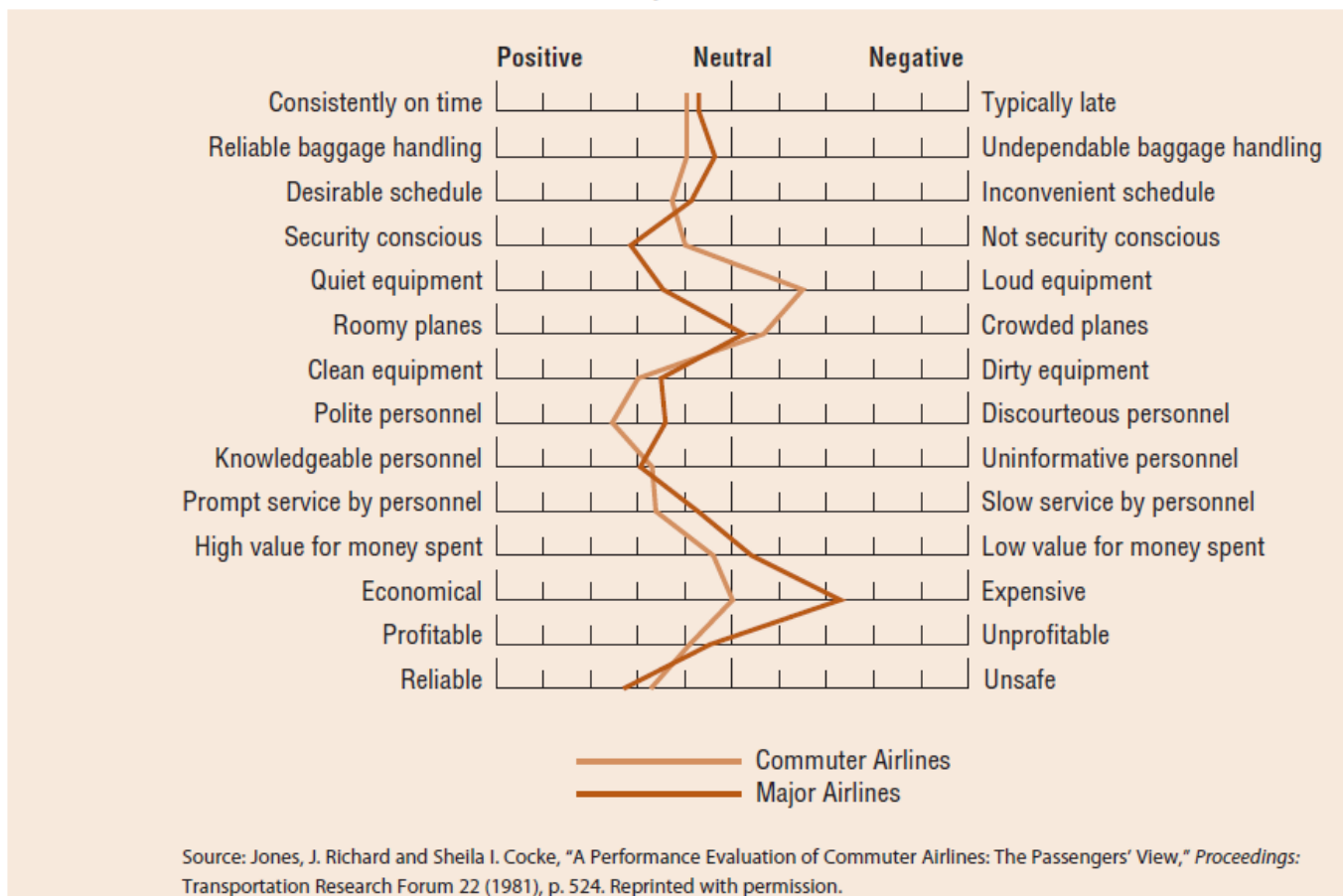
Modern _____ Old-fashioned

The semantic differential technique originally was developed as a method for measuring the meanings of objects or the “semantic space” of interpersonal experience. Researchers have found the semantic differential versatile and useful in business applications. The validity of the semantic differential depends on finding scale anchors that are semantic opposites. This can sometimes prove difficult. However, in attitude or image studies simple anchors such as very unfavorable and very favorable work well.

For scoring purposes, a numerical score is assigned to each position on the rating scale. Traditionally, score ranges such as 1, 2, 3, 4, 5, 6, 7 or —3, —2, —1, 0, +1, +2, +3 are used. Many business researchers find it desirable to assume that the semantic differential provides interval data. This assumption, although widely accepted, has its critics, who argue that the data have only ordinal properties because the numerical scores are arbitrary. Practically speaking, most researchers will treat semantic differential scales as metric (at least interval). This is because the amount of error introduced by assuming the intervals between choices are equal (even though this is uncertain) is fairly small.

Below figure illustrates a typical **image profile** based on semantic differential data. Because the data are assumed to be interval, either the arithmetic mean or the median will be used to compare the profile of one product, brand, or store with that of a competing product, brand, or store.

EXHIBIT 14.4 Image Profile of Commuter Airlines versus Major Airlines



Numerical Scales

A **numerical scale** simply provides numbers rather than a semantic space or verbal descriptions to identify response options or categories (response positions). For example, a scale using five response positions is called a five-point numerical scale. A six-point scale has six positions and a seven-point scale seven positions, and so on. Consider the following numerical scale:

Now that you've had your automobile for about one year, please tell us how satisfied you are with your Ford Taurus.
Extremely Dissatisfied 1 2 3 4 5 6 7 Extremely Satisfied

This numerical scale uses bipolar adjectives in the same manner as the semantic differential.

In practice, researchers have found that a scale with numerical labels for intermediate points on the scale is as effective a measure as the true semantic differential. The Research Snapshot above demonstrates how numerical scales can be helpful in assessing Web site effectiveness.

Stapel Scale

The **Stapel scale**, named after Jan Stapel, was originally developed in the 1950s to measure simultaneously the direction and intensity of an attitude. Modern versions of the scale, with a single adjective, are used as a substitute for the semantic differential when it is difficult to create pairs of bipolar adjectives. The modified Stapel scale places a single adjective in the center of an even number of numerical values (ranging, perhaps, from +3 to -3). The scale measures how close to or distant from the adjective a given stimulus is perceived to be. Exhibit 14.5 illustrates a Stapel scale item used in measurement of a retailer's store image.

The advantages and disadvantages of the Stapel scale are very similar to those of the semantic differential. However, the Stapel scale is markedly easier to administer, especially over the telephone.

Because the Stapel scale does not require bipolar adjectives, it is easier to construct than the semantic differential. Research comparing the semantic differential with the Stapel scale indicates that results from the two techniques are largely the same.

EXHIBIT 14.5
A Stapel Scale for Measuring
a Store's Image

	Bloomingdale's
	13
	12
	11
Wide Selection	
	21
	22
	23

Select a *plus* number for words that you think describe the store accurately. The more accurately you think the word describes the store, the larger the plus number you should choose. Select a *minus* number for words you think do not describe the store accurately. The less accurately you think the word describes the store, the larger the minus number you should choose. Therefore, you can select any number from 13 for words that you think are very accurate all the way to 23 for words that you think are very inaccurate.

Constant-Sum Scale

A **constant-sum scale** requires respondents to divide a fixed number of points among several attributes corresponding to their relative importance or weight. Suppose United Parcel Service (UPS) wishes to determine the importance of the attributes of accurate invoicing, delivery as promised, and price to organizations that use its service in business-to-business settings. Respondents might be asked to divide a constant sum of 100 points to indicate the relative importance of those attributes:

Divide 100 points among the following characteristics of a delivery service according to how important each characteristic is to you when selecting a delivery company.

- ___ Accurate invoicing
- ___ Package not damaged
- ___ Delivery as promised
- ___ Lower price
- ___ 100 points

The constant-sum scale works best with respondents who have high educational levels. If respondents follow the instructions correctly, the results will approximate interval measures. As the number of stimuli increases, this technique becomes increasingly complex.

This technique may be used for measuring brand preference. The approach, which is similar to the paired-comparison method, is as follows:

Divide 100 points among the following brands according to your preference for each brand:

- ___ Brand A
- ___ Brand B
- ___ Brand C
- ___ 100 points

In this case, the constant-sum scale is a rating technique. However, with minor modifications, it can be classified as a sorting technique. Although the constant-sum scale is widely used, strictly speaking, the scale is flawed because the last response is completely determined by the way the respondent has scored the other choices. Although this is probably somewhat complex to understand, the fact is that practical reasons often outweigh this concern.

Graphic Rating Scales

A **graphic rating scale** presents respondents with a graphic continuum. The respondents are allowed to choose any point on the continuum to indicate their attitude. Exhibit 14.6 on the next page shows a traditional graphic scale, ranging from one extreme position to the opposite position. Typically a respondent's score is determined by measuring the length (in millimeters) from one end of the graphic

continuum to the point marked by the respondent. Many researchers believe that scoring in this manner strengthens the assumption that graphic rating scales of this type are interval scales. Alternatively, the researcher may divide the line into predetermined scoring categories (lengths) and record respondents' marks accordingly. In other words, the graphic rating scale has the advantage of allowing the researcher to choose any interval desired for scoring purposes. The disadvantage of the graphic rating scale is that there are no standard answers.

EXHIBIT 14.6

Graphic Rating Scale

Please evaluate each attribute in terms of how important it is to you by placing an X at the position on the horizontal line that most reflects your feelings.

Seating comfort Not important _____ Very important

In-flight meals Not important _____ Very important

Airfare Not important _____ Very important

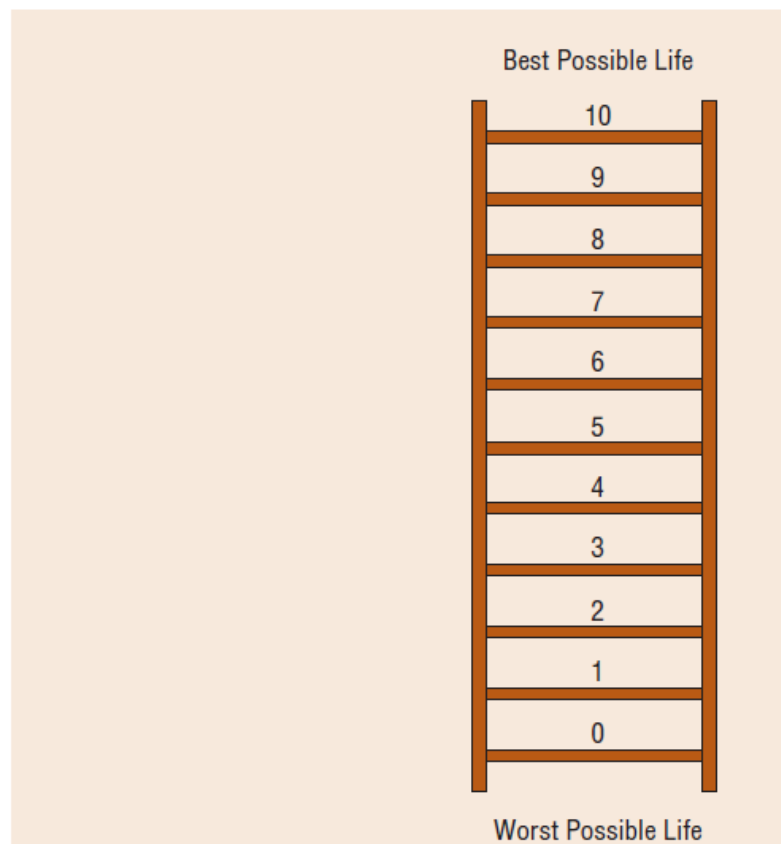
Graphic rating scales are not limited to straight lines as sources of visual communication. Picture response options or another type of graphic continuum may be used to enhance communication with respondents. A variation of the graphic ratings scale is the ladder scale. This scale also includes numerical options:

Here is a ladder scale (response scale is shown in Exhibit 14.7). It represents the "ladder of life." As you see, it is a ladder with eleven rungs numbered 0 to 10. Let's suppose the top of the ladder represents the best possible life for you as you describe it, and the bottom rung represents the worst possible life for you as you describe it.

On which rung of the ladder do you feel your life is today?

0 1 2 3 4 5 6 7 8 9 10

EXHIBIT 14.7

A Ladder Scale

Research to investigate children's attitudes has used happy-face scales (see Exhibit 14.8). The children are asked to indicate which face shows how they feel about candy, a toy, or some other concept. Research with the happy-face scale indicates that children tend to choose the faces at the ends of the scale. Although this may be because children's attitudes fluctuate more widely than adults' or because

they have stronger feelings both positively and negatively, the tendency to select the extremes is a disadvantage of the scale.

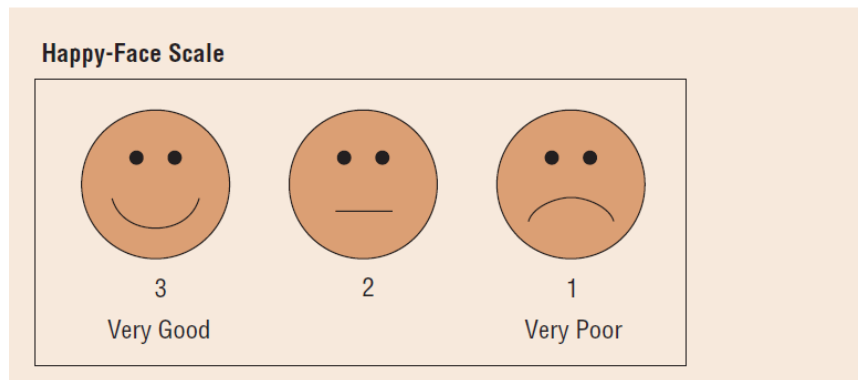


EXHIBIT 14.8

**Graphic Rating Scale
with Picture Response
Categories Stressing Visual
Communication**

Measuring Behavioral Intention

The behavioral component of an attitude involves the behavioral expectations of an individual toward an attitudinal object. The component of interest to researchers may be turnover intentions, a tendency to make business decisions in a certain way, or plans to expand operations or product offerings. For example, category scales for measuring the behavioral component of an attitude ask about a respondent's likelihood of purchase or intention to perform some future action, using questions such as the following:

How likely is it that you will purchase an MP3 player such as an iPod?

- I definitely will buy.
- I probably will buy.
- I might buy.
- I probably will not buy.
- I definitely will not buy.

How likely am I to write a letter to my representative in Congress or other government official in support of this company if it were in a dispute with government?

- Extremely Likely
- Very Likely
- Somewhat Likely
- Likely, about a 50—50 chance
- Somewhat Unlikely
- Very Unlikely
- Absolutely Unlikely

The wording of statements used in these scales often includes phrases such as “I would recommend,” “I would write,” or “I would buy” to indicate action tendencies.

Expectations also may be measured using a scale of subjective probabilities, ranging from 100 for “absolutely certain” to 0 for “absolutely no chance.” Researchers have used the following

subjective probability scale to estimate the chance that a job candidate will accept a position within a company:

100%	(Absolutely certain) I will accept
90%	(Almost sure) I will accept
80%	(Very big chance) I will accept
70%	(Big chance) I will accept
60%	(Not so big a chance) I will accept
50%	(About even) I will accept
40%	(Smaller chance) I will accept
30%	(Small chance) I will accept

20%	(Very small chance) I will accept
10%	(Almost certainly not) I will accept
0%	(Certainly not) I will accept

Behavioral Differential

A general instrument, the **behavioral differential**, is used to measure the behavioral intentions of subjects toward an object or category of objects. As in the semantic differential, a description of the object to be judged is followed by a series of scales on which subjects indicate their behavioral intentions toward this object. For example, one item might be something like this:

A 25-year-old female sales representative
 Would _____ Would not
 ask this person for advice.

Ranking

Consumers often *rank order* their preferences. An ordinal scale may be developed by asking respondents to rank order (from most preferred to least preferred) a set of objects or attributes. Respondents easily understand the task of rank ordering the importance of product attributes or arranging a set of brand names according to preference. Like the constant-sum scale, technically the ranking scale also suffers from inflexibility in that if we know how some ranked five out of six alternatives, we know the answer to the sixth.

Paired Comparisons

Consider a situation in which a chainsaw manufacturer learned that a competitor had introduced a new lightweight (6-pound) chainsaw. The manufacturer's lightest chainsaw weighed 9 pounds. Executives wondered if they needed to introduce a 6-pound chainsaw into the product line. The research design chosen was a **paired comparison**. A 6-pound chainsaw was designed, and a prototype built. To control for color preferences, the competitor's chainsaw was painted the same color as the 9- and 6-pound chainsaws. Respondents were presented with two chainsaws at a time and asked to pick the one they preferred. Three pairs of comparisons were required to determine the most preferred chainsaw.

The following question illustrates the typical format for asking about paired comparisons.

I would like to know your overall opinion of two brands of adhesive bandages. They are Curad and Band-Aid. Overall, which of these two brands—Curad or Band-Aid—do you think is the better one? Or are both the same?

Curad is better.

Band-Aid is better.

They are the same.

If researchers wish to compare four brands of pens on the basis of attractiveness or writing quality, six comparisons $[(n)(n - 1)/2]$ will be necessary.

When comparing only a few items, ranking objects with respect to one attribute is not difficult. As the number of items increases, the number of comparisons increases geometrically. If the number of comparisons is too large, respondents may become fatigued and no longer carefully discriminate among them.

Sorting

Sorting tasks ask respondents to indicate their attitudes or beliefs by arranging items on the basis of perceived similarity or some other attribute. One advertising agency has had consumers sort photographs of people to measure their perceptions of a brand's typical user. Another agency used a sorting technique in which consumers used a deck of 52 cards illustrating elements from advertising for the brand name being studied. The study participants created a stack of cards showing elements

they recalled seeing or hearing, and the interviewer then asked the respondent to identify the item on each of those cards. National City Corporation, a banking company, has used sorting as part of its research into the design of its Web site. Consumers participating in the research were given a set of cards describing various parts of processes that they might engage in when they are banking online. The participants were asked to arrange the cards to show their idea of a logical way to complete these processes. This research method shows the Web site designers how consumers go about doing something—sometimes very differently from the way bankers expect.⁶

A variant of the constant-sum technique uses physical counters (for example, poker chips or coins), to be divided among the items being tested. In an airline study of customer preferences, the following sorting technique could be used:

Here is a sheet that lists several airlines. Next to the name of each airline is a pocket. Here are ten cards. I would like you to put these cards in the pockets next to the airlines you would prefer to fly on your next trip. Assume that all of the airlines fly to wherever you would choose to travel. You can put as many cards as you want next to an airline, or you can put no cards next to an airline.

Cards

American Airlines _____
Delta Airlines _____
United Airlines _____
Southwest Airlines _____
Northwest Airlines _____

Other Methods of Attitude Measurement

Attitudes, as hypothetical constructs, cannot be observed directly. We can, however, infer one's attitude by the way he or she responds to multiple attitude indicators. A summated rating scale can be made up of three indicators of attitude. Consider the following three semantic differential items that may capture a person's attitude towards their immediate supervisor:

very good _____ *very bad*
very unfavorable _____ *very favorable*
very positive _____ *very negative*

The terminology is such that now attitude would be represented as a latent (unobservable) construct indicated by the person's response to these items.

Selecting a Measurement Scale: Some Practical Decisions

Now that we have looked at a number of attitude measurement scales, a natural question arises: "Which is most appropriate?" As in the selection of a basic research design, there is no single best answer for all research projects. The answer to this question is relative, and the choice of scale will depend on the nature of the attitudinal object to be measured, the manager's problem definition, and the backward and forward linkages to choices already made (for example, telephone survey versus mail survey). However, several questions will help focus the choice of a measurement scale:

1. Is a ranking, sorting, rating, or choice technique best?
2. Should a monadic or a comparative scale be used?
3. What type of category labels, if any, will be used for the rating scale?
4. How many scale categories or response positions are needed to accurately measure an attitude?
5. Should a balanced or unbalanced rating scale be chosen?

6. Should a scale that forces a choice among predetermined options be used?
 7. Should a single measure or an index measure be used?

Ranking, Sorting, Rating, or Choice Technique?

The decision whether to use ranking, sorting, rating, or a choice technique is determined largely by the problem definition and especially by the type of statistical analysis desired. For example, ranking provides only ordinal data, limiting the statistical techniques that may be used.

Monadic or Comparative Scale?

If the scale to be used is not a ratio scale, the researcher must decide whether to include a standard of comparison in the verbal portion of the scale. Consider the following rating scale:

Now that you've had your automobile for about one year, please tell us how satisfied you are with its engine power and pickup.

Completely Dissatisfied	Dissatisfied	Somewhat Satisfied	Satisfied	Completely Satisfied
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

This is a **monadic rating scale**, because it asks about a single concept (the brand of automobile the individual actually purchased) in isolation. The respondent is not given a specific frame of reference. A **comparative rating scale** asks a respondent to rate a concept, such as a specific amount of responsibility or authority, in comparison with a benchmark—perhaps another similar concept—explicitly used as a frame of reference. In many cases, the comparative rating scale presents an ideal situation as a reference point for comparison with the actual situation. For example:

Please indicate how the amount of authority in your present position compares with the amount of authority that would be ideal for this position.

Too much ☐ *About right* ☐ *Too little* ☐

What Type of Category Labels, If Any?

We have discussed verbal labels, numerical labels, and unlisted choices. Many rating scales have verbal labels for response categories because researchers believe they help respondents better understand the response positions. The maturity and educational levels of the respondents will influence this decision. The semantic differential, with unlabeled response categories between two bipolar adjectives, and the numerical scale, with numbers to indicate scale positions, often are selected because the researcher wishes to assume interval-scale data.

How Many Scale Categories or Response Positions?

Should a category scale have four, five, or seven response positions or categories? Or should the researcher use a graphic scale with an infinite number of positions? The original developmental research on the semantic differential indicated that five to eight points is optimal. However, the researcher must determine the number of meaningful positions that is best for the specific project. This issue of identifying how many meaningful distinctions respondents can practically make is basically a matter of sensitivity, but at the operational rather than the conceptual level.

Balanced or Unbalanced Rating Scale?

The fixed-alternative format may be balanced or unbalanced. For example, the following question, which asks about parent-child decisions relating to television program watching, is a **balanced rating scale**:

Who decides which television programs your children watch?	
Child decides all of the time.	<input type="checkbox"/>
Child decides most of the time.	<input type="checkbox"/>
Child and parent decide together.	<input type="checkbox"/>
Parent decides most of the time.	<input type="checkbox"/>
Parent decides all of the time.	<input type="checkbox"/>

This scale is balanced because a neutral point, or point of indifference, is at the center of the scale.

Unbalanced rating scales may be used when responses are expected to be distributed at one end of the scale. Unbalanced scales, such as the following one, may eliminate this type of “end piling”:

Completely Dissatisfied	Dissatisfied	Somewhat Satisfied	Satisfied	Completely Satisfied
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Notice that there are three “satisfied” responses and only two “dissatisfied” responses above. The choice of a balanced or unbalanced scale generally depends on the nature of the concept or the researcher’s knowledge about attitudes toward the stimulus to be measured.

Use a Scale That Forces a Choice among Predetermined Options?

In many situations, a respondent has not formed an attitude toward the concept being studied and simply cannot provide an answer. If a **forced-choice rating scale** compels the respondent to answer, the response is merely a function of the question. If answers are not forced, the midpoint of the scale may be used by the respondent to indicate unawareness as well as indifference. If many respondents in the sample are expected to be unaware of the attitudinal object under investigation, this problem may be eliminated by using a non-forced-choice scale that provides a “no opinion” category, as in the following example:

How does the Bank of Commerce compare with the First National Bank?

- ☐ *Bank of Commerce is better than First National Bank.*
- ☐ *Bank of Commerce is about the same as First National Bank.*
- ☐ *Bank of Commerce is worse than First National Bank.*
- ☐ *Can't say.*

Asking this type of question allows the investigator to separate respondents who cannot make an honest comparison from respondents who have had experience with both banks. The argument for forced choice is that people really do have attitudes, even if they are unfamiliar with the banks, and should be required to answer the question. Still, the use of forced-choice questions is associated with

higher incidences of “no answer.” Internet surveys make forced-choice questions easy to implement because the delivery can be set up so that a respondent cannot go to the next question until the previous question is answered. Realize, however, if a respondent truly has no opinion, and the no opinion option is not included, he or she may simply quit responding to the questionnaire.

Single Measure or an Index Measure?

Whether to use a single measure or an index measure depends on the complexity of the issue to be investigated, the number of dimensions the issue contains, and whether individual attributes of the stimulus are part of a holistic attitude or are seen as separate items. Very simple concepts that do not vary from context to context can be measured by single items. However, most psychological concepts are more complex and require multiple-item measurement. Additionally, multiple-item measures are easier to test for construct validity (discussed in later chapters). The researcher’s conceptual definition will be helpful in making this choice.

The researcher has many scaling options. Generally, the choice is influenced by plans for the later stages of the research project. Again, problem definition becomes a determining factor influencing the research design.

QUESTIONNAIRE DESIGN

Introduction

Each stage in the business research process is important and interdependent. The research questionnaire development stage is critically important as the information provided is only as good as the questions asked. However, the importance of question wording is easily, and far too often, overlooked.

Businesspeople who are inexperienced at research frequently believe that constructing a questionnaire is a simple task. Amateur researchers think a short questionnaire can be written in minutes. Unfortunately, newcomers who naively believe that good grammar is all a person needs to construct a questionnaire generally end up with useless results. Ask a bad question, get bad results.

Good questionnaire design requires far more than correct grammar. People don’t understand questions just because they are grammatically correct. Respondents simply may not know what is being asked. They may be unaware of the business issue or topic of interest. They may confuse the subject with something else. The question may not mean the same thing to everyone interviewed. Finally, people may refuse to answer personal questions. Most of these problems can be minimized, however, if a skilled researcher composes the questionnaire.

Questionnaire Quality and Design: Basic Considerations

For a questionnaire to fulfill a researcher’s purposes, the questions must meet the basic criteria of *relevance* and *accuracy*. To achieve these ends, a researcher who is systematically planning a questionnaire’s design will be required to make several decisions—typically, but not necessarily, in the following order:

1. What should be asked?
2. How should questions be phrased?
3. In what sequence should the questions be arranged?
4. What questionnaire layout will best serve the research objectives?
5. How should the questionnaire be pretested? Does the questionnaire need to be revised?

What Should Be Asked?

Certain decisions made during the early stages of the research process will influence the questionnaire design. The preceding chapters stressed good problem definition and clear research questions. This leads to specific research hypotheses that, in turn, clearly indicate what must be measured. Different types of questions may be better at measuring certain things than are others. In addition, the communication medium used for data collection—that is, telephone interview, personal interview, or self-administered questionnaire—must be determined. This decision is another forward linkage that influences the structure and content of the questionnaire. Therefore, the specific questions to be asked will be a function of previous decisions made in the research process. At the same time, the latter stages of the research process will also have an important impact on questionnaire wording and measurement. For example, when designing the questionnaire, the researcher should consider the types of statistical analysis that will be conducted.

Questionnaire Relevancy

A questionnaire is *relevant* to the extent that all information collected addresses a research question that will help the decision maker address the current business problem. Asking a wrong question or an irrelevant question is a common pitfall. If the task is to pinpoint store image problems, questions asking for political opinions are likely irrelevant. The researcher should be specific about data needs and have a rationale for each item requesting information. Irrelevant questions are more than a nuisance because they make the survey needlessly long. In a study where two samples of the same group of businesses received either a one-page or a three-page questionnaire, the response rate was nearly twice as high for the one-page survey.

Conversely, many researchers, after conducting surveys, find that they omitted some important questions. Therefore, when planning the questionnaire design, researchers must think about possible omissions. Is information on the relevant demographic and psychographic variables being collected? Would certain questions help clarify the answers to other questions? Will the results of the study provide the answer to the manager's problem?

Questionnaire Accuracy

Once a researcher decides what should be asked, the criterion of accuracy becomes the primary concern. *Accuracy* means that the information is reliable and valid. While experienced researchers generally believe that questionnaires should use simple, understandable, unbiased, unambiguous, and nonirritating words, no step-by-step procedure for ensuring accuracy in question writing can be generalized across projects. Obtaining accurate answers from respondents depends strongly on the researcher's ability to design a questionnaire that will facilitate recall and motivate respondents to cooperate. Respondents tend to be more cooperative when the subject of the research interests them. When questions are not lengthy, difficult to answer, or ego threatening, there is a higher probability of obtaining unbiased answers.

Question wording and sequence also substantially influence accuracy, which can be particularly challenging when designing a survey for technical audiences. The Department of Treasury commissioned a survey of insurance companies to evaluate their offering of terrorism insurance as required by the government's terrorism reinsurance program. But industry members complained that the survey misused terms such as "contract" and "high risk," which have precise meanings for insurers, and asked for policy information "to date," without specifying which date. These questions caused confusion and left room for interpretation, calling the survey results into question.

Wording Questions

There are many ways to phrase questions, and many standard question formats have been developed in previous research studies. This section presents a classification of question types and provides some helpful guidelines for writing questions.

Open-Ended Response versus Fixed-Alternative Questions

The first decision in questionnaire design is based on the amount of freedom respondents have in answering. Should the question be open-ended, allowing the participants freedom to choose their manner of response, or closed, where the participants choose their response from an already determined fixed set of choices?

Open-ended response questions pose some problem or topic and ask respondents to answer in their own words. If the question is asked in a personal interview, the interviewer may probe for more information, as in the following examples:

- *What names of local banks can you think of?*
- *What comes to mind when you look at this advertisement?*
- *In what way, if any, could this product be changed or improved? I'd like you to tell me anything you can think of no matter how minor it seems.*
- *What things do you like most about working for Federal Express? What do you like least?*
- *Why do you buy more of your clothing in Nordstrom than in other stores?*
- *How would you describe your supervisor's management style?*
- *Please tell us how our stores can better serve your needs.*

Open-ended response questions are free-answer questions. They may be contrasted with **fixed-alternative questions**—sometimes called *closed-ended questions*—which give respondents specific limited-alternative responses and ask them to choose the one closest to their own viewpoints. For example:

Did you use any commercial feed or supplement for livestock or poultry in 2010?

- ☐ Yes
- ☐ No

Would you say that the labor quality in Japan is higher, about the same, or not as good as it was 10 years ago?

- ☐ Higher
- ☐ About the same
- ☐ Not as good

Do you think the Renewable Energy Partnership Program has affected your business?

- ☐ Yes, for the better
- ☐ Not especially
- ☐ Yes, for the worse

How much of your welding supplies do you purchase from our *Tier One* suppliers?

- ☐ All of it
- ☐ Most of it
- ☐ About one-half of it
- ☐ About one-quarter of it
- ☐ Less than one-quarter of it

■ USING OPEN-ENDED RESPONSE QUESTIONS

Open-ended response questions are most beneficial when the researcher is conducting exploratory research, especially when the range of responses is not yet known. Respondents are free to answer with whatever is foremost in their minds. Such questions can be used to learn which words and phrases people spontaneously give to the free-response question. Such responses will reflect the flavor of the language that people use in talking about the issue and thus may provide guidance in the wording of questions and responses for follow up surveys.

Also, open-ended response questions are valuable at the beginning of an interview. They are good first questions because they allow respondents to warm up to the questioning process. They are also good last questions for a fixed-alternative questionnaire, when a researcher can ask the respondent to expand in a manner that provides greater richness to the data. For example, an employee satisfaction survey may collect data with a series of fixed-alternative questions, then conclude with “Can you provide one suggestion on how our organization can enhance employee satisfaction?”

The cost of administering open-ended response questions is substantially higher than that of administering fixed-alternative questions because the job of editing, coding, and analyzing the data is quite extensive. As each respondent’s answer is somewhat unique, there is some difficulty in categorizing and summarizing the answers. The process requires that an editor go over a sample of questions to develop a classification scheme. This scheme is then used to code all answers according to the classification scheme.

Another potential disadvantage of the open-ended response question is the possibility that interviewer bias will influence the answer. While most interviewer instructions state that answers are to be recorded verbatim, rarely does even the best interviewer get every word spoken by the respondent. Interviewers have a tendency to take shortcuts. When this occurs, the interviewer may well introduce error because the final answer may reflect a combination of the respondent’s and interviewer’s ideas.

■ USING FIXED-ALTERNATIVE QUESTIONS

In contrast, fixed-alternative questions require less interviewer skill, take less time, and are easier for the respondent to answer. This is because answers to closed questions are classified into standardized groupings prior to data collection. Standardizing alternative responses to a question provides comparability of answers, which facilitates coding, tabulating, and ultimately interpreting the data.

However, when a researcher is unaware of the potential responses to a question, fixed-alternative questions obviously cannot be used. If the researcher assumes what the responses will be, but is in fact wrong, he or she will have no way of knowing the extent to which the assumption was incorrect. Sometimes this type of error comes to light after the questionnaire has been used. Researchers found cross-cultural misunderstandings in a survey of mothers called the Preschooler Feeding Questionnaire. By talking to a group of African-American mothers, a researcher at the University of Chicago determined that they had experiences with encouraging children to eat more and using food to calm children, but they used different language for these situations than the questionnaire used, so they misinterpreted some questions.⁴

Unanticipated alternatives emerge when respondents believe that closed answers do not adequately reflect their feelings. They may make comments to the interviewer or write additional answers on the questionnaire indicating that the exploratory research did not yield a complete array of responses. After the fact, little can be done to correct a closed question that does not provide the correct responses or enough alternatives. Therefore, a researcher may find exploratory research with open-ended responses valuable before writing a descriptive questionnaire. The researcher should strive to ensure that there are sufficient response choices to include almost all possible answers.

Respondents may check off obvious alternatives, such as *salary* or *health benefits* in an employee survey, if they do not see *opportunities for advancement*, the choice they would prefer. Also, a fixed-alternative question may tempt respondents to check an answer that is more prestigious or socially acceptable than the true answer. Rather than stating that they do not know why they chose a given product, they may select an alternative among those presented, or as a matter of convenience, they may select a given alternative rather than think of the most correct response.

Types of Fixed-Alternative Questions

Here we identify and categorize the various types.

The **simple-dichotomy (dichotomous) question** requires the respondent to choose one of two alternatives. The answer can be a simple “yes” or “no” or a choice between “this” and “that.” For example:

Did you have any overnight travel for work-related activities last month?

- ☐ *Yes* ☐ *No*

Several types of questions provide the respondent with *multiple-choice alternatives*. The **determinant-choice question** requires the respondent to choose one—and only one—response from among several possible alternatives. For example:

Please give us some information about your flight. In which section of the aircraft did you sit?

- ☐ *First class*
☐ *Business class*
☐ *Coach class*

The **frequency-determination question** is a determinant-choice question that asks for an answer about the general frequency of occurrence. For example:

How frequently do you watch MTV?

- ☐ *Every day*
☐ *5—6 times a week*
☐ *2—4 times a week*
☐ *Once a week*
☐ *Less than once a week*
☐ *Never*

Attitude rating scales, such as the Likert scale, semantic differential, Stapel scale, and so on, are also fixed-alternative questions.

The **checklist question** allows the respondent to provide multiple answers to a single question. The respondent indicates past experience, preference, and the like merely by checking off items. In many cases the choices are adjectives that describe a particular object. A typical checklist question might ask the following:

Please check which, if any, of the following sources of information about investments you regularly use.

- ☐ *Personal advice of your broker(s)*
☐ *Brokerage newsletters*
☐ *Brokerage research reports*
☐ *Investment advisory service(s)*
☐ *Conversations with other investors*
☐ *Web page(s)*
☐ *None of these*
☐ *Other (please specify) ____*

A major problem in developing dichotomous or multiple-choice alternatives is establishing the response alternatives. Alternatives should be **totally exhaustive**, meaning that all the response options are covered and that every respondent has an alternative to check. The alternatives should also be **mutually exclusive**, meaning there should be no overlap among categories and only one dimension of an issue should be related to each alternative. So, there is a response category for everyone, but only a single response category for each individual. In other words, a place for everything and everything in its place!

Phrasing Questions for Self-Administered, Telephone, and Personal Interview Surveys

The means of data collection—telephone interview, personal interview, self-administered questionnaire—will influence the question format and question phrasing. In general, questions for telephone in particular, as well as Internet and mail surveys, must be less complex than those used in personal interviews. Questionnaires for telephone and personal interviews should be written in a conversational style. It is particularly important that telephone surveys use easy to understand response categories. Below illustrates how a question may be revised for a different medium.

Mail Form:

How satisfied are you with your community?

- 1 Very satisfied
- 2 Quite satisfied
- 3 Somewhat satisfied
- 4 Slightly satisfied
- 5 Neither satisfied nor dissatisfied
- 6 Slightly dissatisfied
- 7 Somewhat dissatisfied
- 8 Quite dissatisfied
- 9 Very dissatisfied

Revised for Telephone:

How satisfied are you with your community? Would you say you are very satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, or very dissatisfied?

- | | | |
|------------------------------------|---|---|
| Very satisfied | 1 | |
| Somewhat satisfied | 2 | |
| Neither satisfied nor dissatisfied | | 3 |
| Somewhat dissatisfied | 4 | |
| Very dissatisfied | 5 | |

In a telephone survey about attitudes toward police services, the questionnaire not only asked about general attitudes such as how much respondents trust their local police officers and whether the police are “approachable,” “dedicated,” and so on, but also provided basic scenarios to help respondents put their expectations into words. For example, the interviewer asked respondents to imagine that someone had broken into their home and stolen items, and that the respondent called the police to report the crime. The interviewer asked how quickly or slowly the respondent expected the police to arrive. When a question is read aloud, remembering the alternative choices can be difficult.

Guidelines for Constructing Questions

Developing good business research questionnaires is a combination of art and science. Few hard- and-fast rules exist in guiding the development of a questionnaire. Fortunately, research experience has yielded some guidelines that help prevent the most common mistakes. The Research Snapshot above

illustrates problems with question wording in a simple descriptive research project.

Avoid Complexity: Use Simple, Conversational Language

Words used in questionnaires should be readily understandable to all respondents. The researcher usually has the difficult task of adopting the conversational language of people at the lower education levels without talking down to better-educated respondents. Remember, not all people have the vocabulary of a college graduate. In fact, in the U.S., less than 25 percent of the population has a bachelor's degree.

Respondents can probably tell an interviewer whether they are married, single, divorced, separated, or widowed, but providing their *marital status* may present a problem. The technical jargon of top corporate executives should be avoided when surveying retailers or industrial users. "Brand image," "positioning," "marginal analysis," and other corporate language may not have the same meaning for, or even be understood by, a store owner-operator in a retail survey. The vocabulary used in the following question from an attitude survey on social problems probably would confuse many respondents:

*When effluents from a paper mill can be drunk and exhaust from factory smokestacks can be breathed, then humankind will have done a good job in saving the environment. . . .
Don't you agree that what we want is zero toxicity: no effluents?*

Besides being too long and confusing, this question is leading. Survey questions should be short and to the point. Like this:

The stock market is too risky to invest in these days.

Avoid Leading and Loaded Questions

Leading and loaded questions are a major source of bias in question wording. A **leading question** suggests or implies certain answers. A study of the dry cleaning industry asked this question:

*Many people are using dry cleaning less because of improved wash-and-wear clothes.
How do you feel wash-and-wear clothes have affected your use of dry cleaning facilities in the past 4 years?*

- ☐ Use less ☐ No change ☐ Use more

It should be clear that this question leads the respondent to report lower usage of dry cleaning. The potential "bandwagon effect" implied in this question threatens the study's validity.

Partial mention of alternatives is a variation of this phenomenon:

Do accounting graduates who attended state universities, such as Washington State University, make better auditors?

A **loaded question** suggests a socially desirable answer or is emotionally charged. Consider the following question from a survey about media influence on politics:

What most influences your vote in major elections?

- ☐ My own informed opinion
☐ Major media outlets such as CNN
☐ Newspaper endorsements
☐ Popular celebrity opinions
☐ Candidate's physical attractiveness
☐ Family or friends
☐ Video advertising (television or Web video)
☐ Other

The vast majority of respondents chose the first alternative. Although this question is not overly emotionally loaded, many people could be reluctant to say they are swayed by the media or advertising

as opposed to their independent mindset. In fact, a research question dealing with what influences decisions like these may best be done by drawing some inference based on less direct questioning.

Certain answers to questions are more socially desirable than others. For example, a truthful answer to the following classification question might be painful:

Where did you rank academically in your high school graduating class?

- ☐ *Top quarter*
- ☐ *2nd quarter*
- ☐ *3rd quarter*
- ☐ *4th quarter*

When taking personality or psychographic tests, respondents frequently can interpret which answers are most socially acceptable even if those answers do not portray their true feelings. For example, which are the socially desirable answers to the following questions on a self-confidence scale?

I feel capable of handling myself in most social situations.

- ☐ *Agree* ☐ *Disagree*

I fear my actions will cause others to have low opinions of me.

- ☐ *Agree* ☐ *Disagree*

Invoking the status quo is a form of loading that results in bias because most people tend to resist change. An experiment conducted in the early days of polling illustrates the unpopularity of change.

Avoid Ambiguity: Be as Specific as Possible

Items on questionnaires often are ambiguous because they are too general. Consider such indefinite words as *often*, *occasionally*, *regularly*, *frequently*, *many*, *good*, and *poor*. Each of these words has many different meanings. For one consumer, *frequent* reading of *Fortune* magazine may be reading all 25 issues in a year, while another might think 12, or even 6 issues a year is frequent. Earlier, we used the following question as an example of a checklist question:

Please check which, if any, of the following sources of information about investments you regularly use.

What exactly does *regularly* mean? It can certainly vary from respondent to respondent. How exactly does *hardly any* differ from *occasionally*? Where is the cutoff? It is much better to use specific time periods whenever possible.

A brewing industry study on point-of-purchase advertising (store displays) asked their distributors:

How often does the company shut down production for sanitary maintenance?

- ☐ Annually (once a year)
- ☐ Semiannually (once every six months)
- ☐ Quarterly (about every three months)
- ☐ At least once monthly
- ☐ Less frequently (less often than once a year)

Here the researchers clarified the terms *permanent*, *semipermanent*, and *temporary* by defining them for the respondent. However, the question remained somewhat ambiguous. Beer marketers often use a variety of point-of-purchase devices to serve different purposes—in this case, what is the purpose? In addition, analysis was difficult because respondents were merely asked to indicate a preference rather than a *degree* of preference. Thus, the meaning of a question may not be clear because the frame of reference is inadequate for interpreting the context of the question.

A student research group asked this question:

What media do you rely on most?

- ☐ *Television*
- ☐ *Radio*
- ☐ *Internet*
- ☐ *Newspapers*

This question is ambiguous because it does not provide information about the context. “Rely on most” for what—news, sports, entertainment? When—while getting dressed in the morning, driving to work, at home in the evening? Knowing the specific circumstance can affect the choice made.

Each of these examples shows how a question can be ambiguous and interpreted differently by different individuals. While we might not be able to completely eliminate ambiguity, by using words or descriptions that have universal meaning, replacing terms with specific response categories, and defining the situation surrounding the question, we can improve our business research questionnaires.

Avoid Double-Barreled Items

A question covering several issues at once is referred to as a **double-barreled question** and should always be avoided. Making the mistake of asking two questions rather than one is easy—for example, “Do you feel our hospital emergency room waiting area is clean and comfortable?” What do we learn from this question? If the respondent responds positively, we could likely infer that our waiting area is clean and comfortable. However, if the response is negative, is it because the room is not clean, or not comfortable? Or both? Certainly for a manager to make improvements it is important to know which element needs attention. When multiple questions are asked in one question, the results may be exceedingly difficult to interpret.

One of the questions we presented earlier when discussing fixed-alternative questions provides a good example of a double-barreled question:

Did your plant use any commercial feed or supplement for livestock or poultry in 2010?
☐ *Yes* ☐ *No*

Here, the question could actually be thought of as a “double-double-barreled” question. Both *commercial feed or supplement* and *livestock or poultry* are double barreled. Interpreting the answer to this question would be challenging.

A researcher is well served to carefully examine any survey question that includes the words *and* or *or*. While sometimes words such as these may be used to reinforce or clarify a question, they are often a sign of a double-barreled question. If you have two (or three) questions, ask them separately, not all together.

Avoid Making Assumptions

Consider the following question:

Should General Electric continue to pay its outstanding quarterly dividends?
☐ *Yes* ☐ *No*

This question has a built-in assumption: that people believe the dividends paid by General Electric are outstanding. By answering “yes,” the respondent implies that the program is, in fact, outstanding and that things are fine just as they are. When a respondent answers “no,” he or she implies that GE should discontinue the dividends. The researchers should not place the respondent in that sort of bind by including an implicit assumption in the question.

Another frequent mistake is assuming that the respondent had previously thought about an issue. For example, the following question appeared in a survey concerning Jack-in-the-Box: “Do you think

Jack-in-the-Box restaurants should consider changing their name?” Respondents have not likely thought about this question beforehand. Most respondents answered the question even though they had no prior opinion concerning the name change. Research that induces people to express attitudes on subjects they do not ordinarily think about is rather meaningless.

Avoid Burdensome Questions That May Tax the Respondent's Memory

A simple fact of human life is that people forget. Researchers writing questions about past behavior or events should recognize that certain questions may make serious demands on the respondent's memory. Writing questions about prior events requires a conscientious attempt to minimize the problems associated with forgetting.

In many situations, respondents cannot recall the answer to a question. For example, a telephone survey conducted during the 24-hour period following the airing of the Super Bowl might establish whether the respondent watched the Super Bowl and then ask, “Do you recall any commercials on that program?” If the answer is positive, the interviewer might ask, “What brands were advertised?” These two questions measure *unaided recall*, because they give the respondent no clue as to the brand of interest.

If the researcher suspects that the respondent may have forgotten the answer to a question, he or she may rewrite the question in an *aided-recall* format—that is, in a format that provides a clue to help jog the respondent's memory. For instance, the question about an advertised beer in an aided-recall format might be “Do you recall whether there was a brand of beer advertised on that program?” or “I am going to read you a list of beer brand names. Can you pick out the name of the beer that was advertised on the program?” While aided recall is not as strong a test of attention or memory as unaided recall, it is less taxing to the respondent's memory.

Telescoping and squishing are two additional consequences of respondents' forgetting the exact details of their behavior. *Telescoping error* occurs when respondents believe that past event happened more recently than they actually did. For instance, most people will estimate that they have changed the oil in their car more recently than they actually have. The opposite effect, *squishing error*, occurs when respondents think that recent events took place longer ago than they really did. A solution to this problem may be to refer to a specific event that is memorable—for example, “How often have you gone to a sporting event since the World Series?” Because forgetting tends to increase over time, the question may concern a recent period: “How often did you watch HBO on cable television last week?” During pretesting or the questionnaire editing stage, the most appropriate time period can be determined.

Make Certain Questions Generate Variance

We want our variables to vary! It is important that the response categories provided cover the breadth of possibilities (totally exhaustive), but also critical that they yield variance across respondents. In many ways, if all of the respondents check the same box, we have not generated usable information.

For example, the U.S. census uses the following age categories:

Under 5 years

5 to 9 years

10 to 14 years

15 to 19 years

20 to 24 years

25 to 29 years

95 to 99 years

100 years and over

While these five-year age categories do capture the range of ages and provide rather detailed census information regarding the general population, what would happen if they were used for a survey of undergraduate students? In many institutions, 95 percent or more of the respondents would fall into two groups. What might be more appropriate and provide better information in a study of undergraduates?

In practice, it is also often better to use a scaled response than a dichotomous response form. For example, our earlier example of a simple-dichotomy (dichotomous) question asked:

Did you have any overnight travel for work-related activities last month?

☐ *Yes* ☐ *No*

While the respondent could likely answer this question and we may simply desire to place respondents into either the “did travel” or “did not travel” category, we really do not gain much information from this question. It fails to discriminate at all between employees that travel once a month, twice a month, or were gone for 25 days last month. It is likely that these employees have different attitudes and needs regarding business travel. A better approach might be to create multiple categories (0, 1—5, 6—10, 11—15, 16—20, 21—25, 26+ nights) or ask for a specific number of nights away on business travel. From this, we could always recode the respondents into the nominal data categories of yes/no if needed. However, if we collect yes/do data to begin with, we cannot make more detailed distinctions later.

What Is the Best Question Sequence?

The order of questions, or the question sequence, may serve several functions for the researcher. If the opening questions are interesting, simple to comprehend and easy to answer, respondents’ cooperation and involvement can be maintained throughout the questionnaire. Asking easy-to-answer questions teaches respondents their role and builds their confidence.

A mail survey among department store buyers drew an extremely poor return rate. A substantial improvement in response rate occurred, however, when researchers added some introductory questions seeking opinions on pending legislation of great importance to these buyers. Respondents continued on to complete all the questions, not only those in the opening section.

In their attempt to “warm up” respondents toward the questionnaire, student researchers frequently ask demographic or classification questions at the beginning of the survey. This generally is not advisable, because asking for personal information such as income level or education may embarrass or threaten respondents. Asking these questions at the end of the questionnaire usually is better, after rapport has been established between respondent and interviewer.

Order bias can result from a particular answer’s position in a set of answers or from the sequencing of questions. In political elections in which candidates lack high visibility, such as elections for county commissioners and judges, the first name listed on the ballot often receives the highest percentage of votes. For this reason, many election boards print several ballots so that each candidate’s name appears in every possible position on the ballot.

Order bias can also distort survey results. For example, suppose a questionnaire’s purpose is to measure levels of awareness of several charitable organizations. If Big Brothers and Big Sisters is always mentioned first, the American Red Cross second, and the American Cancer Society third, Big Brothers and Big Sisters may receive an artificially high awareness rating because respondents are prone to yea-saying (by indicating awareness of the first item in the list).

Asking specific questions before asking about broader issues is a common cause of order bias. For example, people who are first asked, “Are you satisfied with your marriage?” will respond differently

to a follow-up question that asks, “Are you satisfied with your life?” than if the questions are asked in the reverse order. Generally, researchers should ask general questions before specific questions. This procedure, known as the **funnel technique**, allows the researcher to understand the respondent’s frame of reference before asking more specific questions about the level of the respondent’s information and the intensity of his or her opinions.

Consider how later answers might be biased by previous questions in this questionnaire on environmental pollution:

Please consider each of the following issues. Circle the number for each that best indicates your feelings about the severity of that issue as an environmental problem:

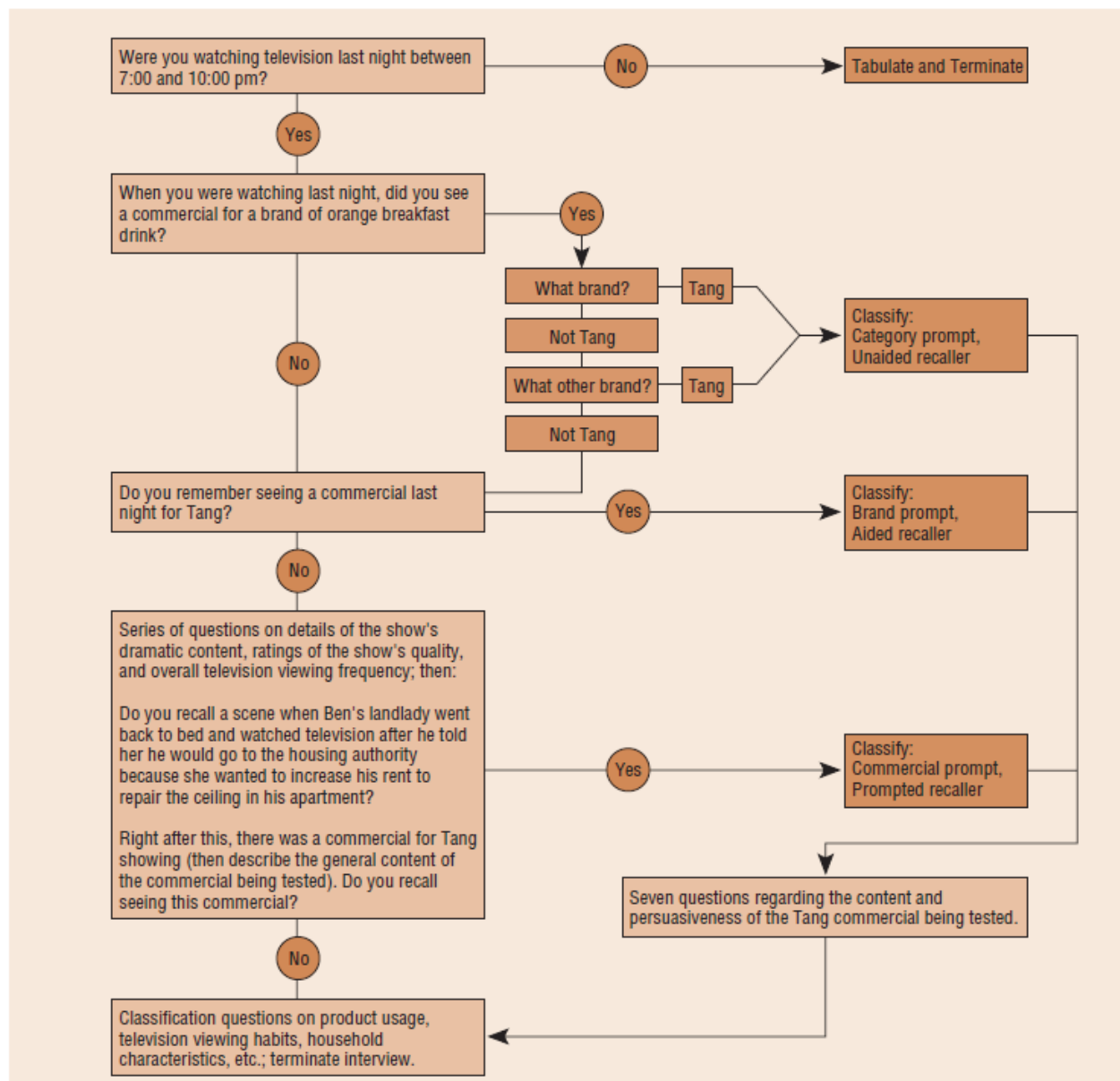
Issue	Not At All A Problem			Very Severe Problem	
<i>Air pollution from automobile exhausts</i>	1	2	3	4	5
<i>Air pollution from open burning</i>	1	2	3	4	5
<i>Air pollution from industrial smoke</i>	1	2	3	4	5
<i>Air pollution from foul odors</i>	1	2	3	4	5
<i>Noise pollution from airplanes</i>	1	2	3	4	5
<i>Noise pollution from cars, trucks,</i>	1	2	3	4	5
<i>Noise pollution from industry</i>	1	2	3	4	5

Not surprisingly, researchers found that the responses to the air pollution questions were highly correlated—in fact, almost identical. What if the first issue was *foul odors* instead of *automobile exhaust*? Do you think it would affect the remaining responses?

With attitude scales, there also may be an *anchoring effect*. The first concept measured tends to become a comparison point from which subsequent evaluations are made. Randomization of items on a questionnaire susceptible to the anchoring effect helps minimize order bias.

A related problem is bias caused by the order of alternatives on closed questions. To avoid this problem, the order of these choices should be rotated if producing alternative forms of the questionnaire is possible. Unfortunately, business researchers rarely print alternative questionnaires to eliminate problems resulting from order bias. With Internet surveys, however, reducing order bias by having the computer randomly order questions and/or response alternatives is quite easy. With complete randomization, question order is random and respondents see response alternatives in different positions/

Asking a question that does not apply to the respondent or that the respondent is not qualified to answer may be irritating or cause a biased response because the respondent wishes to please the interviewer or to avoid embarrassment. Including a **filter question** minimizes the chance of asking questions that are inapplicable. Asking a human resource manager “How would you rate the third party administrator (TPA) of your employee health plan?” may elicit a response even though the organization does not utilize a TPA. The respondent may wish to please the interviewer with an answer. A filter question such as “Does your organization use a third party administrator (TPA) for your employee health plan?” followed by “If you answered *Yes* to the previous question, how would you rate your TPA on . . . ?” would screen out the people who are not qualified to answer.

EXHIBIT 15.2 Flow of Questions to Determine the Level of Prompting Required to Stimulate Recall

What Is the Best Layout?

Good layout and physical attractiveness are crucial in mail, Internet, and other self-administered questionnaires. For different reasons, a good layout in questionnaires designed for personal and telephone interviews is also important.

Traditional Questionnaires

Often rate of return can be increased by using money that might have been spent on an incentive to improve the attractiveness and quality of the questionnaire. Mail questionnaires should never be overcrowded. Margins should be of decent size, white space should be used to separate blocks of print, and the unavoidable columns of multiple boxes should be kept to a minimum. A question should not begin on one page and end on another page. Splitting questions may cause a respondent to read only part of a question, to pay less attention to answers on one of the pages, or to become confused.

Questionnaires should be designed to appear as short as possible. Sometimes it is advisable to use a booklet form of questionnaire rather than stapling a large number of pages together. In situations in which it is necessary to conserve space on the questionnaire or to

facilitate data entry or tabulation of the data, a multiple-grid layout may be used. The **multiple-grid question** presents several similar questions and corresponding response alternatives arranged in a grid format. For example.

Airlines often offer special fare promotions, but they may require connecting flights. On a vacation trip, how often would you take a connecting flight instead of a nonstop flight if you could save \$100 a ticket, but the connecting flight was longer?

	<i>Never</i>	<i>Rare</i>	<i>Sometime</i>	<i>Often</i>	<i>Alwa</i>
<i>Complete trip is one hour</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Complete trip is two hours</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Complete trip is three</i>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Airlines often offer special fare promotions, but they may require connecting flights. On a vacation trip, how often would you take a connecting flight instead of a nonstop flight if you could save \$100 a ticket, but the connecting flight was longer?

Experienced researchers have found that the title of a questionnaire should be phrased carefully. In self-administered and mail questionnaires, a carefully constructed title may capture the respondent's interest, underline the importance of the research ("Nationwide Study of Blood Donors"), emphasize the interesting nature of the study ("Study of Internet Usage"), appeal to the respondent's ego ("Survey of Top Executives"), or emphasize the confidential nature of the study ("A Confidential Survey of Physicians"). At the same time, the researcher should take steps to ensure that the wording of the title will not bias the respondent in the same way that a leading question might.

By using several forms, special instructions, and other tricks of the trade, the researcher can design the questionnaire to facilitate the interviewer's job of following interconnected questions.

Instructions are often capitalized or printed in bold to alert the interviewer that it may be necessary to proceed in a certain way. For example, if a particular answer is given, the interviewer or respondent may be instructed to skip certain questions or go to a special sequence of questions.

■ LAYOUT ISSUES

Even if the questionnaire designer's computer and the respondents' computers are compatible, a Web questionnaire designer should consider several layout issues. The first decision is whether the questionnaire will appear page by page, with individual questions or groups of questions on separate screens (Web pages), or on a scrolling basis, with the entire questionnaire appearing on a single Web page that the respondent scrolls from top to bottom. The *paging layout* (going from screen to screen) greatly facilitates skip patterns. Based on a respondent's answers to filter questions, the computer can automatically insert relevant questions on subsequent pages. If the entire questionnaire appears on one page (the *scrolling layout*), the display should advance smoothly, as if it were a piece of paper being moved up or down. The scrolling layout gives the respondent the ability to read any portion of the questionnaire at any time, but the absence of page boundaries can cause problems. For example, suppose a Likert scale consists of 15 statements in a grid-format layout, with the response categories Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree at the beginning of the questionnaire. Once the respondent has scrolled down beyond the first few statements, he or she may not be able to see both the statements at the end of the list and the response categories at the top of the grid simultaneously. Thus, avoiding the problems associated with splitting questions and response categories may be difficult with scrolling questionnaires.

When a scrolling questionnaire is long, category or section headings are helpful to respondents. It is also a good idea to provide links to the top and bottom parts of each section, so that users can navigate through the questionnaire without having to scroll through the entire document.¹¹

Whether a Web survey is page-by-page or scrolling format a **push button** with a label should clearly describe the actions to be taken. For example, if the respondent is to go to the next page, a large arrow labeled “NEXT” might appear in color at the bottom of the screen.

Decisions must be made about the use of color, graphics, animation, sound, and other special features that the Internet makes possible. One point to remember is that, although sophisticated graphics are not a problem for most people with powerful computers and high speed Internet, many respondents’ computers and/or Internet connections are not powerful enough to deliver complex graphics at a satisfactory speed.

With a paper questionnaire, the respondent knows how many questions he or she must answer. Because many Internet surveys offer no visual clues about the number of questions to be asked, it is important to provide a **status bar** or some other visual indicator of questionnaire length. For example, including a partially filled rectangular box as a visual symbol and a statement such as “The status bar at top right indicates approximately what portion of the survey you have completed” increases the likelihood that the respondent will finish the entire sequence of questions.

An Internet questionnaire uses dialog boxes to display questions and record answers. Exhibit 15.8 portrays four common ways of displaying questions on a computer screen. Many Internet questionnaires require the respondent to activate his or her answer by clicking on the **radio button** for a response. Radio buttons work like push buttons on automobile radios: Clicking on an alternative response deactivates the first choice and replaces it with the new response.

Checklist questions may be followed by **check boxes**, several, none, or all of which may be checked by the respondent. **Open-ended boxes** are boxes in which respondents type their answers to open-ended questions. Open-ended boxes may be designed as *one-line text boxes* or *scrolling text boxes*, depending on the breadth of the expected answer. Of course, open-ended questions require that respondents have both the skill and the willingness to keyboard lengthy answers on the computer. Some open-ended boxes are designed so that respondents can enter numbers for frequency response, ranking, or rating questions. For example,

Below you will see a series of statements that might or might not describe how you feel about your career. Please rate each statement using a scale from 1 to 5, where 1 means “Totally Disagree,” 2 means “Somewhat Disagree,” 3 means “Neither Agree nor Disagree,” 4 means “Somewhat Agree,” and 5 means “Totally Agree.” Please enter your numeric answer in the box provided next to each statement. Would you say that . . .

A lack of business knowledge relevant to my field/career could hurt my career advancement.

My career life is an important part of how I define myself.

I am seriously considering a change in careers.

Pop-up boxes are message boxes that can be used to highlight important information. For example, pop-up boxes may be used to provide a privacy statement, such as the following:

IBM would like your help in making our Web site easier to use and more effective. Choose to complete the survey now or not at all.

Clicking on Privacy Statement opens the following pop-up box:

Survey Privacy Statement

This overall Privacy Statement verifies that IBM is a member of the TRUSTe program and is in compliance with TRUSTe principles. This survey is strictly for market research purposes. The information you provide will be used only to improve the overall content, navigation, and usability of ibm.com.

SOFTWARE THAT MAKES QUESTIONNAIRES INTERACTIVE

Computer code can be written to make Internet questionnaires interactive and less prone to errors. The writing of software programs is beyond the scope of this discussion. However, several of the interactive functions that software makes possible should be mentioned here.

Internet software allows the branching off of questioning into two or more different lines, depending on a particular respondent's answer, and the skipping or filtering of questions. Questionnaire-writing software with skip and branching logic is readily available. Most of these programs have *hidden skip logic* so that respondents never see any evidence of skips. It is best if the questions the respondent sees flow in numerical sequence. However, some programs number all potential questions in numerical order, and the respondent sees only the numbers on the questions he or she answers. Thus, a respondent may answer questions 1 through 11 and then next see a question numbered 15 because of the skip logic.

Software can systematically or randomly manipulate the questions a respondent sees. **Variable piping software** allows variables, such as answers from previous questions, to be inserted into unfolding questions. Other software can randomly rotate the order of questions, blocks of questions, and response alternatives from respondent to respondent.

Researchers can also use software to control the flow of a questionnaire. Respondents can be blocked from backing up, or they can be allowed to stop in mid-questionnaire and come back later to finish. A questionnaire can be designed so that if the respondent fails to answer a question or answers it with an incorrect type of response, an immediate error message appears. This is called **error trapping**. With **forced answering software**, respondents cannot skip over questions as they do in mail surveys. The program will not let them continue if they fail to answer a question. The software may insert a boldfaced error message on the question screen or insert a pop-up box instructing the respondent how to continue. For example, if a respondent does not answer a question and tries to proceed to another screen, a pop-up box might present the following message:

You cannot leave a question blank. On questions without a "Not sure" or "Decline to answer" option, please choose the response that best represents your opinions or experiences.

The respondent must close the pop-up box and answer the question in order to proceed to the next screen.

Some designers include an **interactive help desk** in their Web questionnaire so that respondents can solve problems they encounter in completing a questionnaire. A respondent might e-mail questions to the survey help desk or get live, interactive, real-time support via an online help desk.

Some respondents will leave the questionnaire Web site, prematurely terminating the survey. In many cases sending an e-mail message to these respondents at a later date, encouraging them to revisit the Web site, will persuade them to complete the questionnaire. Through the use of software and cookies, researchers can make sure that the respondent who revisits the Web site will be able to pick up at the point where he or she left off.

Once an Internet questionnaire has been designed, it is important to pretest it to ensure that it works with Internet Explorer, Mozilla Firefox, Safari, Opera, Maxthon, and other browsers.

How Much Pretesting and Revising Are Necessary?

Many novelists write, rewrite, revise, and rewrite again certain chapters, paragraphs, or even sentences. The researcher works in a similar world. Rarely—if ever—does he or she write only a first draft of a questionnaire. Usually the questionnaire is written, revised, shared with others for feedback, then revised again. After that, it is tried out on a group, selected on a convenience basis, that is similar in makeup to the one that ultimately will be sampled. Although the researcher should not select a group too divergent from the target market—for example, selecting business students as surrogates for

businesspeople—pretesting does not require a statistical sample. The pretesting process allows the researcher to determine whether respondents have any difficulty understanding the questionnaire and whether there are any ambiguous or biased questions. This process is exceedingly beneficial. Making a mistake with 25 or 50 subjects can avoid the potential disaster of administering an invalid questionnaire to several hundred individuals. For a questionnaire investigating teaching-students' experience with Web-based instruction, the researcher had the questionnaire reviewed first by university faculty members to ensure the questions were valid, then asked 20 teaching students to try answering the questions and indicate any ambiguities they noticed. Their feedback prompted changes in the format and wording. Pretesting was especially helpful because the English-language questionnaire was used in a school in the United Arab Emirates, where English is spoken but is not the primary language.

Tabulating the results of a pretest helps determine whether the questionnaire will meet the objectives of the research. A **preliminary tabulation** often illustrates that, although respondents can easily comprehend and answer a given question, that question is inappropriate because it does not provide relevant information to help solve the business problem. Consider the following example from a survey among distributors of power-actuated tools such as stud drivers concerning the percentage of sales to given industries:

Please estimate what percentage of your fastener and load sales go to the following industries:

- % heating, plumbing, and air conditioning
- % carpentry
- % electrical
- % maintenance
- % other (please specify)

The researchers were fortunate to learn that asking the question in this manner made it virtually impossible to obtain the information actually desired. The categories are rather vague, a high percentage may fall into the *Other* category, and most respondents' answers did not total 100 percent. As a result, the question had to be revised. In general, getting respondents to add everything correctly is a difficult task, and virtually impossible if they can not see all the categories (not a good idea for a telephone survey!). Pretesting difficult questions such as these is essential.

What administrative procedures should be implemented to maximize the value of a pretest? Administering a questionnaire exactly as planned in the actual study often is not possible. For example, mailing out a questionnaire is quite expensive and might require several weeks that simply cannot be spared. Pretesting a questionnaire in this manner would provide important information on response rate, but may not point out why questions were skipped or what questions are ambiguous or confusing. Personal interviewers can record requests for additional explanation or comments that indicate respondents' difficulty with question sequence or other factors. This is the primary reason why interviewers are often used for pretest work. Self-administered questionnaires are not reworded to be personal interviews, but interviewers are instructed to observe respondents and ask for their comments after they complete the questionnaire. When pretesting personal or telephone interviews, interviewers may test alternative wordings and question sequences to determine which format best suits the intended respondents.

No matter how the pretest is conducted, the researcher should remember that its purpose is to uncover any problems that the questionnaire may cause. Thus, pretests typically are conducted to answer questions about the questionnaire such as the following:

- Can the questionnaire format be followed by the interviewer?
- Does the questionnaire flow naturally and conversationally?
- Are the questions clear and easy to understand?
- Can respondents answer the questions easily?

- Which alternative forms of questions work best?

Pretests also provide means for testing the sampling procedure—to determine, for example, whether interviewers are following the sampling instructions properly and whether the procedure is efficient. Pretests also provide estimates of the response rates for mail surveys and the completion rates for telephone surveys.

Usually a questionnaire goes through several revisions. The exact number of revisions depends on the researcher's and client's judgment. The revision process usually ends when both agree that the desired information is being collected in an unbiased manner.

Designing Questionnaires for Global Markets

Now that business research is being conducted around the globe, researchers must take cultural factors into account when designing questionnaires. The most common problem involves translating a questionnaire into other languages. A questionnaire developed in one country may be difficult to translate because equivalent language concepts do not exist or because of differences in idiom and vernacular. Although Spanish is spoken in both Mexico and Venezuela, one researcher found out that the Spanish translation of the English term *retail outlet* works in Mexico but not in Venezuela. Venezuelans interpreted the translation to refer to an electrical outlet, an outlet of a river into an ocean, or the passageway onto a patio.

Counting on an international audience to speak a common language such as English does not necessarily bridge these gaps, even when the respondents actually do speak more than one language. Cultural differences incorporate many shades of meaning that may not be captured by a survey delivered in a language used primarily for, say, business transactions. In a test of this idea, undergraduate students in 24 countries completed questionnaires about attitudes toward school and career. Half received the questionnaire in English, and half in their native language. The results varied, with country-to-country differences being smaller when students completed the questionnaire in English.¹⁵

International researchers often have questionnaires back translated. **Back translation** is the process of taking a questionnaire that has previously been translated from one language to another and having it translated back again by a second, independent translator. The back translator is often a person whose native tongue is the language that will be used for the questionnaire. This process can reveal inconsistencies between the English version and the translation. For example, when a soft-drink company translated its slogan “Baby, it’s cold inside” into Cantonese for research in Hong Kong, the result read “Small Mosquito, on the inside, it is very cold.” In Hong Kong, *small mosquito* is a colloquial expression for a small child. Obviously the intended meaning of the advertising message had been lost in the translated questionnaire.¹⁶

Literacy rates also influences the designs of self-administered questionnaires and interviews. Knowledge of the literacy rates in foreign countries, especially those that are just developing modern economies, is vital.

SAMPLING DESIGNS AND SAMPLING PROCEDURES

Introduction

Sampling is a familiar part of daily life. A customer in a bookstore picks up a book, looks at the cover, and skims a few pages to get a sense of the writing style and content before deciding whether to buy. A high school student visits a college classroom to listen to a professor's lecture. Selecting a university on the basis of one classroom visit may not be scientific sampling, but in a personal situation, it may be a practical sampling experience. When measuring every item in a population is impossible, inconvenient, or too expensive, we intuitively take a sample.

Although sampling is commonplace in daily activities, these familiar samples are seldom scientific. For researchers, the process of sampling can be quite complex. Sampling is a central aspect of business research, requiring in-depth examination. This chapter explains the nature of sampling and ways to determine the appropriate sample design.

Sampling Terminology

As seen in the chapter vignette above, the process of sampling involves using a portion of a population to make conclusions about the whole population. A **sample** is a subset, or some part, of a larger population. The purpose of sampling is to estimate an unknown characteristic of a population.

Sampling is defined in terms of the population being studied. A **population (universe)** is any complete group—for example, of people, sales territories, stores, or college students—that shares some common set of characteristics. The term **population element** refers to an individual member of the population.

Researchers could study every element of a population to draw some conclusion. A **census** is an investigation of all the individual elements that make up the population—a total enumeration rather than a sample. Thus, if we wished to know whether more adult Texans drive pickup trucks than sedans, we could contact every adult Texan and find out whether or not they drive a pickup truck or a sedan. We would then know the answer to this question definitively.

Why Sample?

At a wine tasting, guests sample wine by having a small taste from each of a number of different wines. From this, the taster decides if he or she likes a particular wine and if it is judged to be of low or high quality. If an entire bottle were consumed to decide, the taster may end up not caring care about the next bottle. However, in a scientific study in which the objective is to determine an unknown population value, why should a sample rather than a complete census be taken?

1-Pragmatic Reasons

Applied business research projects usually have budget and time constraints. If Ford Motor Corporation wished to take a census of past purchasers' reactions to the company's recalls of defective models, the researchers would have to contact millions of automobile buyers. Some of them would be inaccessible (for example, out of the country), and it would be impossible to contact all these people within a short time period.

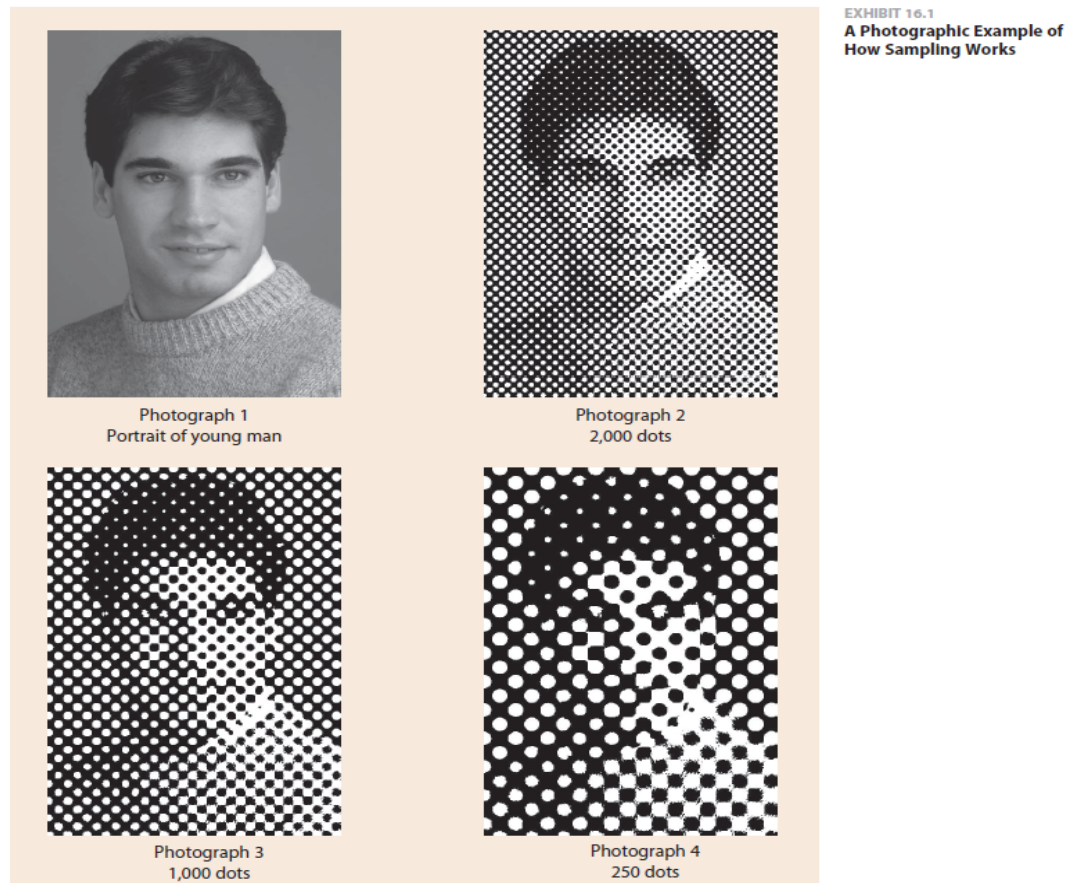
A researcher who wants to investigate a population with an extremely small number of population elements may elect to conduct a census rather than a sample because the cost, labor, and time drawbacks would be relatively insignificant. For a company that wants to assess salespersons' satisfaction with its computer networking system, circulating a questionnaire to all 25 of its employees is practical. In most situations, however, many practical reasons favor sampling. Sampling cuts costs, reduces labor requirements, and gathers vital information quickly. These advantages may be sufficient

2-Accurate and Reliable Results

As seen in the Research, major reason for sampling is that most properly selected samples give results that are reasonably accurate. If the elements of a population are quite similar, only a small sample is necessary to accurately portray the characteristic of interest. Thus, a population consisting of 10,000

eleventh grade students in all-boys Catholic high schools will require a smaller sample than a broader population consisting of 10,000 high school students from coeducational secondary schools.

A visual example of how different-sized samples produce generalizable conclusions is provided in Exhibit 16.1. All are JPEG images that contain different numbers of “dots.” More dots mean more memory is required to store the photo. In this case, the dots can be thought of as sampling units representing the population which can be thought of as all the little pieces of detail that form the actual image.



The first photograph is comprised of thousands of dots resulting in a very detailed photograph. Very little detail is lost and the face can be confidently recognized. The other photographs provide less detail. Photograph 2 consists of approximately 2,000 dots. The face is still very recognizable, but less detail is retained than in the first photograph. Photograph 3 is made up of 1,000 dots, constituting a sample that is only half as large as that in photograph 2. The 1,000-dot sample provides an image that can still be recognized. Photograph 4 consists of only 250 dots. Yet, if you look at the picture at a distance, you can still recognize the face. The 250-dot sample is still useful, although some detail is lost and under some circumstances (such as looking at it from a short distance) we have less confidence in judging the image using this sample. *Precision* has suffered, but *accuracy* has not.

A sample may on occasion be more accurate than a census. Interviewer mistakes, tabulation errors, and other nonsampling errors may increase during a census because of the increased volume of work. In a sample, increased accuracy may sometimes be possible because the fieldwork and tabulation of data can be more closely supervised. In a field survey, a small, well-trained, closely supervised group may do a more careful and accurate job of collecting information than a large group of nonprofessional interviewers who try to contact everyone. An interesting case in point is the use of samples by the Bureau of the Census to check the accuracy of the U.S. Census. If the sample indicates a possible source of error, the census is redone.

3-Destruction of Test Units

Many research projects, especially those in quality-control testing, require the destruction of the items being tested. If a manufacturer of firecrackers wished to find out whether each unit met a specific production standard, no product would be left after the testing. This is the exact situation in many research strategy experiments. For example, if an experimental sales presentation were presented to every potential customer, no prospects would remain to be contacted after the experiment. In other words, if there is a finite population and everyone in the population participates in the research and cannot be replaced, no population elements remain to be selected as sampling units. The test units have been destroyed or ruined for the purpose of the research project.

Practical Sampling Concepts

Defining the Target Population

Once the decision to sample has been made, the first question concerns identifying the target population. What is the relevant population? In many cases this question is easy to answer. Registered voters may be clearly identifiable. Likewise, if a company's 106-person sales force is the population of concern, there are few definitional problems. In other cases the decision may be difficult. One survey concerning organizational buyer behavior incorrectly defined the population as purchasing agents whom sales representatives regularly contacted. After the survey, investigators discovered that industrial engineers within the customer companies rarely talked with the salespeople but substantially affected buying decisions. For consumer-related research, the appropriate population element frequently is the household rather than an individual member of the household. This presents some problems if household lists are not available.

At the outset of the sampling process, the target population must be carefully defined so that the proper sources from which the data are to be collected can be identified. The usual technique for defining the target population is to answer questions about the crucial characteristics of the population. Does the term *comic book reader* include children under six years of age who do not actually read the words? Does *all persons west of the Mississippi* include people in east bank towns that border the river, such as East St. Louis, Illinois? The question to answer is, "Whom do we want to talk to?" The answer may be users, nonusers, recent adopters, or brand switchers.

To implement the sample in the field, tangible characteristics should be used to define the population. A baby food manufacturer might define the population as all women still capable of bearing children. However, a more specific *operational definition* would be women between the ages of 12 and 50. While this definition by age may exclude a few women who are capable of childbearing and include some who are not, it is still more explicit and provides a manageable basis for the sample design.

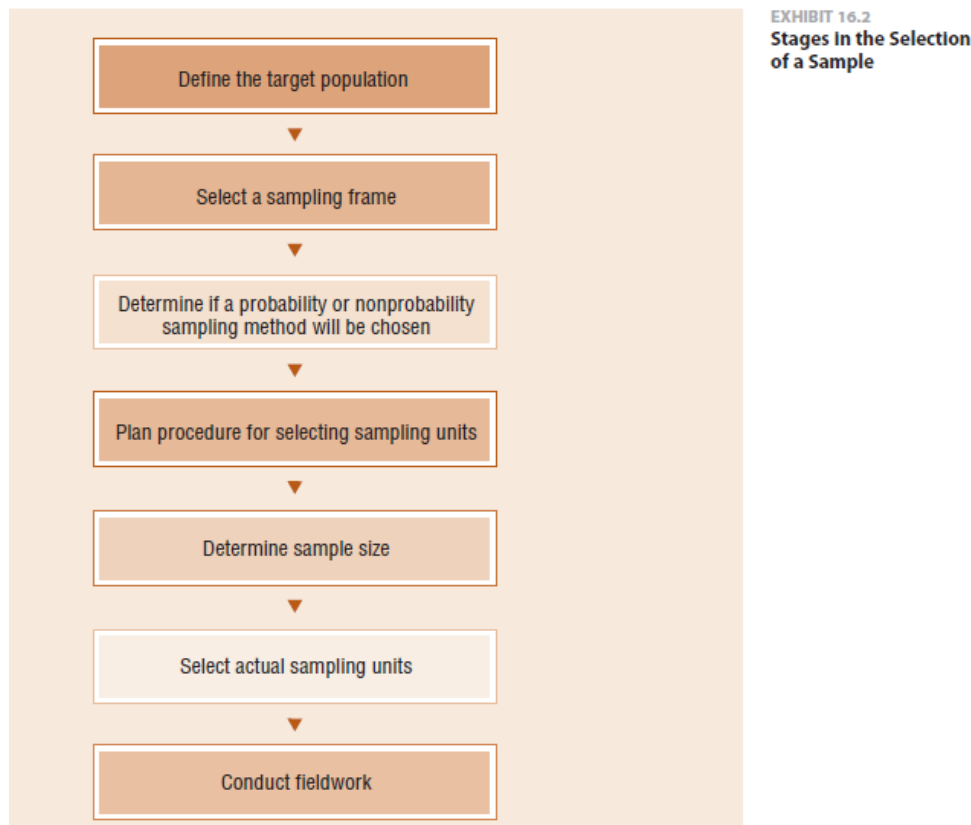
The Sampling Frame

In practice, the sample will be drawn from a list of population elements that often differs somewhat from the defined target population. A list of elements from which the sample may be drawn is called a **sampling frame**. The sampling frame is also called the *working population* because these units will eventually provide units involved in analysis. A simple example of a sampling frame would be a list of all members of the American Medical Association.

In practice, almost every list excludes some members of the population. For example, would a university e-mail directory provide an accurate sampling frame for a given university's student population? Perhaps the sampling frame excludes students who registered late and includes students who have resigned from the university. The e-mail directory also will likely list only the student's official university e-mail address. However, many students may not ever use this address, opting to use a private e-mail account instead. Thus, the university e-mail directory could not be expected to perfectly represent the student population. However, a perfect representation isn't always possible or

needed.

Some firms, called *sampling services* or *list brokers*, specialize in providing lists or databases that include the names, addresses, phone numbers, and e-mail addresses of specific populations. Exhibit 16.3 shows a page from a mailing list company's offerings. Lists offered by companies such as this are compiled from subscriptions to professional journals, credit card applications, warranty card registrations, and a variety of other sources. One sampling service obtained its listing of households with children from an ice cream retailer who gave away free ice cream cones on children's birthdays. The children filled out cards with their names, addresses, and birthdays, which the retailer then sold to the mailing list company.



A **sampling frame error** occurs when certain sample elements are excluded or when the entire population is not accurately represented in the sampling frame. Election polling that used a telephone directory as a sampling frame would be contacting households with listed phone numbers, not households whose members are likely to vote. A better sampling frame might be voter registration records. Another potential sampling frame error involving phone records is the possibility that a phone survey could underrepresent people with disabilities. Some disabilities, such as hearing and speech impairments, might make telephone use impossible.

■ SAMPLING FRAMES FOR INTERNATIONAL RESEARCH

The availability of sampling frames around the globe varies dramatically. Not every country's government conducts a census of population. In some countries telephone directories are incomplete, no voter registration lists exist, and accurate maps of urban areas are unobtainable. However, in Taiwan, Japan, and other Asian countries, a researcher can build a sampling frame relatively easily because those governments release some census information. If a family changes households, updated census information must be reported to a centralized government agency before communal services (water, gas, electricity, education, and so on) are made available.³ This information is then easily accessible in the local *Inhabitants' Register*.

Random Sampling and Nonsampling Errors

An advertising agency sampled a small number of shoppers in grocery stores that used Shopper's Video, an in-store advertising network. The agency hoped to measure brand awareness and purchase intentions. Investigators expected this sample to be representative of the grocery-shopping population. However, if a difference exists between the value of a sample statistic of interest (for example, the sample group's average willingness to buy the advertised brand) and the value of the corresponding population parameter (the population's average willingness to buy), a *statistical error* has occurred. Two basic causes of differences between statistics and parameters were introduced in an earlier chapter and are described below:

1. *random sampling errors*
2. *systematic (nonsampling) error*

An estimation made from a sample is not the same as a census count. **Random sampling error** is the difference between the sample result and the result of a census conducted using identical procedures. Of course, the result of a census is unknown unless one is taken, which is rarely done. Other sources of error also can be present. Random sampling error occurs because of chance variation in the scientific selection of sampling units. The sampling units, even if properly selected according to sampling theory, may not perfectly represent the population, but generally they are reliable estimates.

Random Sampling Error

Random sampling error is a function of sample size. As sample size increases, random sampling error decreases. Of course, the resources available will influence how large a sample may be taken. It is possible to estimate the random sampling error that may be expected with various sample sizes. Suppose a survey of approximately 1,000 people has been taken in Fresno to determine the feasibility of a new soccer franchise. Assume that 30 percent of the respondents favor the idea of a new professional sport in town. The researcher will know, based on the laws of probability, that 95 percent of the time a survey of slightly fewer than 900 people will produce results with an error of approximately plus or minus 3 percent. If the survey were conducted with only 325 people, the margin of error would increase to approximately plus or minus 5 percentage points. This example illustrates random sampling errors.

Systematic Sampling Error

Systematic (nonsampling) errors result from nonsampling factors, primarily the nature of a study's design and the correctness of execution. These errors are *not* due to chance fluctuations. For example, highly educated respondents are more likely to cooperate with mail surveys than poorly educated ones, for whom filling out forms is more difficult and intimidating. Sample biases such as these account for a large portion of errors in marketing research. The term *sample bias* is somewhat unfortunate, because many forms of bias are not related to the selection of the sample.

Less Than Perfectly Representative Samples

Random sampling errors and systematic errors associated with the sampling process may combine to yield a sample that is less than perfectly representative of the population. The total population is represented by the area of the largest square. Sampling frame errors eliminate some potential respondents. Random sampling error (due exclusively to random, chance fluctuation) may cause an imbalance in the representativeness of the group. Additional errors will occur if individuals refuse to be interviewed or cannot be contacted. Such nonresponse error may also cause the sample to be less than perfectly representative. Thus, the actual sample is drawn from a population different from (or smaller than) the ideal.

Probability versus Nonprobability Sampling

Several alternative ways to take a sample are available. The main alternative sampling plans may be grouped into two categories: probability techniques and nonprobability techniques.

In **probability sampling**, every element in the population has a *known, nonzero probability* of selection. The simple random sample, in which each member of the population has an equal probability of being selected, is the best-known probability sample.

In **nonprobability sampling**, the probability of any particular member of the population being chosen is unknown. The selection of sampling units in nonprobability sampling is quite arbitrary, as researchers rely heavily on personal judgment. Technically, no appropriate statistical techniques exist for measuring random sampling error from a nonprobability sample. Therefore, projecting the data beyond the sample is, technically speaking, statistically inappropriate. Nevertheless, as the Research Snapshot on prescription drug costs shows, researchers sometimes find nonprobability samples best suited for a specific researcher purpose. As a result, nonprobability samples are pragmatic and are used in market research.

Nonprobability Sampling

Although probability sampling is preferred, we will discuss nonprobability sampling first to illustrate some potential sources of error and other weaknesses in sampling.

Convenience Sampling

As the name suggests, **convenience sampling** refers to sampling by obtaining people or units that are conveniently available. A research team may determine that the most convenient and economical method is to set up an interviewing booth from which to intercept consumers at a shopping center. Just before elections, television stations often present person-on-the-street interviews that are presumed to reflect public opinion. (Of course, the television station generally warns that the survey was “unscientific and random” [sic].) The college professor who uses his or her students has a captive sample—convenient, but perhaps not so representative.

Researchers generally use convenience samples to obtain a large number of completed questionnaires quickly and economically, or when obtaining a sample through other means is impractical. For example, many Internet surveys are conducted with volunteer respondents who, either intentionally or by happenstance, visit an organization’s Web site. Although this method produces a large number of responses quickly and at a low cost, selecting all visitors to a Web site is clearly convenience sampling. Respondents may not be representative because of the haphazard manner by which many of them arrived at the Web site or because of self-selection bias.

Similarly, research looking for cross-cultural differences in organizational or consumer behavior typically uses convenience samples. Rather than selecting cultures with characteristics relevant to the hypothesis being tested, the researchers conducting these studies often choose cultures to which they have access (for example, because they speak the language or have contacts in that culture’s organizations). Further adding to the convenience, cross-cultural research often defines “culture” in terms of nations, which are easier to identify and obtain statistics for, even though many nations include several cultures and some people in a given nation may be more involved with the international business or academic community than with a particular ethnic culture.⁴ Here again, the use of convenience sampling limits how well the research represents the intended population.

The user of research based on a convenience sample should remember that projecting the results beyond the specific sample is inappropriate. Convenience samples are best used for exploratory research when additional research will subsequently be conducted with a probability sample.

Judgment Sampling

Judgment (purposive) sampling is a nonprobability sampling technique in which an experienced individual selects the sample based on his or her judgment about some appropriate characteristics required of the sample member. Researchers select samples that satisfy their specific purposes, even if they are not fully representative. The consumer price index (CPI) is based on a judgment sample of market-basket items, housing costs, and other selected goods and services expected to reflect a representative sample of items consumed by most Americans. Test-market cities often are selected because they are viewed as typical cities whose demographic profiles closely match the national profile. A fashion manufacturer regularly selects a sample of key accounts that it believes are capable of providing information needed to predict what may sell in the fall. Thus, the sample is selected to achieve this specific objective.

Judgment sampling often is used in attempts to forecast election results. People frequently wonder how a television network can predict the results of an election with only 2 percent of the votes reported. Political and sampling experts judge which small voting districts approximate overall state returns from previous election years; then these *bellwether precincts* are selected as the sampling units. Of course, the assumption is that the past voting records of these districts are still representative of the political behavior of the state's population.

Quota Sampling

Suppose a firm wishes to investigate consumers who currently subscribe to an HDTV (high definition television) service. The researchers may wish to ensure that each brand of HDTV televisions is included proportionately in the sample. Strict probability sampling procedures would likely underrepresent certain brands and overrepresent other brands. If the selection process were left strictly to chance, some variation would be expected.

As seen in the Research Snapshot above, the purpose of **quota sampling** is to ensure that the various subgroups in a population are represented on pertinent sample characteristics to the exact extent that the investigators desire. Stratified sampling, a probability sampling procedure described in the next section, also has this objective, but it should not be confused with quota sampling. In quota sampling, the interviewer has a quota to achieve. For example, an interviewer in a particular city may be assigned 100 interviews, 35 with owners of Sony TVs, 30 with owners of Samsung TVs, 18 with owners of Panasonic TVs, and the rest with owners of other brands. The interviewer is responsible for finding enough people to meet the quota. Aggregating the various interview quotas yields a sample that represents the desired proportion of each subgroup.

■ POSSIBLE SOURCES OF BIAS

The logic of classifying the population by pertinent subgroups is essentially sound. However, because respondents are selected according to a convenience sampling procedure rather than on a probability basis (as in stratified sampling), the haphazard selection of subjects may introduce bias. For example, a college professor hired some of his students to conduct a quota sample based on age. When analyzing the data, the professor discovered that almost all the people in the "under 25 years" category were college-educated. Interviewers, being human, tend to prefer to interview people who are similar to themselves.

Quota samples tend to include people who are easily found, willing to be interviewed, and middle class. Fieldworkers are given considerable leeway to exercise their judgment concerning selection of actual respondents. Interviewers often concentrate their interviewing in areas with heavy pedestrian traffic such as downtowns, shopping malls, and college campuses. Those who interview door-to-door learn quickly that quota requirements are difficult to meet by interviewing whoever happens to appear at the door. People who are more likely to stay at home generally share a less active lifestyle and are less likely to be meaningfully employed. One interviewer related a story of working in an upper-middle-class neighborhood. After a few blocks, he arrived in a

neighborhood of mansions. Feeling that most of the would-be respondents were above his station, the interviewer skipped these houses because he felt uncomfortable knocking on doors that would be answered by these people or their hired help.

■ ADVANTAGES OF QUOTA SAMPLING

The major advantages of quota sampling over probability sampling are speed of data collection, lower costs, and convenience. Although quota sampling has many problems, carefully supervised data collection may provide a representative sample of the various subgroups within a population. Quota sampling may be appropriate when the researcher knows that a certain demographic group is more likely to refuse to cooperate with a survey. For instance, if older men are more likely to refuse, a higher quota can be set for this group so that the proportion of each demographic category will be similar to the proportions in the population. A number of laboratory experiments also rely on quota sampling because it is difficult to find a sample of the general population willing to visit a laboratory to participate in an experiment.

Snowball Sampling

A variety of procedures known as **snowball sampling** involve using probability methods for an initial selection of respondents and then obtaining additional respondents through information provided by the initial respondents. This technique is used to locate members of rare populations by referrals. Suppose a manufacturer of sports equipment is considering marketing a mahogany croquet set for serious adult players. This market is certainly small. An extremely large sample would be necessary to find 100 serious adult croquet players. It would be much more economical to survey, say, 300 people, find 15 croquet players, and ask them for the names of other players.

Reduced sample sizes and costs are clear-cut advantages of snowball sampling. However, bias is likely to enter into the study because a person suggested by someone also in the sample has a higher probability of being similar to the first person. If there are major differences between those who are widely known by others and those who are not, this technique may present some serious problems. However, snowball sampling may be used to locate and recruit heavy users, such as consumers who buy more than 50 compact discs per year, for focus groups. As the focus group is not expected to be a generalized sample, snowball sampling may be appropriate.

Probability Sampling

All probability sampling techniques are based on chance selection procedures. Because the probability sampling process is random, the bias inherent in nonprobability sampling procedures is eliminated. Note that the term *random* refers to the procedure for selecting the sample; it does not describe the data in the sample. *Randomness* characterizes a procedure whose outcome cannot be predicted because it depends on chance. Randomness should not be thought of as unplanned or unscientific—it is the basis of all probability sampling techniques. This section will examine the various probability sampling methods.

Simple Random Sampling

The sampling procedure that ensures each element in the population will have an equal chance of being included in the sample is called **simple random sampling**. Examples include drawing names from a hat and selecting the winning raffle ticket from a large drum. If the names or raffle tickets are thoroughly stirred, each person or ticket should have an equal chance of being selected. In contrast to other, more complex types of probability sampling, this process is simple because it requires only one stage of sample selection.

Although drawing names or numbers out of a fishbowl, using a spinner, rolling dice, or turning a roulette wheel may be an appropriate way to draw a sample from a small population, when populations consist of large numbers of elements, sample selection is based on tables of random numbers or computer-generated random numbers

Suppose a researcher is interested in selecting a simple random sample of all the Honda dealers in California, New Mexico, Arizona, and Nevada. Each dealer's name is assigned a number from 1 to 105. The numbers can be written on paper slips, and all the slips can be placed in a bowl. After the slips of paper have been thoroughly mixed, one is selected for each sampling unit. Thus, if the sample size is 35, the selection procedure must be repeated 34 times after the first slip has been selected. Mixing the slips after each selection will ensure that those at the bottom of the bowl will continue to have an equal chance of being selected in the sample.

The random-digit dialing technique of sample selection requires that the researcher identify the exchange or exchanges of interest (the first three numbers) and then use a table of numbers to select the next four numbers. In practice, the exchanges are not always selected randomly. Researchers who wanted to find out whether Americans of African descent prefer being called "black" or "African-American" narrowed their sampling frame by selecting exchanges associated with geographic areas where the proportion of the population (African-Americans/blacks) was at least 30 percent. The reasoning was that this made the survey procedure far more efficient, considering that the researchers were trying to contact a group representing less than 15 percent of U.S. households. This initial judgment sampling raises the same issues we discussed regarding nonprobability sampling. In this study, the researchers found that respondents were most likely to prefer the term *black* if they had attended schools that were about half black and half white. If such experiences influence the answers to the question of interest to the researchers, the fact that blacks who live in predominantly white communities are underrepresented may introduce bias into the results.

Systematic Sampling

Suppose a researcher wants to take a sample of 1,000 from a list of 200,000 names. With **systematic sampling**, every 200th name from the list would be drawn. The procedure is extremely simple. A starting point is selected by a random process; then every *n*th number on the list is selected. To take a sample of consumers from a rural telephone directory that does not separate business from residential listings, every 23rd name might be selected as the *sampling interval*. In the process, Mike's Restaurant might be selected. This unit is inappropriate because it is a business listing rather than a consumer listing, so the next eligible name would be selected as the sampling unit, and the systematic process would continue.

While systematic sampling is not actually a random selection procedure, it does yield random results if the arrangement of the items in the list is random in character. The problem of *periodicity* occurs if a list has a systematic pattern—that is, if it is not random in character. Collecting retail sales information every seventh day would result in a distorted sample because there would be a systematic pattern of selecting sampling units—sales for only one day of the week (perhaps Monday) would be sampled. If the first 50 names on a list of contributors to a charity were extremely large donors, periodicity bias might occur in sampling every 200th name. Periodicity is rarely a problem for most sampling in marketing research, but researchers should be aware of the possibility.

Stratified Sampling

The usefulness of dividing the population into subgroups, or *strata*, whose members are more or less equal with respect to some characteristic was illustrated in our discussion of quota sampling. The first step is the same for both stratified and quota sampling: choosing strata on the basis of existing information—for example, classifying retail outlets based on annual sales volume. However, the process of selecting sampling units within the strata differs substantially. In **stratified sampling**, a subsample is drawn using simple random sampling within each stratum. This is not true of quota sampling.

The reason for taking a stratified sample is to obtain a more efficient sample than would be possible with simple random sampling. Suppose, for example, that urban and rural groups have widely

different attitudes toward energy conservation, but members within each group hold very similar attitudes. Random sampling error will be reduced with the use of stratified sampling, because each group is internally homogeneous but there are comparative differences between groups. More technically, a smaller standard error may result from this stratified sampling because the groups will be adequately represented when strata are combined.

Another reason for selecting a stratified sample is to ensure that the sample will accurately reflect the population on the basis of the criterion or criteria used for stratification. This is a concern because occasionally simple random sampling yields a disproportionate number of one group or another and the sample ends up being less representative than it could be.

A researcher can select a stratified sample as follows. First, a variable (sometimes several variables) is identified as an efficient basis for stratification. A stratification variable must be a characteristic of the population elements known to be related to the dependent variable or other variables of interest. The variable chosen should increase homogeneity within each stratum and increase heterogeneity between strata. The stratification variable usually is a categorical variable or one easily converted into categories (that is, subgroups). For example, a pharmaceutical company interested in measuring how often physicians prescribe a certain drug might choose physicians' training as a basis for stratification. In this example the mutually exclusive strata are MDs (medical doctors) and ODs (osteopathic doctors).

Proportional versus Disproportional Sampling

If the number of sampling units drawn from each stratum is in proportion to the relative population size of the stratum, the sample is a **proportional stratified sample**. Sometimes, however, a disproportional stratified sample will be selected to ensure an adequate number of sampling units in every stratum. Sampling more heavily in a given stratum than its relative population size warrants is not a problem if the primary purpose of the research is to estimate some characteristic separately for each stratum and if researchers are concerned about assessing the differences among strata. Consider, however, the percentages of retail outlets presented in Exhibit 16.5. A proportional sample would have the same percentages as in the population. Although there is a small percentage of warehouse club stores, the average dollar sales volume for the warehouse club store stratum is quite large and varies substantially from the average store size for the smaller independent stores. To avoid overrepresenting the chain stores and independent stores (with smaller sales volume) in the sample, a disproportional sample is taken.

In a **disproportional stratified sample** the sample size for each stratum is not allocated in proportion to the population size but is dictated by analytical considerations, such as variability in store sales volume. The logic behind this procedure relates to the general argument for sample size: As variability increases, sample size must increase to provide accurate estimates. Thus, the strata that exhibit the greatest variability are sampled more heavily to increase sample efficiency—that is, produce smaller random sampling error. Complex formulas (beyond the scope of an introductory course in business research) have been developed to determine sample size for each stratum. A simplified rule of thumb for understanding the concept of optimal allocation is that the stratum sample size increases for strata of larger sizes with the greatest relative variability. Other complexities arise in determining population estimates. For example, when disproportional stratified sampling is used, the estimated mean for each stratum has to be weighed according to the number of elements in each stratum in order to calculate the total population mean.

Cluster Sampling

The purpose of **cluster sampling** is to sample economically while retaining the characteristics of a probability sample. Consider a researcher who must conduct five hundred personal interviews with consumers scattered throughout the United States. Travel costs are likely to be enormous because the amount of time spent traveling will be substantially greater than the time spent in the interviewing

process. If an aspirin marketer can assume the product will be equally successful in Phoenix and Baltimore, or if a frozen pizza manufacturer assumes its product will suit the tastes of Texans equally as well as Oregonians, cluster sampling may be used to represent the United States.

In a cluster sample, the primary sampling unit is no longer the individual element in the population (for example, grocery stores) but a larger cluster of elements located in proximity to one another (for example, cities). The *area sample* is the most popular type of cluster sample. A grocery store researcher, for example, may randomly choose several geographic areas as primary sampling units and then interview all or a sample of grocery stores within the geographic clusters. Interviews are confined to these clusters only. No interviews occur in other clusters. Cluster sampling is classified as a probability sampling technique because of either the random selection of clusters or the random selection of elements within each cluster. Some examples of clusters appear in Exhibit 16.6 on the next page.

Cluster samples frequently are used when lists of the sample population are not available. For example, when researchers investigating employees and self-employed workers for a downtown revitalization project found that a comprehensive list of these people was not available, they decided to take a cluster sample, selecting organizations (business and government) as the clusters. A sample of firms within the central business district was developed, using stratified probability sampling to identify clusters. Next, individual workers within the firms (clusters) were randomly selected and interviewed concerning the central business district.

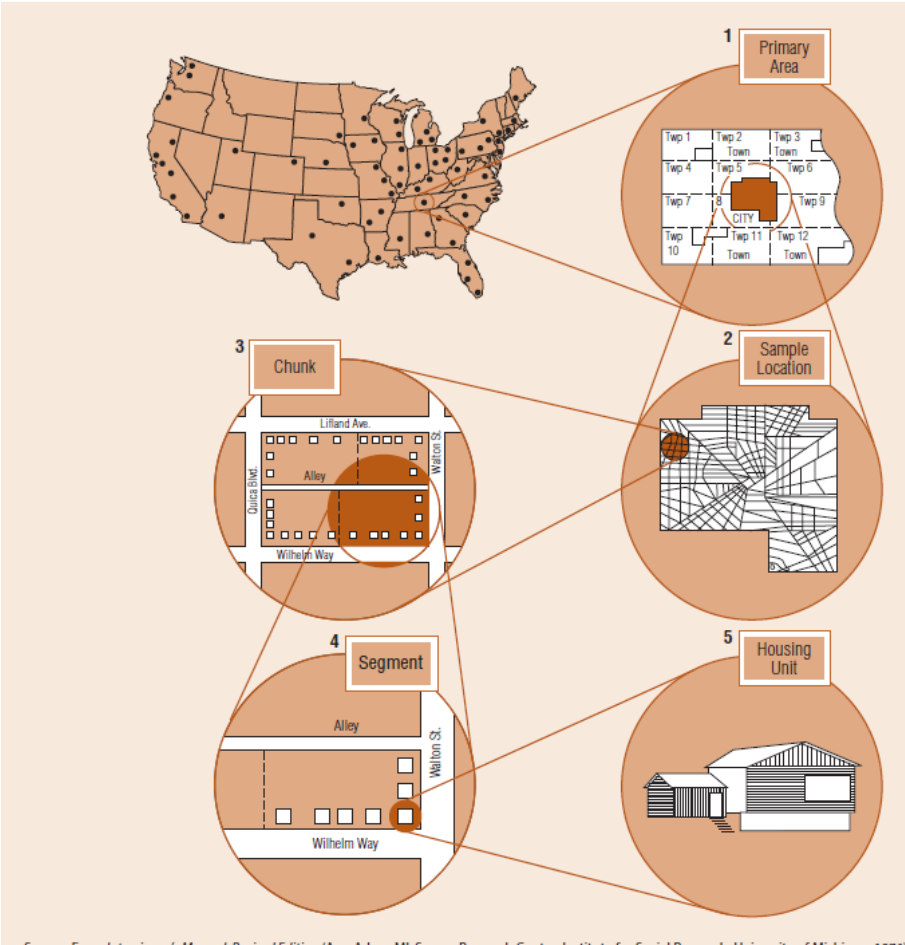
Ideally a cluster should be as heterogeneous as the population itself—a mirror image of the population. A problem may arise with cluster sampling if the characteristics and attitudes of the elements within the cluster are too similar. For example, geographic neighborhoods tend to have residents of the same socioeconomic status. Students at a university tend to share similar beliefs. This problem may be mitigated by constructing clusters composed of diverse elements and by selecting a large number of sampled clusters.

Multistage Area Sampling

Multistage area sampling involves two or more steps that combine some of the probability techniques already described. Typically, geographic areas are randomly selected in progressively smaller (lower-population) units. For example, a political pollster investigating an election in Arizona might first choose counties within the state to ensure that the different areas are represented in the sample. In the second step, precincts within the selected counties may be chosen. As a final step, the pollster may select blocks (or households) within the precincts, then interview all the blocks (or households) within the geographic area. Researchers may take as many steps as necessary to achieve a representative sample. Exhibit 16.7 graphically portrays a multistage area sampling process frequently used by a major academic research center. Progressively smaller geographic areas are chosen until a single housing unit is selected for interviewing.

The Bureau of the Census provides maps, population information, demographic characteristics for population statistics, and so on, by several small geographical areas; these may be useful in sampling. Census classifications of small geographical areas vary, depending on the extent of urbanization within Metropolitan Statistical Areas (MSAs) or counties.

EXHIBIT 16.7 Illustration of Multistage Area Sampling



What Is the Appropriate Sample Design?

A researcher who must decide on the most appropriate sample design for a specific project will identify a number of sampling criteria and evaluate the relative importance of each criterion before selecting a sampling design. This section outlines and briefly discusses the most common criteria. Exhibit 16.9 summarizes the advantages and disadvantages of each nonprobability sampling technique, and Exhibit 16.10 does the same for the probability sampling techniques.

EXHIBIT 16.9 Comparison of Sampling Techniques: Nonprobability Samples

Nonprobability Samples			
Description	Cost and Degree of Use	Advantages	Disadvantages
1. <i>Convenience</i> : The researcher uses the most convenient sample or economical sample units.	Very low cost, extensively used	No need for list of population	Unrepresentative samples likely; random sampling error estimates cannot be made; projecting data beyond sample is relatively risky
2. <i>Judgment</i> : An expert or experienced researcher selects the sample to fulfill a purpose, such as ensuring that all members have a certain characteristic.	Moderate cost, average use	Useful for certain types of forecasting; sample guaranteed to meet a specific objective	Bias due to expert's beliefs may make sample unrepresentative; projecting data beyond sample is risky
3. <i>Quota</i> : The researcher classifies the population by pertinent properties, determines the desired proportion to sample from each class, and fixes quotas for each interviewer.	Moderate cost, very extensively used	Introduces some stratification of population; requires no list of population	Introduces bias in researcher's classification of subjects; nonrandom selection within classes means error from population cannot be estimated; projecting data beyond sample is risky
4. <i>Snowball</i> : Initial respondents are selected by probability samples; additional respondents are obtained by referral from initial respondents.	Low cost, used in special situations	Useful in locating members of rare populations	High bias because sample units are not independent; projecting data beyond sample is risky

EXHIBIT 16.10 Comparison of Sampling Techniques: Probability Samples

Probability Samples			
Description	Cost and Degree of Use	Advantages	Disadvantages
1. <i>Simple random</i> : The researcher assigns each member of the sampling frame a number, then selects sample units by random method.	High cost, moderately used in practice (most common in random digit dialing and with computerized sampling frames)	Only minimal advance knowledge of population needed; easy to analyze data and compute error	Requires sampling frame to work from; does not use knowledge of population that researcher may have; larger errors for same sampling size than in stratified sampling; respondents may be widely dispersed, hence cost may be higher
2. <i>Systematic</i> : The researcher uses natural ordering or the order of the sampling frame, selects an arbitrary starting point, then selects items at a preselected interval.	Moderate cost, moderately used	Simple to draw sample; easy to check	If sampling interval is related to periodic ordering of the population, may introduce increased variability
3. <i>Stratified</i> : The researcher divides the population into groups and randomly selects subsamples from each group. Variations include proportional, disproportional, and optimal allocation of subsample sizes.	High cost, moderately used	Ensures representation of all groups in sample; characteristics of each stratum can be estimated and comparisons made; reduces variability for same sample size	Requires accurate information on proportion in each stratum; if stratified lists are not already available, they can be costly to prepare
4. <i>Cluster</i> : The researcher selects sampling units at random, then does a complete observation of all units or draws a probability sample in the group.	Low cost, frequently used	If clusters geographically defined, yields lowest field cost; requires listing of all clusters, but of individuals only within clusters; can estimate characteristics of clusters as well as of population	Larger error for comparable size than with other probability samples; researcher must be able to assign population members to unique cluster or else duplication or omission of individuals will result
5. <i>Multistage</i> : Progressively smaller areas are selected in each stage by some combination of the first four techniques.	High cost, frequently used, especially in nationwide surveys	Depends on techniques combined	Depends on techniques combined

Degree of Accuracy

Selecting a representative sample is important to all researchers. However, the degree of accuracy required or the researcher's tolerance for sampling and nonsampling error may vary from project to project, especially when cost savings or another benefit may be a trade-off for a reduction in accuracy.

Resources

The cost associated with the different sampling techniques varies tremendously. If the researcher's financial and human resources are restricted, certain options will have to be eliminated. For a graduate student working on a master's thesis, conducting a national survey is almost always out of the question because of limited resources. Managers concerned with the cost of the research versus the value of the information often will opt to save money by using a nonprobability sampling design rather than make the decision to conduct no research at all.

Time

A researcher who needs to meet a deadline or complete a project quickly will be more likely to select a simple, less time-consuming sample design.

Advance Knowledge of the Population

Advance knowledge of population characteristics, such as the availability of lists of population members, is an important criterion. In many cases, however, no list of population elements will be available to the researcher. This is especially true when the population element is defined by ownership of a particular product or brand, by experience in performing a specific job task, or on a qualitative dimension. A lack of adequate lists may automatically rule out systematic sampling, stratified sampling, or other sampling designs, or it may dictate that a preliminary study, such as a short telephone survey using random digit dialing, be conducted to generate information to build a

sampling frame for the primary study. In many developing countries, things like reverse directories are rare. Thus, researchers planning sample designs have to work around this limitation.

National versus Local Project

Geographic proximity of population elements will influence sample design. When population elements are unequally distributed geographically, a cluster sample may become much more attractive.

Internet Sampling Is Unique

Internet surveys allow researchers to reach a large sample rapidly—both an advantage and a disadvantage. Sample size requirements can be met overnight or in some cases almost instantaneously. A researcher can, for instance, release a survey during the morning in the Eastern Standard Time zone and have all sample size requirements met before anyone on the West Coast wakes up. If rapid response rates are expected, the sample for an Internet survey should be metered out across all time zones. In addition, people in some populations are more likely to go online during weekend than on a weekday. If the researcher can anticipate a day-of-the-week effect, the survey should be kept open long enough so that all sample units have the opportunity to participate in the research project.

The ease and low cost of an Internet survey also has contributed to a flood of online questionnaires, some more formal than others. As a result, frequent Internet users may be more selective about which surveys they bother answering. Researchers investigating college students' attitudes toward environmental issues found that those who responded to an e-mail request that had been sent to all students tended to be more concerned about the environment than students who were contacted individually through systematic sampling. The researchers concluded that students who cared about the issues were more likely to respond to the online survey.

Another disadvantage of Internet surveys is the lack of computer ownership and Internet access among certain segments of the population. A sample of Internet users is representative only of Internet users, who tend to be younger, better educated, and more affluent than the general population. This is not to say that all Internet samples are unrepresentative of all target populations. Nevertheless, when using Internet surveys, researchers should be keenly aware of potential sampling problems that can arise due to systematic characteristics of heavy computer users.

Web Site Visitors

As noted earlier, many Internet surveys are conducted with volunteer respondents who visit an organization's Web site intentionally or by happenstance. These *unrestricted samples* are clearly convenience samples. They may not be representative because of the haphazard manner by which many respondents arrived at a particular Web site or because of self-selection bias.

A better technique for sampling Web site visitors is to randomly select sampling units. Survey- Site, a company that specializes in conducting Internet surveys, collects data by using its "pop-up survey" software. The software selects Web visitors at random and "pops up" a small JavaScript window asking the person if he or she wants to participate in an evaluation survey. If the person clicks yes, a new window containing the online survey opens up. The person can then browse the site at his or her own pace and switch to the survey at any time to express an opinion.⁷

Randomly selecting Web site visitors can cause a problem. It is possible to overrepresent frequent visitors to the site and thus represent site visits rather than visitors. Several programming techniques and technologies (using cookies, registration data, or prescreening) are available to help accomplish more representative sampling based on site traffic.⁸ Details of these techniques are beyond the scope of this discussion.

This type of random sampling is most valuable if the target population is defined as visitors to a particular Web site. Evaluation and analysis of visitors' perceptions and experiences of the Web site

would be a typical survey objective with this type of sample. Researchers who have broader interests may obtain Internet samples in a variety of other ways.

Panel Samples

Drawing a probability sample from an established consumer panel or other prerecruited membership panel is a popular, scientific, and effective method for creating a sample of Internet users. Typically, sampling from a panel yields a high response rate because panel members have already agreed to cooperate with the research organization's e-mail or Internet surveys. Often panel members are compensated for their time with a sweepstakes, a small cash incentive, or redeemable points. Further, because the panel has already supplied demographic characteristics and other information from previous questionnaires, researchers are able to select panelists based on product ownership, lifestyle, or other characteristics. As seen in the Research Snapshots on the Current Population Survey and student adjustment, a variety of sampling methods and data transformation techniques can be applied to ensure that sample results are representative of the general public or a targeted population.

Recruited Ad Hoc Samples

Another means of obtaining an Internet sample is to obtain or create a sampling frame of e-mail addresses on an *ad hoc* basis. Researchers may create the sampling frame offline or online. Databases containing e-mail addresses can be compiled from many sources, including customer/client lists, advertising banners on pop-up windows that recruit survey participants, online sweepstakes, and registration forms that must be filled out in order to gain access to a particular Web site. Researchers may contact respondents by "snail mail" or by telephone to ask for their e-mail addresses and obtain permission for an Internet survey. Using offline techniques, such as random-digit dialing and short telephone screening interviews, to recruit respondents can be a very practical way to get a representative sample for an Internet survey. Companies anticipating future Internet research can develop a valuable database for sample recruitment by including e-mail addresses in their customer relationship databases (by inviting customers to provide that information on product registration cards, in telephone interactions, through on-site registration, etc.).

DETERMINATION OF SAMPLE SIZE: A REVIEW OF STATISTICAL THEORY

Descriptive and Inferential Statistics

The *Statistical Abstract of the United States* presents table after table of figures associated with numbers of births, number of employees in each county of the United States, and other data that the average person calls "statistics." Technically, these are **descriptive statistics**, which describe basic characteristics and summarize the data in a straightforward and understandable manner. Another type of statistics, **inferential statistics**, is used to make inferences or to project from a sample to an entire population. For example, when a firm test-markets a new product in Peoria and Fort Worth, it is not only concerned about how customers in these two cities feel, but they want to make an inference from these sample markets to predict what will happen throughout the United States. So, two applications of statistics exist: (1) descriptive statistics which describe characteristics of the population or sample and (2) inferential statistics which are used to generalize from a sample to a population.

Sample Statistics and Population Parameters

A sample is a subset or relatively small portion of the total number of elements in a given population. **Sample statistics** are measures computed from sample data. Since business researchers typically deal with samples—we rarely talk to every consumer, manager, or organization—we normally base our

decisions off of sample data. The primary purpose of inferential statistics is to make a judgment about a population, or the total collection of all elements about which a researcher seeks information, based from a subset of that population.

Population parameters are measured characteristics of a specific population. In other words, information about the entire universe of interest. Sample statistics are used to make inferences (guesses) about population parameters based on sample data.² In our notation, we will generally represent population parameters with Greek lowercase letters—for example, μ or σ —and sample statistics with English letters, such as \bar{X} or S .

Making Data Usable

Suppose a telephone survey has been conducted for a savings and loan association. The data have been recorded on a large number of questionnaires. To make the data usable, this information must be organized and summarized. Methods for doing this include **frequency distributions, proportions, measures of central tendency, and measures of dispersion.**

Frequency Distributions

One of the most common ways to summarize a set of data is to construct a *frequency table*, or **frequency distribution**. The process begins with recording the number of times a particular value of a variable occurs. This is the frequency of that value. Using an example of a telephone survey for a savings and loan association, Exhibit 17.1 on the next page represents a frequency distribution of respondents' answers to a question that asked how much money customers had deposited in the institution. In this case, we can see that more respondents (811) checked the highest box of \$12,000 or more.

EXHIBIT 17.1
Frequency Distribution
of Deposits

Amount	Frequency (Number of People Who Hold Deposits in Each Range)
Under \$3,000	499
\$3,000–\$5,999	530
\$6,000–\$8,999	562
\$9,000–\$11,999	718
\$12,000 or more	811
	3,120

EXHIBIT 17.2
Percentage Distribution
of Deposits

Amount	Percent (Percentage of People Who Hold Amount Deposits in Each Range)
Under \$3,000	16%
\$3,000–\$5,999	17%
\$6,000–\$8,999	18%
\$9,000–\$11,999	23%
\$12,000 or more	26%
	100%

A similar method of describing the data is to construct a distribution of relative frequency, or a **percentage distribution**. To develop a frequency distribution of percentages, divide the frequency of **Probability** is the long-run relative frequency with which an event will occur. Inferential statistics uses the concept of a probability distribution, which is conceptually the same as a percentage distribution except that the data are converted into probabilities. Exhibit 17.3 shows the probability

distribution of the savings and loan deposits. We know that the probability of a respondent falling into the top category of \$12,000 or more is the highest, 0.26.

Amount	Probability
Under \$3,000	0.16
\$3,000–\$5,999	0.17
\$6,000–\$8,999	0.18
\$9,000–\$11,999	0.23
\$12,000 or more	0.26
	<u>1.00</u>

EXHIBIT 17.3
Probability Distribution
of Deposits

Proportions

When a frequency distribution portrays only a single characteristic in terms of a percentage of the total, it defines the proportion of occurrence. A proportion, such as the proportion of CPAs at an accounting firm, indicates the percentage of population elements that successfully meet some standard concerning the particular characteristic. A proportion may be expressed as a percentage (25%), a fraction (1/4), or a decimal value (0.25).

Measures of Central Tendency

On a typical day, a sales manager counts the number of sales calls each sales representative makes. She may want to inspect the data to find the average, center, or middle area, of the frequency distribution. Central tendency can be measured in three ways—the mean, median, or mode—each of which has a different meaning.

■ THE MEAN

We all have been exposed to the average known as the **mean**. The mean is simply the arithmetic average, and it is perhaps the most common measure of central tendency. More likely than not, you already know how to calculate a mean. However, knowing how to distinguish among the symbols Σ , μ , and X is helpful to understand statistics.

To express the mean mathematically, we use the summation symbol, the capital Greek letter *sigma* (Σ). A typical use might look like this:

$$\sum_{i=1}^n X_i$$

which is a shorthand way to write the sum

$$X_1 + X_2 + X_3 + X_4 + X_5 + \cdots + X_n$$

Below the Σ is the initial value of an index, usually, i , j , or k , and above it is the final value, in this case n , the number of observations. The shorthand expression says to replace i in the formula with the values from 1 to 8 and total the observations obtained. Without changing the basic formula, the initial and final index values may be replaced by other values to indicate different starting and stopping points.

Suppose our sales manager supervises the eight salespeople listed in Exhibit 17.4. To express the sum of the salespeople's calls in Σ notation, we just number the salespeople (this number becomes the

index number) and associate subscripted variables with their numbers of calls: the initial and final index values may be replaced by other values to indicate different starting and stopping points. Suppose our sales manager supervises the eight salespeople listed in Exhibit 17.4. To express the sum of the salespeople's calls in Σ notation, we just number the salespeople (this number becomes the index number) and associate subscripted variables with their numbers of calls:

Index		Salesperson	Variable		Number of Calls
1	=	Mike	X_1	=	4
2	=	Patty	X_2	=	3
3	=	Billie	X_3	=	2
4	=	Bob	X_4	=	5
5	=	John	X_5	=	3
6	=	Frank	X_6	=	3
7	=	Chuck	X_7	=	1
8	=	Samantha	X_8	=	5

We then write an appropriate Σ formula and evaluate it:

$$\begin{aligned}
 \sum_{i=1}^8 X_i &= X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 \\
 &= 4 + 3 + 2 + 5 + 3 + 3 + 1 + 5 \\
 &= 26
 \end{aligned}$$

This notation is the numerator in the formula for the arithmetic mean:

$$\text{Mean} = \frac{\sum_{i=1}^n X_i}{n} = \frac{26}{8} = 3.25$$

The sum $\sum_{i=1}^n X_i$ tells us to add all the X s whose subscripts are between 1 and n inclusive, where n equals the number of observations. The formula shows that the mean number of sales calls in this example is 3.25.

Researchers generally wish to know the population mean, μ (lowercase Greek letter *mu*), which is calculated as follows:

$$\mu = \frac{\sum_{i=1}^n X_i}{N}$$

where

N = number of all observations in the population

Often we will not have the data to calculate the population mean, μ , so we will calculate a sample mean, \bar{X} (read "X bar"), with the following formula:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

where

n = number of observations made in the sample

In this introductory discussion of the summation sign (Σ), we have used very detailed notation that includes the subscript for the initial index value (i) and the final index value (n). However, from this point on, references to Σ will sometimes omit the subscript for the initial index value (i) and the final index value (n).

THE MEDIAN

The next measure of central tendency, the median, is the midpoint of the distribution, or the 50th percentile. In other words, the median is the value below which half the values in the sample fall, and above which half of the values fall. In the sales manager example, 3 is the median because half the observations are greater than 3 and half are less than 3.

THE MODE

In apparel, mode refers to the most popular fashion. In statistics the mode is the measure of central tendency that identifies the value that occurs most often. In our example of sales calls, Patty, John, and Frank each made three sales calls. The value 3 occurs most often, so 3 is the mode. The mode is determined by listing each possible value and noting the number of times each value occurs

Measures of Dispersion

The mean, median, and mode summarize the central tendency of frequency distributions. Accurate analysis of data also requires knowing the tendency of observations to depart from the central tendency. What is the spread across the observations? Thus, another way to summarize the data is to calculate the dispersion of the data, or how the observations vary from the mean. Consider, for instance, the 12-month sales patterns of the two products shown in Exhibit 17.5. Both have a mean monthly sales volume of 200 units, as well as a median and mode of 200, but the dispersion of observations for product B is much greater than that for product A. There are several measures of dispersion.

THE RANGE

The simplest measure of dispersion is the range. It is the distance between the smallest and the largest values of a frequency distribution. In Exhibit 17.5, the range for product A is between 196 units and 202 units (6 units), whereas for product B the range is between 150 units and 261 units (111 units). The range does not take into account all the observations; it merely tells us about the extreme values of the distribution.

EXHIBIT 17.5
Sales Levels for Two
Products with Identical
Average Sales

	Units Product A	Units Product B
January	196	150
February	198	160
March	199	175
April	200	181
May	200	192
June	200	200
July	200	200
August	201	202
September	201	213
October	201	224
November	202	240
December	202	261
Average	200	200

■ WHY USE THE STANDARD DEVIATION?

Statisticians have derived several quantitative indexes to reflect a distribution's spread, or variability. The *standard deviation* is perhaps the most valuable index of spread, or dispersion. Students often have difficulty understanding it. Learning about the standard deviation will be easier if we first look at several other measures of dispersion that may be used. Each of these has certain limitations that the standard deviation does not.

First is the *deviation*. Deviation is a method of calculating how far any observation is from the mean. To calculate a deviation from the mean, use the following formula:

$$d_{i_i} = X_i - \bar{X}$$

For the value of 150 units for product B for the month of January, the deviation score is -50; that is, $150 - 200 = -50$. If the deviation scores are large, we will have a fat distribution because the distribution exhibits a broad spread.

Next is the *average deviation*. We compute the average deviation by calculating the deviation score of each observation value (that is, its difference from the mean), summing these scores, and then dividing by the sample size (n):

$$\text{Average deviation} = \frac{\sum(X_i - \bar{X})}{n}$$

While this measure of spread may seem initially interesting, it is never used. Positive deviation scores are canceled out by negative scores with this formula, leaving an average deviation value of zero no matter how wide the spread may be. Hence, the average deviation is a useless spread measure.

One might correct for the disadvantage of the average deviation by computing the absolute values of the deviations, termed *mean absolute deviation*. In other words, we ignore all the positive and negative signs and use only the absolute value of each deviation. The formula for the mean absolute deviation is

$$\text{Average deviation} = \frac{\sum(X_i - \bar{X})}{n}$$

While this procedure eliminates the problem of always having a zero score for the deviation measure, some technical mathematical problems make it less valuable than some other measures.

The *mean squared deviation* provides another method of eliminating the positive/negative sign problem. In this case, the deviation is squared, which eliminates the negative values. The mean squared deviation is calculated by the following formula

$$\text{Mean squared deviation} = \frac{\sum(X_i - \bar{X})^2}{n}$$

This measure is quite useful for describing the sample variability.

Variance

However, we typically wish to make an inference about a population from a sample, and so the divisor $n - 1$ is used rather than n in most pragmatic marketing research problems.³ This new measure of spread, called **variance**, has the following formula:

$$\text{Variance} = S^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}$$

Variance is a very good index of dispersion. The variance, S^2 , will equal zero if and only if each and every observation in the distribution is the same as the mean. The variance will grow larger as the observations tend to differ increasingly from one another and from the mean.

Standard Deviation

While the variance is frequently used in statistics, it has one major drawback. The variance reflects a unit of measurement that has been squared. For instance, if measures of sales in a territory are made in dollars, the mean number will be reflected in dollars, but the variance will be in squared dollars. Because of this, statisticians often take the square root of the variance. Using the square root of the variance for a distribution, called the **standard deviation**, eliminates the drawback of having the measure of dispersion in squared units rather than in the original measurement units. The formula for the standard deviation is

$$S = \sqrt{S^2} = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n - 1}}$$

Exhibit 17.7 illustrates that the calculation of a standard deviation requires the researcher to first calculate the sample mean. In the example with eight salespeople's sales calls (Exhibit 17.4), we calculated the sample mean as 3.25. Exhibit 17.7 illustrates how to calculate the standard deviation for these data

Calculating a Standard Deviation: Number of Sales Calls per Day for Eight Salespeople

X	$(X - \bar{X})$	$(X - \bar{X})^2$
4	$(4 - 3.25) = .75$.5625
3	$(3 - 3.25) = -.25$.0625
2	$(2 - 3.25) = -1.25$	1.5625
5	$(5 - 3.25) = 1.75$	3.0625
3	$(3 - 3.25) = -.25$.0625
3	$(3 - 3.25) = -.25$.0625
1	$(1 - 3.25) = -2.25$	5.0625
<u>5</u>	<u>$(5 - 3.25) = 1.75$</u>	<u>3.0625</u>
Σ^2	0	13.5000

$n = 8 \quad \bar{X} = 3.25$

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{13.5}{8 - 1}} = \sqrt{\frac{13.5}{7}} = \sqrt{1.9286} = 1.3887$$

The Normal Distribution

One of the most common probability distributions in statistics is the **normal distribution**, commonly represented by the *normal curve*. This mathematical and theoretical distribution describes the expected distribution of sample means and many other chance occurrences. The normal curve is bell shaped, and almost all (99 percent) of its values are within ± 3 standard deviations from its mean. An example of a normal curve, the distribution of IQ scores, appears in Exhibit 17.8 on the next page. In this example, 1 standard deviation for IQ equals 15. We can identify the proportion of the curve by measuring a score's distance (in this case, standard deviation) from the mean (100).

The **standardized normal distribution** is a specific normal curve that has several characteristics:

1. It is symmetrical about its mean; the tails on both sides are equal.
2. The mode identifies the normal curve's highest point, which is also the mean and median, and the vertical line about which this normal curve is symmetrical.
3. The normal curve has an infinite number of cases (it is a continuous distribution), and the area under the curve has a probability density equal to 1.0.
4. The standardized normal distribution has a mean of 0 and a standard deviation of 1.

Exhibit 17.9 on the next page illustrates these properties. Exhibit 17.10 on the next page is a summary version of the typical standardized normal table found at the end of most statistics textbooks. A more complex table of areas under the standardized normal distribution appears in Table A. 2 in the appendix.

The standardized normal distribution is a purely theoretical probability distribution, but it is the most useful distribution in inferential statistics.

Statisticians have spent a great deal of time and effort making it convenient for researchers to find the probability of any portion of the area under the standardized normal distribution. All we have to do is transform, or convert, the data from other observed normal distributions to the standardized normal curve. In other words, the standardized normal distribution is extremely valuable because we can translate, or transform, any normal variable, X , into the standardized value, Z .

EXHIBIT 17.8
Normal Distribution:
Distribution of Intelligence
Quotient (IQ) Scores

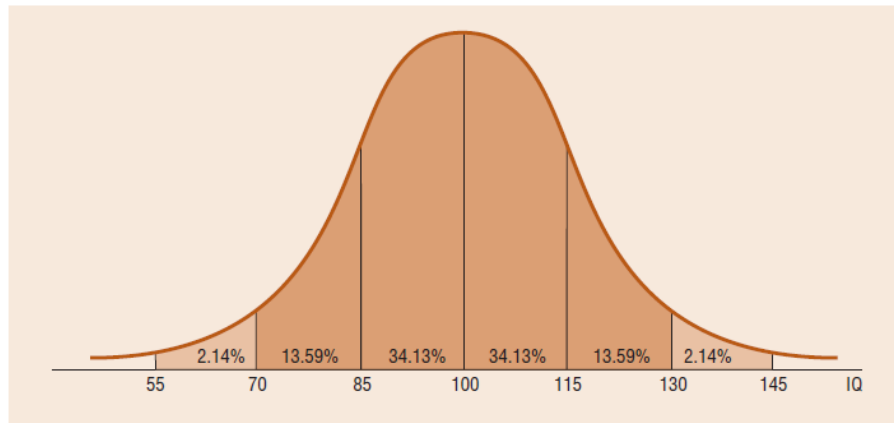


EXHIBIT 17.9
Standardized Normal
Distribution

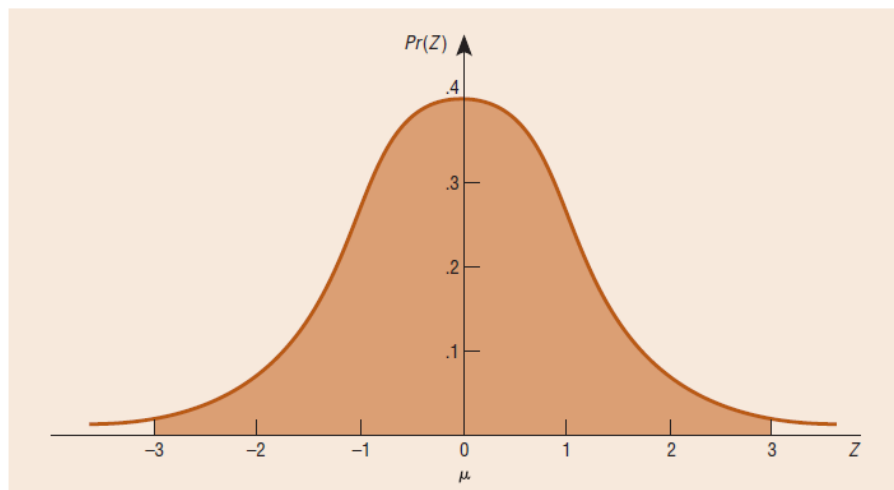


EXHIBIT 17.10 Standardized Normal Table: Area under Half of the Normal Curve^a

Z Standard Deviations from the Mean (Units)	Z Standard Deviations from the Mean (Tenths of Units)									
	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0.0	.000	.040	.080	.118	.155	.192	.226	.258	.288	.315
1.0	.341	.364	.385	.403	.419	.433	.445	.455	.464	.471
2.0	.477	.482	.486	.489	.492	.494	.495	.496	.497	.498
3.0	.499	.499	.499	.499	.499	.499	.499	.499	.499	.499

^aArea under the segment of the normal curve extending (in one direction) from the mean to the point indicated by each row-column combination. For example, about 68 percent of normally distributed events can be expected to fall within 1.0 standard deviation on either side of the mean (0.341×2). An interval of almost 2.0 standard deviations around the mean will include 95 percent of all cases.

Computing the standardized value, Z , of any measurement expressed in original units is simple: Subtract the mean from the value to be transformed, and divide by the standard deviation (all expressed in original units). The formula for this procedure and its verbal statement follow. In the formula, note that σ , the population standard deviation, is used for calculation. Also note that we do not use an absolute value, but rather allow the Z value to be either negative (below the mean) or positive (above the mean).

$$\text{Standardized value} = \frac{\text{Value to be transformed} - \text{Mean}}{\text{Standard deviation}}$$

$$Z = \frac{X - \mu}{\sigma}$$

where

μ = hypothesized or expected value of the mean

Suppose that in the past a toy manufacturer has experienced mean sales, μ , of 9,000 units and a standard deviation, σ , of 500 units during September. The production manager wishes to know whether wholesalers will demand between 7,500 and 9,625 units during September of the upcoming year. Because no tables are available showing the distribution for a mean of 9,000 and a standard deviation of 500, we must transform our distribution of toy sales, X , into the standardized form using our simple formula:

$$Z = \frac{X - \mu}{\sigma} = \frac{7,500 - 9,000}{500} = -3.00$$

$$Z = \frac{X - \mu}{\sigma} = \frac{9,625 - 9,000}{500} = 1.25$$

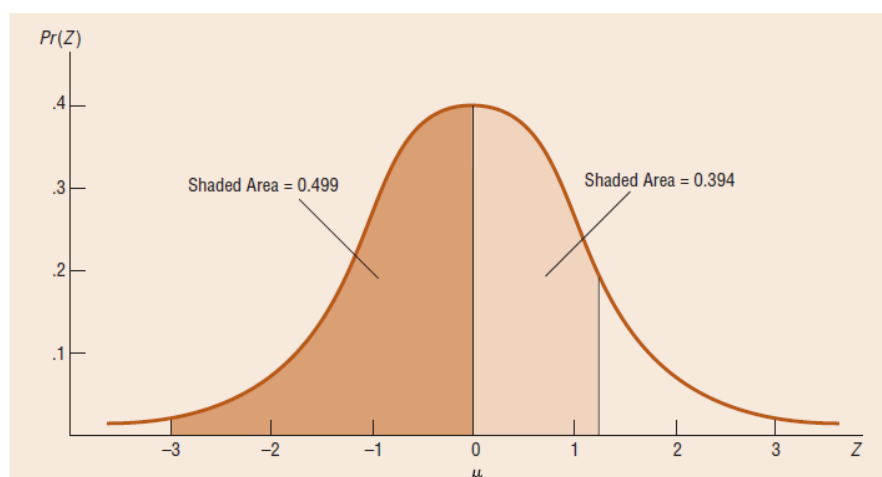
The —3.00 indicates the standardized Z for sales of 7,500, while the 1.25 is the Z score for 9,625. Using Exhibit 17.10 (or Table A.2 in the appendix), we find that

When $Z = -3.00$, the area under the curve (probability) equals 0.499.

When $Z = 1.25$, the area under the curve (probability) equals 0.394.

Thus, the total area under the curve is $0.499 + 0.394 = 0.893$. In other words, the probability (Pr) of obtaining sales in this range is equal to 0.893. This is illustrated in Exhibit 17.12 in the shaded area. The sales manager, therefore, knows there is a 0.893 probability that sales will be between 7500 and 9,625. We can go a step further here by comparing the area under the curve to the total. Since the distribution is symmetrical, 0.500 of the distribution is on either side of the center line. For the 7,500 figure the area under our curve is 0.499, so the probability of sales being *less* than 7,500 is 0.001 ($0.500 - 0.499$). Similarly, the probability of sales being *more* than 9,625 is 0.106 ($0.500 - 0.394$).

Standardized Distribution Curve



Population Distribution, Sample Distribution, and Sampling Distribution

A frequency distribution of the population elements is called a **population distribution**. The mean and standard deviation of the population distribution are represented by the Greek letters μ and σ . A frequency distribution of a sample is called a **sample distribution**. The sample mean is designated \bar{X} , and the sample standard deviation is designated S .

The concepts of population distribution and sample distribution are relatively simple. However, we must now introduce another distribution, which is the crux of understanding statistics: the *sampling distribution of the sample mean*. The sampling distribution is a theoretical probability distribution that in actual practice would never be calculated. Hence, practical, business-oriented students have difficulty understanding why the notion of the sampling distribution is important. Statisticians, with their mathematical curiosity, have asked themselves, "What would happen if we were to draw a large number of samples (say, 50,000), each having n elements, from a specified population?" Assuming that the samples were randomly selected, the sample means, \bar{X} s, could be arranged in a frequency distribution. Because different people or sample units would be selected in the different samples, the sample means would not be exactly equal. The shape of the sampling distribution is of considerable importance to statisticians. If the sample size is sufficiently large and if the samples are randomly drawn, we know from the central-limit theorem (discussed below) that the sampling distribution of the mean will be approximately normally distributed.

A formal definition of the sampling distribution is as follows:

A **sampling distribution** is a theoretical probability distribution that shows the functional relation between the possible values of some summary characteristic of n cases drawn at random and the probability (density) associated with each value over all possible samples of size n from a particular population.

The sampling distribution's mean is called the *expected value* of the statistic. The expected value of the mean of the sampling distribution is equal to μ . The standard deviation of a sampling distribution of \bar{X} is called **standard error of the mean** ($S_{\bar{X}}$) and is approximately equal to

$$S_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

mean

The standard deviation of the sampling distribution.

To review, for us to make an inference about a population from a sample, we must know about three important distributions: the population distribution, the sample distribution, and the sampling distribution. They have the following characteristics:

	Mean	Standard Deviation
Population distribution	μ	σ
Sample distribution	\bar{X}	S
Sampling distribution	$\mu_{\bar{X}} = \mu$	$S_{\bar{X}}$

We now have much of the information we need to understand the concept of statistical inference. To clarify why the sampling distribution has the characteristic just described, we will elaborate on two concepts: the standard error of the mean and the central-limit theorem. You may be wondering why the standard error of the mean, $S_{\bar{X}}$, is defined as $S_{\bar{X}} = \sigma/\sqrt{n}$. The reason is based on the notion that the variance or dispersion within the sampling distribution of the mean will be less if we have a larger sample size for independent samples. It should make intuitive sense that a larger sample size allows the researcher to be more confident that the sample mean is closer to the population mean. In actual practice, the standard error of the mean is estimated using the sample's standard deviation. Thus, $S_{\bar{X}}$ is estimated using S/\sqrt{n} .

Point Estimates

Our goal in using statistics is to make an estimate about population parameters. A population mean, μ , and standard deviation, σ , are constants, but in most instances of business research, they are unknown. To estimate population values, we are required to sample. As we have discussed, \bar{X} and S are random variables that will vary from sample to sample with a certain probability (sampling) distribution.

Our previous example of statistical inference was somewhat unrealistic because the population had only six individuals. Consider the more realistic example of a prospective racquetball entrepreneur who wishes to estimate the average number of days players participate in this sport each week. When statistical inference is needed, the population mean, μ , is a constant but unknown parameter. To estimate the average number of playing days, we could take a sample of three hundred racquetball players throughout the area where our entrepreneur is thinking of building club facilities. If the sample mean, \bar{X} , equals 2.6 days per week, we might use this figure as a **point estimate**. This single value, 2.6, would be the best estimate of the population mean. However, we would be extremely lucky if the sample estimate were exactly the same as the population value. A less risky alternative would be to calculate a confidence interval. An example of a point estimate and confidence interval is provided in the Research Snapshot on the next page.

Confidence Intervals

If we specify a range of numbers, or interval, within which the population mean should lie, we can be more confident that our inference is correct. A **confidence interval estimate** is based on the knowledge that $\mu = \bar{X} \pm$ a small sampling error. After calculating an interval estimate, we can determine how probable it is that the population mean will fall within this range of statistical values. In the racquetball project, the researcher, after setting up a confidence interval, would be able to make a statement such as “With 95 percent confidence, I think that the average number

of days played per week is between 2.3 and 2.9.” This information can be used to estimate market demand because the researcher has a certain confidence that the interval contains the value of the true population mean.

The crux of the problem for a researcher is to determine how much random sampling error to tolerate. In other words, what should the confidence interval be? How much of a gamble should be taken that μ will be included in the range? Do we need to be 80 percent, 90 percent, 95 percent, or 99 percent sure? The **confidence level** is a percentage or decimal that indicates the long-run probability that the results will be correct. Traditionally, researchers have used the 95 percent confidence level. While there is nothing magical about the 95 percent confidence level, it is useful to select this confidence level in our examples.

As mentioned, the point estimate gives no information about the possible magnitude of random sampling error. The confidence interval gives the estimated value of the population parameter, plus or minus an estimate of the error. We can express the idea of the confidence interval as follows:

$$\mu = \bar{X} \pm \text{a small sampling error}$$

More formally, assuming that the researchers select a large sample (more than 30 observations), the small sampling error is given by

$$\text{Small sampling error} = Z_{c.l.} S_{\bar{X}}$$

where

$Z_{c.l.}$ = value of Z , or standardized normal variable, at a specified confidence level ($c.l.$)

$S_{\bar{X}}$ = standard error of the mean

The precision of our estimate is indicated by the value of $Z_{c.l.} S_{\bar{X}}$. It is useful to define the range of possible error, E , as follows:

$$E = Z_{c.l.} S_{\bar{X}}$$

Thus,

$$\mu = \bar{X} \pm E$$

where

\bar{X} = sample mean

E = range of sampling error

or

$$\mu = \bar{X} \pm Z_{c.l.} S_{\bar{X}}$$

The confidence interval $\pm E$ is always stated as one-half (thus the plus or minus) of the total confidence interval.

The following step-by-step procedure can be used to calculate confidence intervals:

1. Calculate \bar{X} from the sample.
2. Assuming σ is unknown, estimate the population standard deviation by finding S , the sample standard deviation.
3. Estimate the standard error of the mean, using the following formula:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

4. Determine the Z -value associated with the desired confidence level. The confidence level should be divided by 2 to determine what percentage of the area under the curve to include on each side of the mean.
5. Calculate the confidence interval.

Sample Size

Random Error and Sample Size

When asked to evaluate a business research project, most people, even those with little research training, begin by asking, "How big was the sample?" Intuitively we know that the larger the sample, the more accurate the research. This is in fact a statistical truth; random sampling error varies with samples of different sizes. In statistical terms, increasing the sample size decreases the width of the confidence interval at a given confidence level. Obviously if we collect information from every member of the population, we know the population parameters, so there would be no interval. When the standard deviation of the population is unknown, a confidence interval is calculated using the following formula:

$$\text{Confidence interval} = \bar{X} \pm Z \frac{S}{\sqrt{n}}$$

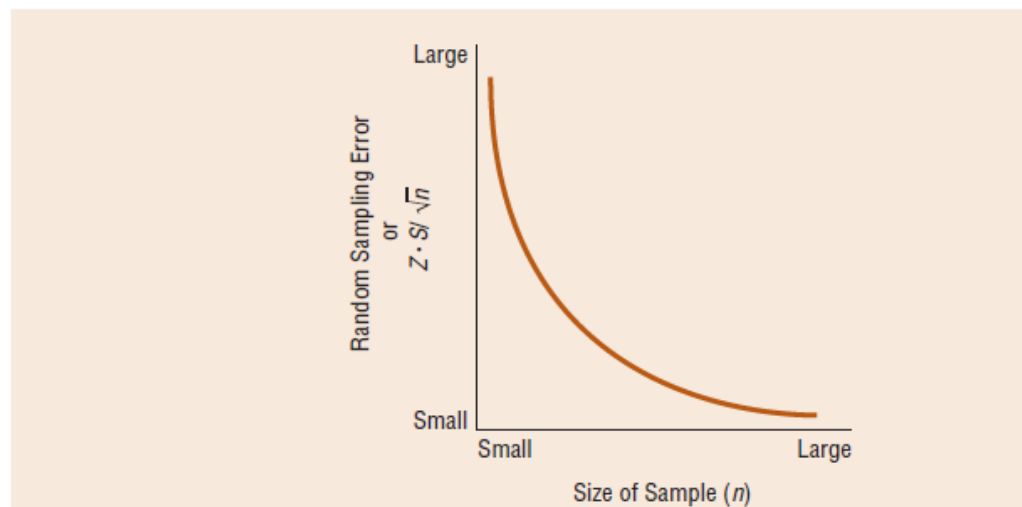
Observe that the equation for the plus or minus error factor in the confidence interval includes n , the sample size:

$$E = Z \frac{S}{\sqrt{n}}$$

If n increases, E is reduced. Exhibit 17.18 illustrates that the confidence interval (or magnitude of error) decreases as the sample size, n , increases.

We already noted that it is not necessary to take a census of all elements of the population to conduct an accurate study. The laws of probability give investigators sufficient confidence regarding the accuracy of data collected from a sample. Knowledge of the characteristics of the sampling distribution helps researchers make reasonably precise estimates.

EXHIBIT 17.18

Relationship between Sample Size and Error

Increasing the sample size reduces the sampling error. However, those familiar with the law of diminishing returns in economics will easily grasp the concept that increases in sample size reduce sampling error at a *decreasing rate*. For example, doubling a sample of 1,000 will reduce random sampling error by 1 percentage point, but doubling the sample from 2,000 to 4,000 will reduce random sampling error by only another half percentage point. More technically, random sampling error is inversely proportional to the square root of n . (Exhibit 17.18 gives an approximation of the relationship between sample size and error.) Thus, the main issue becomes one of determining the optimal sample size. The Research Snapshot above discusses sample size and shows that some samples are extremely large.

Factors in Determining Sample Size for Questions Involving Means

Three factors are required to specify sample size: (1) the heterogeneity (i.e., variance) of the population; (2) the magnitude of acceptable error (i.e., \pm some amount); and (3) the confidence level (i.e., 90 percent, 95 percent, 99 percent).

The determination of sample size heavily depends on the variability within the sample. The *variance*, or *heterogeneity*, of the population is the first necessary bit of information. In statistical terms, this refers to the *standard deviation* of the population. Only a small sample is required if the population is homogeneous. For example, predicting the average age of college students requires a smaller sample than predicting the average age of people who visit the zoo on a given Sunday afternoon. As *heterogeneity* increases, so must sample size. Thus, to test the effectiveness of an employee training program, the sample must be large enough to cover the range of employee work experience (for example).

The *magnitude of error*, or the confidence interval, is the second necessary bit of information. Defined in statistical terms as E , the magnitude of error indicates how precise the estimate must be. It indicates a certain precision level. From a managerial perspective, the importance of the decision in terms of profitability will influence the researcher's specifications of the range of error. If, for example, favorable results from a test-market sample will result in the construction of a new plant and unfavorable results will dictate not marketing the product, the acceptable range of error probably will be small; the cost of an error would be too great to allow much room for random sampling errors. In other cases, the estimate need not be extremely precise. Allowing an error of $\pm \$1,000$ in total family income instead of $E = \pm 50$ may be acceptable in most market segmentation studies.

The third factor of concern is the *confidence level*. In our examples, as in most business research, we will typically use the 95 percent confidence level. This, however, is an arbitrary decision based on convention; there is nothing sacred about the 0.05 chance level (that is, the probability of 0.05 of the true population parameter being incorrectly estimated).

Estimating Sample Size for Questions Involving Means

Once the preceding concepts are understood, determining the actual size for a simple random sample is quite easy. The researcher must follow three steps:

1. Estimate the standard deviation of the population.
2. Make a judgment about the allowable magnitude of error.
3. Determine a confidence level.

The judgment about the allowable error and the confidence level are the manager's decision to make. Thus, the only problem is estimating the standard deviation of the population. Ideally, similar studies conducted in the past will give a basis for judging the standard deviation. In practice, researchers who lack prior information may conduct a pilot study to estimate the population parameters so that another, larger sample of the appropriate sample size may be drawn. This procedure is called *sequential sampling* because researchers take an initial look at the pilot study results before deciding on a larger sample to provide more precise information.

A rule of thumb for estimating the value of the standard deviation is to expect it to be about one-sixth of the range. If researchers conducting a study on television purchases expected the price paid to range from \$100 to \$700, a rule-of-thumb estimate for the standard deviation would be \$100. This is also useful when the question is a scaled response on a questionnaire. For example, if we plan on using a 10-point purchase intention scale, we can use our rule to determine the estimate for the standard deviation ($10/6 = 1.67$).

For the moment, assume that the standard deviation has been estimated in some preliminary work. If our concern is to estimate the mean of a particular population, the formula for sample size is

$$n = \left(\frac{ZS}{E} \right)^2$$

where

Z = standardized value that corresponds to the confidence level

S = sample standard deviation or estimate of the population standard deviation

E = acceptable magnitude of error, plus or minus error factor (range is one-half of the total confidence interval)⁷

Suppose a survey researcher studying annual expenditures on lipstick wishes to have a 95 percent confidence level ($Z = 1.96$) and a range of error (E) of less than \$2. If the estimate of the standard deviation is \$29, the sample size can be calculated as follows:

$$n = \left(\frac{ZS}{E} \right)^2 = \left(\frac{(1.96)(29)}{2} \right)^2 = \left(\frac{56.84}{2} \right)^2 = 28.42^2 = 808$$

If a range of error (E) of \$4 is acceptable, the necessary sample size will be reduced:

$$n = \left(\frac{ZS}{E} \right)^2 = \left(\frac{(1.96)(29)}{4} \right)^2 = \left(\frac{56.84}{4} \right)^2 = 14.21^2 = 202$$

Thus, doubling the range of acceptable error reduces sample size to approximately one-quarter of its original size. Stated conversely in a general sense, doubling sample size will reduce error by only approximately one-quarter.

The Influence of Population Size on Sample Size

The ACNielsen Company estimates television ratings. Throughout the years, it has been plagued with questions about how it is possible to rate 98 million or more television homes with such a small sample (approximately 5,000 households). The answer to that question is that in most cases the size of the population does not have an effect on the sample size. As we have indicated, the variance of the population has the largest effect on sample size. However, a finite correction factor may be needed to adjust a sample size that is more than 5 percent of a finite population. If the sample is large relative to the population, the foregoing procedures may overestimate sample size, and the researcher may need to adjust sample size. The finite correction factor is

$$\sqrt{\frac{(N-n)}{(N-1)}}$$

where

N = population size and n = sample size.

Factors in Determining Sample Size for Proportions

Researchers frequently are concerned with determining sample size for problems that involve estimating population proportions or percentages. When the question involves the estimation of a proportion, the researcher requires some knowledge of the logic for determining a confidence interval around a sample proportion estimation (p) of the population proportion (π). For a confidence interval to be constructed around the sample proportion (p), an estimate of the standard error of the proportion (S_p) must be calculated and a confidence level specified.

The precision of the estimate is indicated by the value $Z_{cl}S_p$. Thus, the plus-or-minus estimate of the population proportion is

$$\text{Confidence interval} = p \pm Z_{cl}S_p$$

If the researcher selects a 95 percent probability for the confidence interval, Z_{cl} will equal 1.96 (see Table A.2 in the appendix). The formula for S_p is

$$S_p = \sqrt{\frac{pq}{n}} \text{ or } S_p = \sqrt{\frac{p(1-p)}{n}}$$

where

S_p = estimate of the standard error of the proportion

p = proportion of successes

$q = 1 - p$, or proportion of failures

Suppose that 20 percent of a sample of 1,200 television viewers recall seeing an advertisement. The proportion of successes (p) equals 0.2, and the proportion of failures (q) equals 0.8. We estimate the 95 percent confidence interval as follows:

$$\begin{aligned} \text{Confidence Interval} &= p \pm Z_{cl}S_p \\ &= 0.2 \pm 1.96S_p \\ &= 0.2 \pm 1.96\sqrt{\frac{p(1-p)}{n}} \\ &= 0.2 \pm 1.96\sqrt{\frac{0.2(1-0.2)}{1,200}} \\ &= 0.2 \pm 1.96\sqrt{\frac{0.16}{1,200}} = 0.2 \pm 1.96(0.0115) \\ &= 0.2 \pm 0.022 \end{aligned}$$

Thus, the population proportion who see an advertisement is estimated to be included in the interval between 0.178 and 0.222, or roughly between 18 and 22 percent, with a 95 percent confidence coefficient.

To determine *sample size* for a proportion, the researcher must make a judgment about confidence level and the maximum allowance for random sampling error. Furthermore, the size of the proportion influences random sampling error, so an estimate of the expected proportion of successes must be made, based on intuition or prior information. The formula is

$$n = \frac{Z_{c.l.}^2 pq}{E^2}$$

where

n = number of items in sample

$Z_{c.l.}^2$ = square of the confidence level in standard error units

p = estimated proportion of successes

$q = 1 - p$, or estimated proportion of failures

E^2 = square of the maximum allowance for error between the true proportion and the sample proportion, or $Z_{c.l.} S_p$ squared

Suppose a researcher believes that a simple random sample will show that 60 percent of the population (p) recognizes the name of an automobile dealership. The researcher wishes to estimate with 95 percent confidence ($Z_{c.l.} = 1.96$) that the allowance for sampling error is not greater than 3.5 percentage points (E). Substituting these values into the formula gives

$$\begin{aligned} n &= \frac{(1.96)^2(0.6)(0.4)}{0.035^2} \\ &= \frac{(3.8416)(0.24)}{0.001225} \\ &= \frac{0.922}{0.001225} \\ &= 753 \end{aligned}$$

Determining Level of Precision after Data Collection

Up to this point, we have discussed the process for determining how large of a sample we need to collect given the estimated variance among the responses and our desired level of precision and acceptable error. This is a very important consideration for researchers. However, after we have collected the data, we also want to determine our level of precision, given the size of the sample, the variance, and the confidence level. In this case, we can rewrite our equation for determining sample size:

$$n = \left(\frac{ZS}{E} \right)^2$$

Rather than solving for n , we now know n and instead want to solve for E , the magnitude of error. Our new equation would be:

$$E^2 = \frac{(ZS)^2}{n}$$

So, we could solve for E^2 , and then take the square root of this to determine our level of precision. This is a useful approach to use after-the-fact to show our final level of precision. In our earlier example of sample size regarding lipstick expenditures, we found that if we wanted to be 95% confident (Z value of 1.96) that our estimate of expenditures was within \$2.00 and we had a standard deviation of \$29.00, we would need a sample size of 808. Using the same situation, let's assume we had already collected the data, but were not certain of our level of precision. Our formula above would show:

$$(1.96*29)^2/808 = 4$$

The square root of 4 is 2.

When completing a research project it is often a good idea to provide managers with the level of precision for key measures. This formula will allow you to do so.