Bringing Research to Life

The executive director of White Ice Compound gestured broadly at the still snow-capped Canadian Rockies that enveloped the complex and discouraged casual visitation for most of the year. "It has been three very happy years for me here, though not easy on my ego since I let corporate North America intrude on our idyllic existence. Not that I blame them for my shortcomings."

"You mean the MindWriter people?" prompted Jason. "The ones who flew me up here? My clients?"

The executive director propelled Jason straight across a manicured lawn toward the refreshment tent, where her faculty and paying guests were basking in postconcert euphoria, following a stirring performance of Beethoven by the White Ice Summer Festival Orchestra.

"Please, don't misunderstand," said the executive director. "They have been fine tenants . . . good corporate citizens . . . generous contributors to our little community. When I rented them a part of our compound for use in corporate education, they quite generously insisted that I avail myself of some of their training for midlevel managers. And I have to admit, now, that their managerial style makes me feel like an inadequate cellist with a stiff wrist, not an executive director evolving toward competence."

"Well, if we are going to help you," ventured Jason, "you had better tell me quickly what you do here. I have a 4 P.M. flight out."

"Surely. We in White Ice have a simple, never-varying rhythm of activity. By the middle of September, the paying guests, the visiting artists, the musicians, and the tourists have left White Ice and I bring in artisans for two weeks of intensive repairs and renovations. Then I prepare financial and artistic reports for the three foundations that have endowed us and also draw up an agenda for capital improvements and special events, which becomes the basis for frantic proposal writing during the weeks preceding Christmas. From January to April, I am on the phone to travel agents from Mexico City to Juneau to arrange a tight reservation and scheduling process, so that we maximize the use of the facilities during our season. I have developed the ability to keep track of the cash flow, which is not easy, with Canadian and U.S. dollars mingled.

"During the winter my artistic directors, Frances Braun and Igor Starvinsky—they have been Mr. and Mrs. Braun-Starvinsky for 30 years—prepare the program and hire the musicians, coordinating closely with me on the budget. This is quite complicated, as most of the performing artists spend only two weeks with us, so that fully 600 artists are part of this orchestra over the course of a summer.

"Then in the early spring I hear from the colleges in British Columbia, who send me their music scholarship students for summer employment as dishwashers, waiters, cleaners, and the like."

"Sounds as if you are right on top of the finances," said Jason, "and I suppose in your seminars with the MindWriter people they told you cash flow is one of three things that must be watched most carefully."

"Oh, yes indeed," the executive director laughed. "Gauging cash flow is not the problem. I descend from a line of genteel poverty. Measuring customer satisfaction—the second of the critical three factors for the MindWriter folks—now that was a problem for me—at first. The care and frequency with which they measure customer satisfaction in the MindWriter seminars dumbfounded me. Throughout a seminar, morning, afternoon, or evening, everyone breaks for coffee and is required to fill out a critique of the speaker. The results are tabulated by the time the last coffee cup has been picked up, and the seminar leader has been given feedback. Is he or she pre-

senting material too slowly or too quickly? Are there too many jokes or not enough? Are concrete examples being used often enough? Do the participants want a hard copy of the slides? They measure attitudinal data six times a day and even query you about the meals, including taste, appearance, cleanliness and speed, friendliness, and accuracy of service.

"My problem is employee commitment, specifically commitment of the orchestral performers to the White Ice Festival. None of the other directors in my North American association of artistic and executive directors has nearly the rapid turnover of performers we have here, so they are not much help."

"Quoting the MindWriter standard line, 'Employee satisfaction is the third leg of the three- legged stool on which performance is based,'" said Jason. "The Braun-Starvinskys have the most contact with performers; they could be your eyes and ears. Ask them to listen carefully."

The executive director laughed ironically. "Jason," she said, "look over my shoulder. Directly behind me is a couple in their mid-sixties. Please describe as exactly as you can the behavior you observe."

"You mean the fellow in the sweatshirt and the woman with her hair in a bun?" whispered Jason. "He is sleeping. And the woman is nevertheless talking to him nonstop. Ah! Now she's shaking him awake. But he appears to fall right back to sleep. Does she ever stop talking?"

"There you have them, Jason—the Braun-Starvinskys, my artistic directors. He stays up all night composing, and all day, when he is not conducting, he snoozes. And she never stops expressing her opinions, be they lifelong prejudices or vagrant musings. Therefore she never listens well enough to later give a coherent report of anything she has heard or been told. If I have to rely on them for feedback, everything would be filtered and distorted beyond recognition."

"It is just as well. Untrained observers can be highly unreliable and inaccurate in measuring and reporting behavior," said Jason. "Have you tried a suggestion box?"

"No, but I do send a letter to each visiting performer soliciting bouquets and brickbats. Do you want to know what some of the performers have written?"

"Shoot," said Jason. "But, quickly, please. I don't want to fly through these mountains at night in a small plane."

"Here is just a sample: 'Starvinsky never listens to our ideas.' 'A day under Braun feels like a week on a Los Angeles freeway.' 'We are all highly trained college teachers of music, but we are treated like children.'

"Clearly our performers aren't our only concern. The restaurant employees, the hospitality staff, the stage carpenters, the . . ."

"Hold on," said Jason, scribbling furiously on a napkin. "I can see you have a problem. I'm making a note to send you some research indexes on work innovation and job motivation. In fact, I believe one identifies and measures five different dimensions of worker attitudes. You'll find it interesting and maybe it is something you can use. Meanwhile, will you send me your customer satisfaction instrument for concert goers?"

"Of course, Jason. And be sure I shall act quickly on your suggestions. The *Vancouver Sun* has commented on our inability to sustain a steady tempo and tonation. When a businessperson fouls up, the mistake may not be evident for days or weeks. But when our orchestra strikes a sour note, 600 audience members receive the message at the speed of sound."

The Nature of Measurement

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

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

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

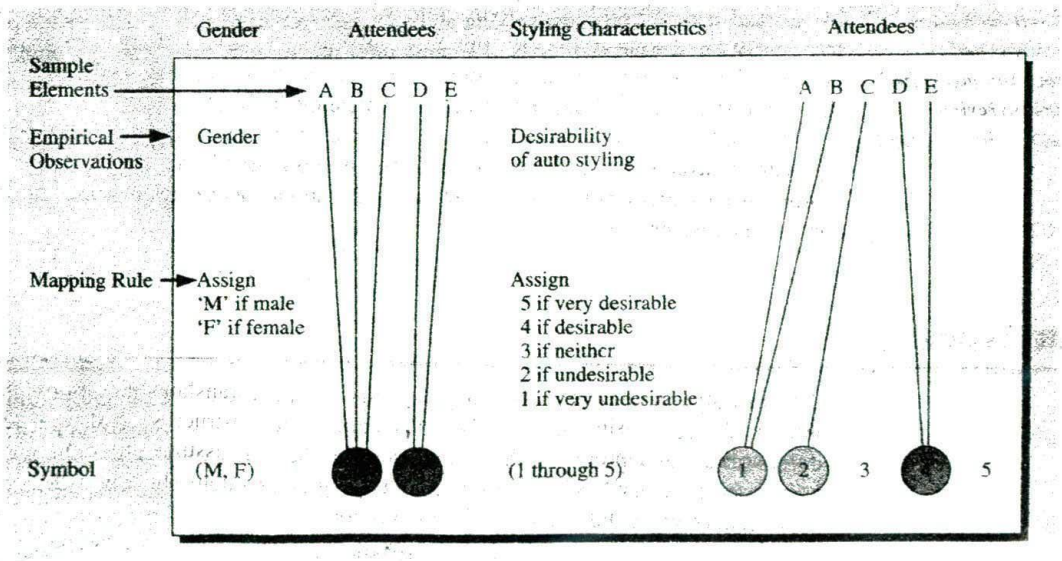
Assume you are studying people who attend an auto show where all of the year's new models are on display. You are interested in learning the male-to-female ratio among attendees. You observe those who enter the show area. If a person is female, you record an F; if male, an M. Any other symbols such as 0 and 1 or # and % also may be used if you know what group the symbol identifies. Exhibit 8–1 uses this example to illustrate the above components.

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

What is your opinion of the styling of the Espace van?

Very desirable — 5 — 4 — 3 — 2 — 1 — Very undesirable

EXHIBIT 8–1 Characteristics of Measurement



All measurement theorists would call the opinion rating scale on page 221 a form of measurement, but some would challenge the male-female classification. Their argument is that measurement must involve quantification—that is, "the assignment of numbers to objects to represent amounts or degrees of a property possessed by all of the objects."[2] Our discussion endorses the more general view that numbers as symbols within a mapping rule can reflect both qualitative and quantitative concepts.

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

What Is Measured?

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

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

Earlier we discussed operational definitions for constructs and concepts. You might find it helpful to revisit Exhibit 2–4 in Chapter 2.

In contrast, it is not easy to measure properties like "motivation to succeed," "ability to stand stress," "problem-solving ability," and "persuasiveness." Since each property cannot be measured directly, one must infer its presence or absence by observing some indicant or pointer measurement. When you begin to make these inferences, there is often disagreement about how to operationalize the indicants.

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

Data Types

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

EXHIBIT 8–2 Types of Data and Their Measurement Characteristics

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

Mapping rules have four characteristics:

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

Combinations of these characteristics of classification, order, distance, and origin provide four widely used classification of measurement scales: (1) nominal, (2) ordinal, (3) interval, and (4) ratio.

The characteristics of these measurement scales are summarized in Exhibit 8–2. Deciding which data type is appropriate for your research needs should be seen as a process (see Exhibit 8–3).

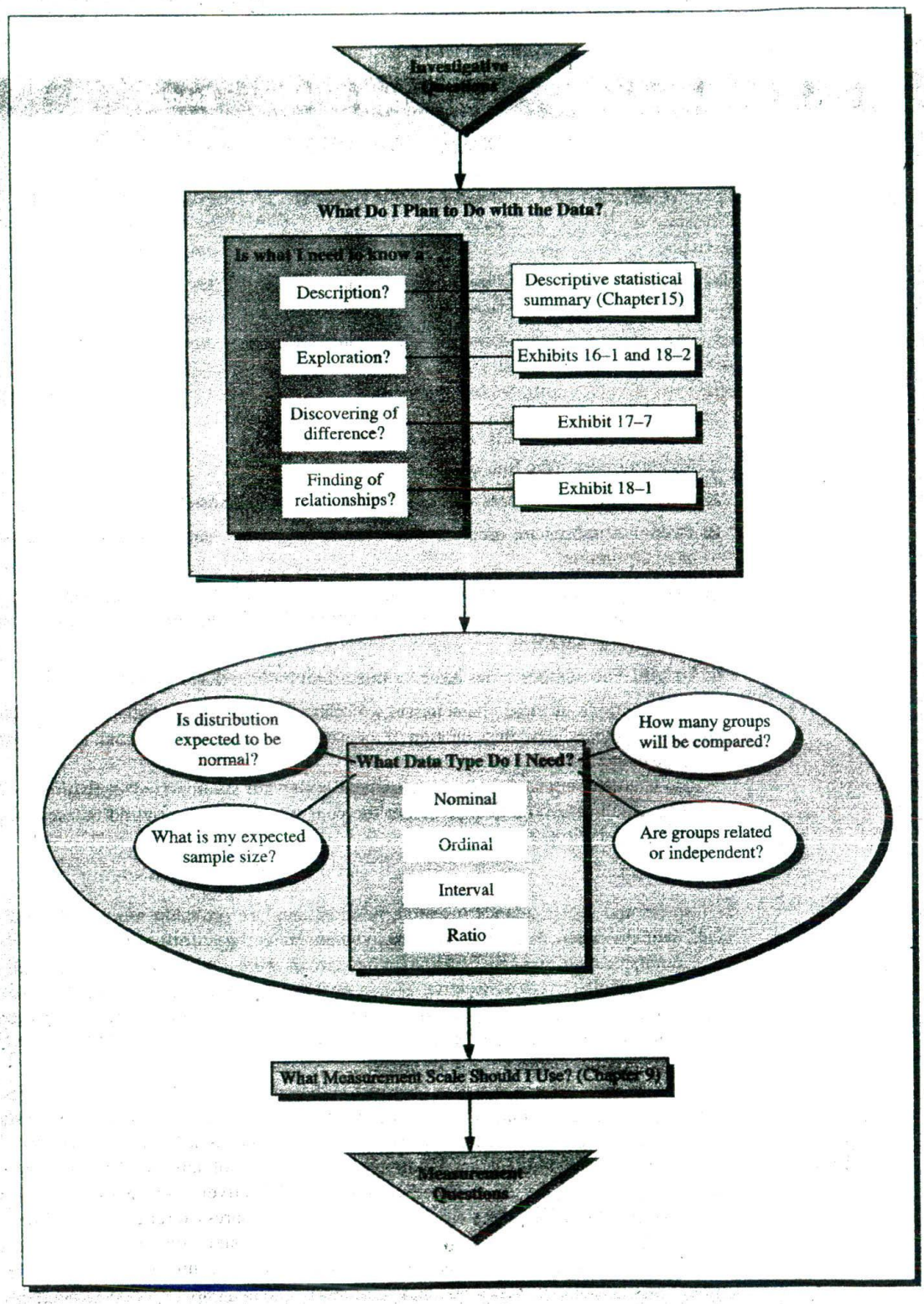
Nominal Data

In business and social science research, nominal data are probably more widely collected than any other. With **nominal data,** you are collecting information on a variable that naturally or by design can be grouped into two or more categories that are mutually exclusive and collectively exhaustive. If data were collected from the performing artists at the White Ice Compound, each artist could be classified by whether he or she stayed the summer or departed early. Every performer would fit into one of the two groups within the variable *duration of employment.*

MANAGEMENT Tip

The counting of members in each group is the only possible arithmetic operation when a nominal scale is employed. If we use numerical symbols within our mapping rule to identify categories, these numbers are recognized as labels only and have no quantitative value. Nominal classifications may consist of any number of separate groups if the groups are mutually exclusive and collectively exhaustive. Thus, one might classify the residents of a city according to their expressed religious preferences. Mapping Rule A given in the table on page 225 is not a sound nominal scale because it is not collectively exhaustive. Mapping Rule B meets the minimum requirements, although this classification may be more useful for some research purposes than others.

EXHIBIT 8–3 Moving from Investigative to Measurement Questions



Religious Preferences	
Mapping Rule A	**Mapping Rule B**
1 = Baptist	1 = Protestant
2 = Catholic	2 = Catholic
3 = Jewish	3 = Jewish
4 = Lutheran	4 = Other
5 = Methodist	
6 = Presbyterian	
7 = Protestant	

Nominal scales are the least powerful of the four data types. They suggest no order or distance relationship and have no arithmetic origin. The scale wastes any information a sample element might share about varying degrees of the property being measured.

MANAGEMENT

We discuss significance tests and measures of association in Chapters 17 and 18.

Since the only quantification is the number count of cases in each category (the frequency distribution), the researcher is restricted to the use of the mode as the measure of central tendency.[3] You can conclude which category has the most members, but that is all. There is no generally used measure of dispersion for nominal scales. Several tests for statistical significance may be utilized; the most common is the chi-square test. For measures of association, phi, lambda, or other measures may be appropriate.

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

Jason visited White Ice because of MindWriter's extensive research into customer satisfaction related to White Ice's manager training. His visit revealed White Ice's need for some exploratory nominal data on employee satisfaction. Orchestra performers could be divided into groups based on their appreciation of the Braun-Starvinskys (favorable, unfavorable), on their attitude toward facilities (suitable, not suitable), or on their perception of how performers were treated (as adults, as children).

Ordinal Data

Ordinal data include the characteristics of the nominal scale plus an indicator of order. Ordinal data are possible if the transitivity postulate is fulfilled. This postulate states: If a is greater than b and b is greater than c, then a is greater than c.[4] The use of an ordinal scale implies a statement of "greater than" or "less than" (an equality statement is also acceptable) without stating how much greater or less. While ordinal measurement speaks of "greater than" and "less than" measurements, other descriptors may be used—"superior to," "happier than," "poorer than," or "above." Like a rubber yardstick, it can stretch varying amounts at different places along its length. Thus, the real difference between ranks 1 and 2 on a happiness scale may be more or less than the difference between ranks 2 and 3.

An ordinal concept can be generalized beyond the three cases used in the simple illustration of $a > b > c$. Any number of cases can be ranked.

A third extension of the ordinal concept occurs when more than one property is of interest. We may ask a taster to rank varieties of carbonated soft drinks by flavor, color, carbonation, and a combination of these characteristics. We can secure the combined ranking either by asking the respondent to base his or her ranking on the combination of properties or by constructing a combination ranking of the individual rankings on each property. To develop this overall index, the researcher typically adds and averages ranks for each of the three properties. This procedure is technically incorrect for ordinal data and, especially for a given respondent, may yield misleading results. When the number of respondents is large, however, these errors average out. A more sophisticated way to combine a number of dimensions into a total index is to use a multidimensional scale (see Chapter 19).

The researcher faces another difficulty when combining the rankings of several respondents. Here again, it is not uncommon to use weighted sums of rank values for a combined index. If there are many observations, this approach will probably give adequate results, though it is not theoretically correct. A better way is to convert ordinal data into interval data, the values of which can then be added and averaged. One well-known example is *Thurstone's Law of Comparative Judgment.*[5] In its simplest form, Thurstone's procedure says the distance between scale positions of two objects, *A* and *B*, depends on the percentage of judgments in which *A* is preferred to *B*.

MANAGEMENT

Examples of ordinal data include opinion and preference scales. Because the numbers of such scales have only a rank meaning, the appropriate measure of central tendency is the median. A percentile or quartile measure reveals the dispersion. Correlation is restricted to various rank-order methods. Measures of statistical significance are technically confined to that body of methods known as *nonparametric methods.*[6]

Researchers in the behavioral sciences differ about whether more powerful parametric significance tests are appropriate with ordinal measures. One position is that this use of parametric tests is incorrect on both theoretical and practical grounds:

> If the measurement is weaker than that of an interval scale, by using parametric methods tests the researcher would "add information" and thereby create distortions.[7]

At the other extreme, some behavioral scientists argue that parametric tests are usually acceptable for ordinal data:

> The differences between parametric and rank-order tests were not great insofar as significance level and power were concerned.[8]

A view between these extremes recognizes that there are risks in using parametric procedures on ordinal data, but these risks are usually not great:

> The best procedure would seem to be to treat ordinal measurements as though they were interval measurements but to be constantly alert to the possibility of gross inequality of intervals.[9]

Because nonparametric tests are abundant, simple to calculate, have good power efficiencies, and do not force the researcher to accept the assumptions of parametric testing, we advise their use with nominal and ordinal data. It is understandable, however, that because parametric tests (such as the *t*-test or analysis of variance) are so versatile, accepted, and understood, they will continue to be used with ordinal data when those data approach interval data characteristics.

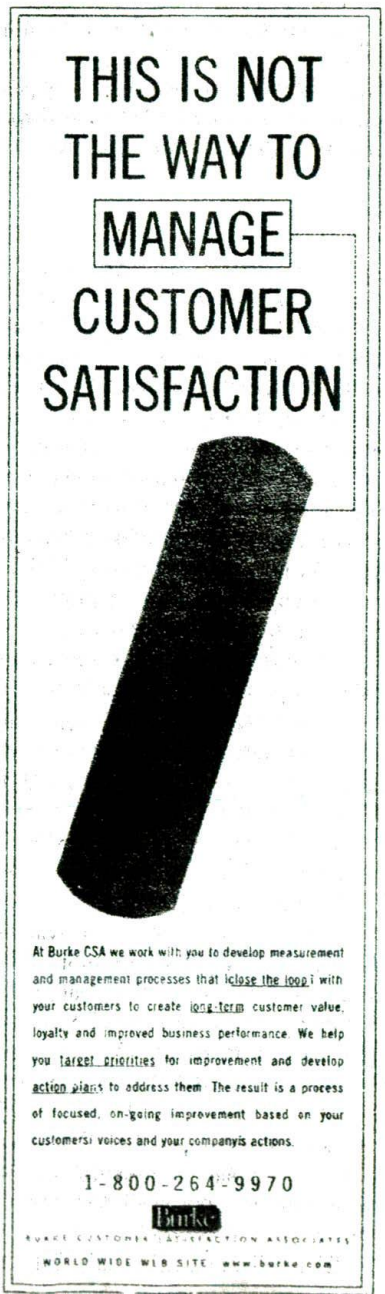
Jason believed White Ice could potentially benefit by using professionally developed, well-tested work evaluation and job motivation indexes (see the opening vignette). Because of the constructs measured (work innovation, job motivation), we know after applying the test that one employee is more motivated than another, that one

employee generates more ideas than another. By applying numerical scores to the variation in motivation, we can assume the collection of interval data.

Interval Data

Interval data have the power of nominal and ordinal data plus one additional strength: They incorporate the concept of equality of interval (the distance between 1 and 2 equals the distance between 2 and 3). Calendar time is such a scale. For example, the

Burke wants companies to think about monitoring customer service before a problem occurs through careful development of measurement and management processes. www.burke.com



elapsed time between 3 and 6 A.M. equals the time between 4 and 7 A.M. One cannot say, however, 6 A.M. is twice as late as 3 A.M. because "zero time" is an arbitrary origin. Centigrade and Fahrenheit temperature scales are other examples of classical interval scales. Both have an arbitrarily determined zero point. Many attitude scales are presumed to be interval. Thurstone's differential scale was an early effort to develop such a scale.[10] Users also treat intelligence scores, semantic differential scales, and many other multipoint graphical scales as interval.

MANAGEMENT Tip

When a scale is interval, you use the arithmetic mean as the measure of central tendency. You can compute the average time of first arrival of trucks at a warehouse or the average attitude value for union workers versus nonunion workers on an election. The standard deviation is the measure of dispersion for arrival times or worker opinions. Product moment correlation, *t*-tests, *F*-tests, and other parametric tests are the statistical procedures of choice.[11]

When the distribution of scores computed from interval data lean in one direction or the other (skewed right or left) we use the median as the measure of central tendency and the interquartile range as the measure of dispersion. The reasons for this are discussed in Chapter 15.

Ratio Data

Ratio data incorporate all of the powers of the previous data types plus the provision for absolute zero or origin. Ratio data represent the actual amounts of a variable. Measures of physical dimensions such as weight, height, distance, and area are examples. In the behavioral sciences, few situations satisfy the requirements of the ratio scale—the area of psychophysics offering some exceptions. In business research, we find ratio scales in many areas. There are money values, population counts, distances, return rates, productivity rates, and amounts of time in a time-period sense.

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

With the White Ice project, Jason could measure the relationship of job satisfaction with a performer's age, the number of years he or she has played professionally, and the number of times he or she has participated in the White Ice summer festival. Each of these examples represents ratio data. For practical purposes, however, the analyst would make the same choice of statistical technique as with interval data.

MANAGEMENT Tip

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

Researchers often encounter the problem of evaluating variables that have been measured at different data levels. The possession of a CPA by an accountant is a nominal, dichotomous variable, and salary is a ratio variable. Certain statistical techniques require the measurement levels to be the same. Since the nominal variable does not have the characteristics of order, distance, or point of origin, we cannot create them artificially after the fact. The ratio-based salary variable, on the other hand, can be reduced.

MANAGEMENT Tip

Rescaling salary downward into high-low, high-medium-low, or another set of categories simplifies the comparison of nominal data. This example may be generalized to other measurement situations—that is, converting or rescaling a variable involves reducing the measure from the more powerful and robust level to a lesser one.[13] The

loss of measurement power accompanying this decision is sometimes costly in that only nonparametric statistics can then be used in data analysis. Thus, the design of the measurement questions should anticipate such problems and avoid them when possible.

Sources of Measurement Differences

The ideal study should be designed and controlled for precise and unambiguous measurement of the variables. Since 100 percent control is unattainable, error does occur. Much potential error is systematic (results from a bias) while the remainder is random (occurs erratically). One authority has pointed out several sources from which measured differences can come.[14]

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

Assume you are conducting an ex post facto study of the residents of a major city. The study concerns the Prince Corporation, a large manufacturer with its headquarters and several major plants located in the city. The objective of the study is to discover the public's opinions about the company and the origin of any generally held adverse opinions.

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

Error Sources

The Respondent Opinion differences that affect measurement come from relatively stable characteristics of the respondent. Typical of these are employee status, ethnic group membership, social class, and nearness to plants. The skilled researcher will anticipate many of these dimensions, adjusting the design to eliminate, neutralize, or otherwise deal with them. However, even the skilled researcher may not be as aware of less obvious dimensions. The latter variety might be a traumatic experience a given respondent had with the Prince Corporation or its personnel. Respondents may be reluctant to express strong negative (or positive) feelings, express opinions which they perceive as different from those of others, or they may have little knowledge about Prince but be reluctant to admit ignorance. This reluctance can lead to an interview of "guesses."

Respondents may also suffer from temporary factors like fatigue, boredom, anxiety, or other distractions; these limit the ability to respond accurately and fully. Hunger, impatience, or general variations in mood may also have an impact.

The portable compact disk player, in combination with CD-burning technology, has made the custom CD not only desirable but also feasible. Measuring attitudes about copyright and its protection is important for publishers, entertainers, and distributors as they search for new business models to accommodate rapidly advancing technology.

SNAPSHOT

Measuring Attitudes about Copyright Infringement

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

> Do you know someone who has done or tried to do any of the following?
>
> 1. Copying software not licensed for personal use.
> 2. Copying a prerecorded videocassette such as a rental or purchased video.
> 3. Copying a prerecorded audiocassette or compact disk.
> 4. Downloading music free or charge from the Internet.
> 5. Photcopying pages from a book or magazine.

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

www.americandemographics.com

www.intersearch.tnsofres.com

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

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

The Instrument A defective instrument can cause distortion in two major ways. First, it can be too confusing and ambiguous. The use of complex words and syntax beyond respondent comprehension is typical. Leading questions, ambiguous meanings, mechanical defects (inadequate space for replies, response choice omissions, and poor printing), and multiple questions suggest the range of problems.

A more elusive type of instrument deficiency is poor selection from the universe of content items. Seldom does the instrument explore all the potentially important issues. The Prince Corporation study might treat company image in areas of employment and ecology but omit the company management's civic leadership, its support of local edu-

cation programs, or its position on minority issues. Even if the general issues are studied, the questions may not cover enough aspects of each area of concern. While we might study the Prince Corporation's image as an employer in terms of salary and wage scales, promotion opportunities, and work stability, perhaps such topics as working conditions, company management relations with organized labor, and retirement and other benefit programs should also be included.

The Characteristics of Sound Measurement

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

- *Validity* refers to the extent to which a test measures what we actually wish to measure.
- *Reliability* has to do with the accuracy and precision of a measurement procedure.
- *Practicality* is concerned with a wide range of factors of economy, convenience, and interpretability.[15]

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

Validity

Many forms of validity are mentioned in the research literature, and the number grows as we expand the concern for more scientific measurement. This text features two major forms: external and internal validity.[16] The external validity of research findings refers to the data's ability to be generalized across persons, settings, and times; we discussed this in reference to sampling in Chapter 7, and more will be said about this in Chapter 14.[17] In this chapter, we discuss only internal validity. Internal validity is further limited in this discussion to the ability of a research instrument to measure what it is purported to measure. Does the instrument really measure what its designer claims it does?

Validity in this context is the extent to which differences found with a measuring tool reflect true differences among respondents being tested. We want the measurement tool to be sensitive to all the nuances of meaning in the variable and to changes in nuances of meaning over time. The difficulty in meeting the test of validity is that usually one does not know what the true differences are. Without direct knowledge of the dimension being studied, you must face the question, "How can one discover validity without directly confirming knowledge?" A quick answer is to seek other relevant evidence that confirms the answers found with the measurement device, but this leads to a second question, "What constitutes relevant evidence?" There is no quick answer this time. What is relevant depends on the nature of the research problem and the researcher's judgment. One way to approach this question is to organize the answer according to measure-relevant types. One widely accepted classification consists of three major forms of validity: (1) content validity, (2) criterion-related validity, and (3) construct validity (see Exhibit 8–4).[18]

Content Validity The **content validity** of a measuring instrument (the composite of measurement scales) is the extent to which it provides adequate coverage of the

EXHIBIT 8–4 Summary of Validity Estimates

Type	What Is Measured	Methods
Content	Degree to which the content of the items adequately represents the universe of all relevant items under study.	Judgmental or panel evaluation with content validity ratio
Criterion-related	Degree to which the predictor is adequate in capturing the relevant aspects of the criterion.	Correlation
Concurrent	Description of the present; criterion data are available at same time as predictor scores.	
Predictive	Prediction of the future; criterion data are measured after the passage of time.	
Construct	Answers the question, "What accounts for the variance in the measure?" Attempts to identify the underlying construct(s) being measured and determine how well the test represents it (them).	Judgmental Correlation of proposed test with established one Convergent-discriminant techniques Factor analysis Multitrait-multimethod analysis

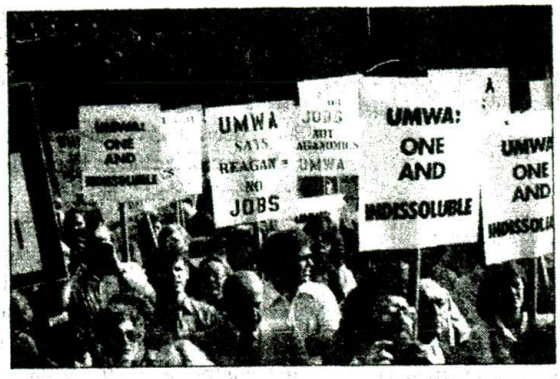
investigative questions guiding the study. If the instrument contains a representative sample of the universe of subject matter of interest, then content validity is good. To evaluate the content validity of an instrument, one must first agree on what elements constitute adequate coverage. In the Prince Corporation study, one must decide what knowledge, attitudes, and opinions are relevant to the measurement of corporate public image and then decide which forms of these opinions are relevant positions on these topics. In the White Ice study, Jason must first determine what factors are influencing employee satisfaction before determining if published indexes can be of value. If the data collection instrument adequately covers the topics that have been defined as the relevant dimensions, we conclude the instrument has good content validity.

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

Determination of content validity is judgmental and can be approached in several ways. First, the designer may determine it through a careful definition of the topic of concern, the items to be scaled, and the scales to be used. This logical process is often intuitive and unique to each research designer.

A second way to determine content validity is to use a panel of persons to judge how well the instrument meets the standards. A panel independently assesses the test items for a performance test. It judges each item to be essential, useful but not essential, or not necessary in assessing performance of a relevant behavior. The "essential" responses on each item from each panelist are evaluated by a content validity ratio, and those meeting a statistical significance value are retained. In both informal judgments and in this systematic process, "content validity is primarily concerned with inferences about test *construction* rather than inferences about test *scores*."[19]

It is important not to define *content* too narrowly. If you were to secure only superficial expressions of opinion in the Prince Corporation public opinion survey, it would probably not have adequate content coverage. The research should delve into the processes by which these opinions came about. How did the respondents come to feel as they do, and what is the intensity of feeling? The same would be true of Mind-



The picket line of strikers, like those seen here for the United Mine Workers of America, is a sight commonly associated with union membership. Peter D. Hart Research Associates, conducting a Society for Human Resource Management (SHRM) worker motivation study for the AFL-CIO, measured via an ordinal measurement scale that younger workers' interest in forming unions was growing not in numbers but in strength of conviction. Comparing a 1999 study with two previous studies conducted in 1997 and 1996, 7 percent more participants selected "would definitely/probably vote for union representation." What would you look for in assessing the soundness of this measurement?

Writer's evaluation of service quality and satisfaction. It is not enough to know a customer is dissatisfied. The manager charged with enhancing or correcting the program needs to know what processes, employees, parts, and time sequences within the CompleteCare program have led to that dissatisfaction.

Criterion-Related Validity **Criterion-related validity** reflects the success of measures used for prediction or estimation. You may want to predict an outcome or estimate the existence of a current behavior or condition. These are *predictive* and *concurrent* validity, respectively. They differ only in a time perspective. An opinion questionnaire that correctly forecasts the outcome of a union election has predictive validity. An observational method that correctly categorizes families by current income class has concurrent validity. While these examples appear to have simple and unambiguous validity criteria, there are difficulties in estimating validity. Consider the problem of estimating family income. There clearly is a knowable true income for every family. However, we may find it difficult to secure this figure. Thus, while the criterion is conceptually clear, it may be unavailable.

In other cases, there may be several criteria, none of which is completely satisfactory. Consider again the problem of judging success among the sales force at SalesPro. A researcher may want to develop a pre-employment test that will predict sales success. There may be several possible criteria, none of which individually tells the full story. Total sales per salesperson may not adequately reflect territory market potential, competitive conditions, or the different profitability rates of various products. One might rely on the sales manager's overall evaluation, but how unbiased and accurate are those impressions? The researcher must ensure that the validity criterion used is itself "valid." One source suggests that any criterion measure must be judged in terms of four qualities: (1) relevance, (2) freedom from bias, (3) reliability, and (4) availability.[20]

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

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

A *reliable* criterion is stable or reproducible. An erratic criterion (using monthly sales, which are highly variable from month to month) can hardly be considered a reliable standard by which to judge performance on a sales employment test. Yet if an unreliable criterion is the only one available, it is often chosen for the study's purpose. In such a case, it is possible to use a *correction for attenuation* formula that lets you see what the correlation between the test and the criterion would be if they were made perfectly reliable.[21]

Finally, the information specified by the criterion must be *available.* If it is not available, how much will it cost and how difficult will it be to secure? The amount of money and effort that should be spent on development of a criterion depends on the importance of the problem for which the test is used.

Chapter 18 describes statistical techniques used to find correlation between variables.

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

Construct Validity One may also wish to measure or infer the presence of abstract characteristics for which no empirical validation seems possible. Attitude scales and aptitude and personality tests generally concern concepts that fall in this category. Although this situation is much more difficult, some assurance is still needed that the measurement has an acceptable degree of validity.

In attempting to evaluate **construct validity,** we consider both the theory and the measuring instrument being used. If we were interested in measuring the effect of ceremony on organizational culture, the way in which "ceremony" was operationally defined would have to correspond to an empirically grounded theory. Once assured that the construct was meaningful in a theoretical sense, we would next investigate the adequacy of the instrument. If a known measure of ceremony in organizational culture was available, we might correlate the results obtained using this measure with those derived from our new instrument. Such an approach would provide us with preliminary indications of *convergent* validity. If Jason were to develop a work innovation index for artistic personnel at White Ice and, when compared, the results revealed the same indications as a predeveloped, established index, Jason's instrument would have convergent validity. Similarly, if Jason and Myra developed an instrument to measure satisfaction with the CompleteCare program and the derived measure could be confirmed with a standardized customer satisfaction measure, convergent validity would exist.

An example of factor analysis is described in Chapter 19.

Returning to our example above, another method of validating the ceremony construct would be to separate it from other constructs in the theory or related theories. To the extent that ceremony could be separated from stories or symbols, we would have completed the first steps toward *discriminant* validity. Established statistical tools such as factor analysis and multitrait-multimethod analysis help determine the construct adequacy of a measuring device.[22]

In the Prince Corporation study, you may be interested in securing a judgment of "how good a citizen" the corporation is. Variations in respondent ratings may be drastically affected if substantial differences exist among the respondents regarding what constitutes proper corporate citizenship. One respondent may believe that any company is an economic organization designed to make profits for its stockholders. She sees relatively little role for corporations in the wide-ranging social issues of the day. At the other end of the continuum, another respondent views the corporation as a leader in solving social problems, even at the cost of profits.

Both of these respondents might understand Prince's role in the community but judge it quite differently in light of their differing views about what its role should be. If these different views were held, you would theorize that other information about these respondents would be logically compatible with their judgments. You might expect the first respondent to oppose high corporate taxes, to be critical of increased involvement of government in family affairs, and to believe that a corporation's major responsibility is to its stockholders. The second respondent would be more likely to favor high corporate income taxes, to opt for more governmental involvement in daily life, and to believe that a corporation's major responsibility is a social one.

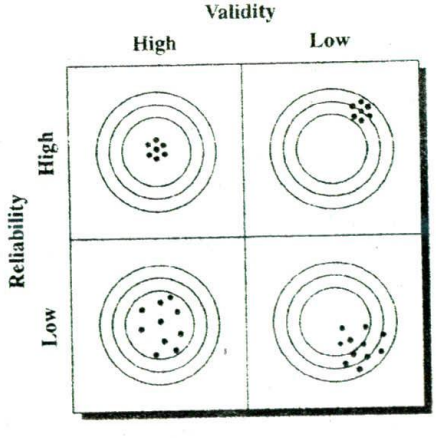
Respondents may not be consistent on all questions because the measurements may be crude and the "theory" may be deficient. When hypothesized tests do not confirm the measurement scale, you are faced with a two-sided question: Is your measurement instrument invalid, or is your theory invalid? These answers require more information or the exercise of judgment.

We discuss the three forms of validity separately, but they are interrelated, both theoretically and operationally. Predictive validity is important for a test designed to predict employee success. In developing such a test, you would probably first postulate the factors (constructs) that provide the basis for useful prediction. For example, you would advance a theory about the variable in employee success—an area for construct validity. Finally, in developing the specific items for inclusion in the success prediction test, you would be concerned with how well the specific items sample the full range of each construct (a matter of content validity).

In the corporate image study for the Prince Corporation, both content and construct validity considerations have been discussed, but what about criterion-related validity? The criteria are less obvious than in the employee success prediction, but judgments will be made of the quality of evidence about the company's image. The criteria used may be both subjective—Does the evidence agree with what we believe?—and objective—Does the evidence agree with other research findings?

Looking at Exhibit 8–5, we can approach the concepts of validity and reliability by using an archer's bow and target as an analogy. High reliability means that repeated arrows shot from the same bow would hit the target in essentially the same place—although not necessarily the intended place (first row of the graphic). If we had a bow

EXHIBIT 8–5
Understanding Validity and Reliability



with high validity as well, then every arrow would hit the bull's-eye (upper left panel). If reliability is low or decreases for some reason, arrows would be more scattered (lacking similarity or closeness like those shown in the second row)

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

Reliability

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

Reliability is concerned with estimates of the degree to which a measurement is free of random or unstable error. Reliable instruments can be used with confidence that transient and situational factors are not interfering. Reliable instruments are robust; they work well at different times under different conditions. This distinction of time and condition is the basis for frequently used perspectives on reliability—stability, equivalence, and internal consistency (see Exhibit 8–6).

SNAPSHOT

Surfing for the Perfect Measurement

When Web banner ads, and the newer superstitial ads or interactive marketing units (IMUs)—larger ads with voice and motion in pop-up windows similar to TV commercials—were first aired, they were heralded as the first truly measurable advertising medium. With a measurement called the *click-through rate*, advertisers could track the number of potential customers who saw an ad, clicked on the ad, arrived at the advertiser's website, and then bought a product online. But while Web advertising has grown faster than any other medium, to an estimated $5.4 billion, advertisers are no longer sure whether the click-through rate measures anything meaningful. "Click-through rates are a misleading statistic—they aren't indicative of raised awareness or of consumer interest," says Scot McLemon, who is executive vice president for ad sales at MarketWatch.com. While advertisers can compute cost measures (cost per click or cost per conversion to purchase from ad click), Web advertising has substantiated rather than disproved a well known adage: Response to advertising is often delayed and certainly not direct. Some firms, like MarketWatch.com, aren't counting click-through rates at all, hoping to persuade "the advertising industry to view online ads as tools for branding rather than direct marketing." Others, like Interactive Advertising Bureau (IAB), demonstrate that advertisers are resorting to tried and true, but far less exacting, ad effectiveness measures: conducting online surveys among those who recall having seen ads and measuring brand awareness, product association with message, purchase intent, and brand favorability. In its latest study by "Dynamic Logic, [IAB] found that the new larger ad units (IMUs) are 25 percent more effective in lifting key brand metrics such as brand awareness and message association—even at one exposure." What type of data and what measurement scales have firms involved with Web advertising been using to measure effectiveness? What should advertisers use?

www.cbs.marketwatch.com

www.dynamiclogic.com

www.iab.net

EXHIBIT 8–6 Summary of Reliability Estimates

Type	Coefficient	What Is Measured	Methods
Test-retest	Stability	Reliability of a test or instrument inferred from examinee scores. Same test is administered twice to same subjects over an interval of less than six months.	Correlation
Parallel forms	Equivalence	Degree to which alternative forms of the same measure produce same or similar results. Administered simultaneously or with a delay.	Correlation
		Interrater estimates of the similarity of judges' observations or scores.	
Split-half KR20 Cronbach's alpha	Internal consistency	Degree to which instrument items are homogeneous and reflect the same underlying construct(s).	Specialized correlational formulas

Maritz recognizes that understanding customers, employees, and other stakeholders starts with solid measurement even when the delivery is automated. www.maritz.com/mmr

Stability A measure is said to possess **stability** if you can secure consistent results with repeated measurements of the same person with the same instrument. An observational procedure is stable if it gives the same reading on a particular person when repeated one or more times. It is often possible to repeat observations on a subject and to compare them for consistency. When there is much time between measurements, there is a chance for situational factors to change, thereby affecting the observations. The change would appear incorrectly as a drop in the reliability of the measurement process.

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

MANAGEMENT Tip

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

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

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

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

The major interest with equivalence is typically not how respondents differ from item to item but how well a given set of items will categorize individuals. There may be

many differences in response between two samples of items, but if a person is classified the same way by each test, then the tests have good equivalence.

One tests for item sample equivalence by using alternative or parallel forms of the same test administered to the same persons simultaneously. The results of the two tests are then correlated. Under this condition, the length of the testing process is likely to affect the subjects' responses through fatigue, and the inferred reliability of the parallel form will be reduced accordingly. Some measurement theorists recommend an interval between the two tests to compensate for this problem. This approach, called *delayed equivalent forms*, is a composite of test-retest and the equivalence method. As in test-retest, one would administer form X followed by form Y to half the examinees and form Y followed by form X to the other half to prevent "order-of-presentation" effects.[23]

MANAGEMENT Tip

The researcher can include only a limited number of measurement questions in an instrument. This limitation implies that a sample of measurement questions from a content domain has been chosen and another sample producing a similar number will need to be drawn for the second instrument. It is frequently difficult to create this second set. Yet if the pool is initially large enough, the items may be randomly selected for each instrument. Even with more sophisticated procedures used by publishers of standardized tests, it is rare to find fully equivalent and interchangeable questions.[24]

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

The Spearman-Brown correction formula is used to adjust for the effect of test length and to estimate reliability of the whole test. A problem with this approach is that the way the test is split may influence the internal consistency coefficient. To remedy this, other indexes are used to secure reliability estimates without splitting the test's items. The *Kuder-Richardson Formula 20 (KR20)* and *Cronbach's coefficient alpha* are two frequently used examples. Cronbach's alpha has the most utility for multi-item scales at the interval level of measurement. The KR20 is the method from which alpha was generalized and is used to estimate reliability for dichotomous items (see Exhibit 8–6).

Improving Reliability The researcher can improve reliability by choosing among the following:

MANAGEMENT Tip

- Minimize external sources of variation.
- Standardize conditions under which measurement occurs.
- Improve investigator consistency by using only well-trained, supervised, and motivated persons to conduct the research.
- Broaden the sample of measurement questions used by adding similar questions to the data collection instrument or adding more observers or occasions to an observational study.
- Improve internal consistency of an instrument by excluding data from analysis drawn from measurement questions eliciting extreme responses. This approach

requires the assumption that a high total score reflects high performance and a low total score, low performance. One selects the extreme scorers—say, the top 20 percent and bottom 20 percent—for individual analysis. By this process, you can distinguish those items that differentiate high and low scorers. Items that have little discriminatory power can then be dropped from the test.

Practicality

The scientific requirements of a project call for the measurement process to be reliable and valid, while the operational requirements call for it to be practical. **Practicality** has been defined as *economy, convenience,* and *interpretability.*[25] While this definition refers to the development of educational and psychological tests, it is meaningful for business measurements as well.

Economy Some trade-off usually occurs between the ideal research project and the budget. Instrument length is one area where economic pressures dominate. More items give more reliability, but in the interest of limiting the interview or observation time (and therefore costs), we hold down the number of measurement questions. The choice of data collection method is also often dictated by economic factors. The rising cost of personal interviewing first led to an increased use of long-distance telephone surveys and subsequently to the current rise in online surveys. In standardized tests, the cost of test materials alone can be such a significant expense that it encourages multiple reuse. Add to this the need for fast and economical scoring, and we see why computer scoring and scanning are attractive.

MANAGEMENT Tip

Convenience A measuring device passes the convenience test if it is easy to administer. A questionnaire with a set of detailed but clear instructions, with examples, is easier to complete correctly than one that lacks these features. In a well-prepared study, it is not uncommon for the interviewer instructions to be several times longer than the interview questions. Naturally, the more complex the concepts, the greater is the need for clear and complete instructions. We can also make the instrument easier to administer by giving close attention to its design and layout. Crowding of material, poor reproductions of illustrations, and the carryover of items from one page to the next make completion of the instrument more difficult.

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

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

Close-Up

Earlier, Jason Henry agreed to send the executive director at White Ice some useful tools for measuring job satisfaction and motivation. In reviewing his files, he found a piece of research conducted at five geographically separate units of the Tennessee Valley Authority, three divisions of an electronics company, and five departments of an appliance manufacturing company. The procedure for developing the measures was first to hold a number of informal interviews with supervisory and nonsupervisory employees. From the knowledge acquired, the researchers constructed the questions. These were then pretested and revised twice on separate groups of TVA employees. Out of this process came the six-item questionnaire on interest in work innovation shown in Exhibit 8–7. This instrument and the others were completed by employees of the three companies. The reliability of the Interest in Work Innovation Index was measured by a test-retest of individual questions. The retest was done one month after the first test. Correlating the test-retest scores question by question gave the following results (see the Pearson correlation coefficient in Chapter 18 for more information on how these correlation coefficients were computed):

Question	*r*
Q1	.72
Q2	.72
Q3	.64
Q4	.67
Q5	.54
Q6	.85

EXHIBIT 8–7 Interest in Work Innovation Index*

1. In your kind of work, if a person tries to change his usual way of doing things, how does it generally turn out?
   (1) _____ Usually turns out worse; the tried and true methods work best in my work.
   (3) _____ Usually doesn't make much difference.
   (5) _____ Usually turns out better; our methods need improvement.
2. Some people prefer doing a job in pretty much the same way because this way they can count on always doing a good job. Others like to go out of their way in order to think up new ways of doing things. How is it with you on your job?
   (1) _____ I always prefer doing things pretty much in the same way.
   (2) _____ I mostly prefer doing things pretty much in the same way.
   (4) _____ I mostly prefer doing things in new and different ways.
   (5) _____ I always prefer doing things in new and different ways.
3. How often do you try out, on your own, a better or faster way of doing something on the job?
   (5) _____ Once a week or more often.
   (4) _____ Two or three times a month.
   (3) _____ About once a month.
   (2) _____ Every few months.
   (1) _____ Rarely or never.
4. How often do you get chances to try out your own ideas on the job, either before or after checking with your supervisor?
   (5) _____ Several times a week or more.
   (4) _____ About once a week.
   (3) _____ Several times a month.
   (2) _____ About once a month.
   (1) _____ Less than once a month.

EXHIBIT 8–7 Concluded

5. In my kind of job, it's usually better to let my supervisor worry about new or better ways of doing things.
   (1) _____ Strongly agree.
   (2) _____ Mostly agree.
   (4) _____ Mostly disagree.
   (5) _____ Strongly disagree.

6. How many times in the past year have you suggested to your supervisor a different or better way of doing something on the job?
   (1) _____ Never had occasion to do this during the past year.
   (2) _____ Once or twice.
   (3) _____ About three times.
   (4) _____ About five times.
   (5) _____ Six to ten times.
   (6) _____ More than ten times had occasion to do this during the past year.

*Numbers in parentheses preceding each response category indicate the score assigned to each response.
Source: Martin Patchen, *Some Questionnaire Measures of Employee Motivation and Morale*, Monograph No. 41 (Ann Arbor: Institute for Social Research, The University of Michigan, 1965), pp. 15–16.

The researchers measured criterion-based validity by comparing worker scores on the six questions to ratings of the same workers by their supervisors. Supervisors were asked to "think of specific instances where employees in their units had suggested new or better ways of doing the job. They then ranked employees they personally knew on 'looking out for new ideas.'"[26] The median correlation between the index scores and the supervisor ratings was about .35. At TVA, where there was an active suggestion system in operation, they also found that the index scores of those making suggestions were significantly higher than those not making suggestions.

Construct validity was evaluated by comparing scores on the Interest in Work Innovation Index to other job-related variables. Mean scores on the index were computed for 90 work groups at TVA. These means were then correlated with group scores on other variables that were hypothesized to relate to interest in innovation. The results are shown in Exhibit 8–8.

The researchers concluded, "The Index of Interest in Work Innovation, while a rough one, shows adequate reliability and sufficient evidence of validity to warrant its use in making rough distinctions among groups of people (or among units)."[27] In addition, they tested a short version of the index (items 1, 5, and 6) and found its validity to be almost equal to that of the longer form.

Having reviewed this research study with its derived indexes, Jason forwarded what he found to the symphony director. She would decide if it was adaptable or if she should develop her own instrument. Managers and researchers frequently assume they need a device tailored to their unique situation. This decision can be costly and time-consuming. Reliability testing may be ignored and validity assessments may be confined to impressions about content. Typically, there is no comparable evidence from other studies by which to calibrate the findings.

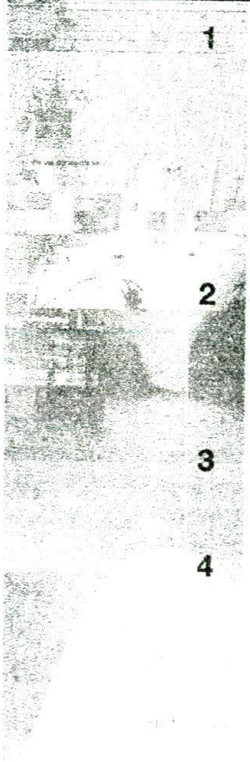
If Jason's recommendation proves to be inadequate, a further search of existing measures will reveal many established ones that might fit the director's needs. Most are copyrighted but available from commercial sources.

EXHIBIT 8–8 Relation of Scores on Interest in Work Innovation Index* to Scores on Other Job-Related Variables† for 90 Work Groups at TVA (Pearson product-moment correlation coefficient, *r*)

Correlation	Variable Name	Correlation	Variable Name
.44‡	Job difficulty	.05	Pressure from peers to do a good job
.39‡	Identification with own occupation	.36‡	General job motivation
.29‡	Control over work methods	.36‡	Willingness to disagree with supervisors
.28‡	Perceived opportunity for achievement	.12	Acceptance of changes in work situation
.19	Feedback on performance	.00	Identification with TVA
.13	Control over goals in work	.21¶	Overall satisfaction (with pay, promotion, supervisors, and peers)
.06	Need for achievement§		

*The shorter three-item Index B was used for these correlations.
†Variables listed are all indexes; each index is composed of several specific questions.
‡$p < .01$, 2-tailed *t*-test.
§This is the Achievement Risk Preference Scale developed by P. O'Connor and J. W. Atkinson (1960).
¶$p < .05$, 2-tailed t-test.
Source: Martin Patchen, *Some Questionnaire Measures of Employee Motivation and Morale*, Monograph No. 41 (Ann Arbor: Institute for Social Research, The University of Michigan, 1965), p. 24.

SUMMARY



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

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

3 Instruments may yield incorrect readings of an indicant for many reasons. These may be classified according to error sources: (1) the respondent or subject, (2) situational factors, (3) the measurer, and (4) the instrument.

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

most complex and abstract. A measure has construct validity to the degree that it conforms to predicted correlations of other theoretical propositions.

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

KEY TERMS

EXAMPLES

Company	Scenario	Page
American Demographics	Sponsored a study on the attitudes toward copyright infringement.	230
Burke CSA	A research company using measurement scales to provide companies with customer feedback.	227
Dynamic Logic	Research firm that studied ad effectiveness measures being used to evaluate Internet advertising.	236
Espace Van*	Measuring attendees' reactions at an auto show.	221
Interactive Advertising Bureau	A trade association's research reveals what ad sellers are using to measure ad effectiveness.	236
MarketWatch.com	An ad seller trying to determine the best way to evaluate ad effectiveness.	236
Peter D. Hart Research Associates, AFL-CIO	A study to determine motivating factors for retaining and recruiting workers in a tight job market; further analysis regarding young workers' interest in unions.	233
Prince Corporation*	A study to discover the public's opinions about the company and the origin of any generally held adverse opinions.	229
SalesPro*	A study to evaluate sales performance.	233
Society for Human Resource Management (SHRM)	A study to determine motivating factors for retaining and recruiting workers in a tight job market; further analysis regarding young workers' interest in unions.	233
Swatch Co.	The use of BeatTime as a ratio scale.	228
TaylorNelson Sofres (TNS) Intersearch	A study for *American Demographics* about adult attitudes related to copyright infringement, included in its special issue on privacy.	230
Tennessee Valley Authority	Instrument development; reliability and validity emphasis.	241

White Ice Summer Festival Orchestra*	A study of conditions influencing the rapid turnover of performers.	BRTL, Close-Up, throughout

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

DISCUSSION QUESTIONS

Terms in Review

1. What can we measure about the four objects listed below? Be as specific as possible.
   a. Laundry detergent
   b. Employees
   c. Factory output
   d. Job satisfaction
2. What are the essential differences among nominal, ordinal, interval, and ratio scales? How do these differences affect the statistical analysis techniques we can use?
3. What are the four major sources of measurement error? Illustrate by example how each of these might affect measurement results in a face-to-face interview situation.
4. Do you agree or disagree with the following statements? Explain.
   a. Validity is more critical to measurement than reliability.
   b. Content validity is the most difficult type of validity to determine.
   c. A valid measurement is reliable, but a reliable measurement may not be valid.
   d. Stability and equivalence are essentially the same thing.

Making Research Decisions

5. You have data from a corporation on the annual salary of each of its 200 employees.
   a. Illustrate how the data can be presented as ratio, interval, ordinal, and nominal data.
   b. Describe the successive loss of information as the presentation changes from ratio to nominal.
6. Below are listed some objects of varying degrees of abstraction. Suggest properties of each of these objects that can be measured by each of the four basic types of scales.
   a. Store customers.
   b. Voter attitudes.
   c. Hardness of steel alloys.
   d. Preference for a particular common stock.
   e. Profitability of various divisions in a company.
7. You have been asked by the head of marketing to design an instrument by which your private, for-profit school can evaluate the quality and value of its various curricula and courses. How might you try to ensure that your instrument has
   a. Stability?
   b. Equivalence?
   c. Internal consistency?
   d. Content validity?
   e. Predictive validity?
   f. Construct validity?

8. A new hire at Mobil Oil, you are asked to assume the management of the Mobil Restaurant Guide. Each restaurant striving to be included in the guide needs to be evaluated. Only a select few restaurants may earn the five-star status. What dimensions would you choose to measure to apply the one to five stars in the Mobil Restaurant Guide?
9. You have been asked to develop an index of student morale at your school.
   a. What constructs or concepts might you employ?
   b. Choose several of the major concepts and specify their dimensions.
   c. Select observable indicators that you might use to measure these dimensions.
   d. How would you compile these various dimensions into a single index?
   e. How would you judge the reliability and/or validity of these measurements?

From Concept to Practice

10. Using Exhibits 8–7 and 8–2, match each question to its appropriate data type. For each data type not represented, develop a measurement question that would obtain that type of data.

WWW Exercises

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

CASES*

A GEM OF A STUDY

CALLING UP ATTENDANCE

DATA DEVELOPMENT, INC.

NCR: TEEING UP A NEW STRATEGIC DIRECTION

PEBBLE BEACH CO.

RAMADA DEMONSTRATES ITS PERSONAL BEST

STATE FARM: DANGEROUS INTERSECTIONS

THE CATALYST FOR WOMEN IN FINANCIAL SERVICES

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

REFERENCE NOTES

1. Fred N. Kerlinger, *Foundations of Behavioral Research*, 3rd ed. (New York: Holt, Rinehart & Winston, 1986), p. 396; and S. Stevens, "Measurement, Statistics, and the Schemapiric View," *Science* (August 1968), p. 384.
2. W. S. Torgerson, *Theory and Method of Scaling* (New York: Wiley, 1958), p. 19.
3. We assume the reader has had an introductory statistics course in which measures of central tendency such as arithmetic mean, median, and mode have been treated. Similarly, we assume familiarity with measures of dispersion such as the standard deviation, range, and interquartile range. For a brief review of these concepts, refer to the Descriptive Statistics section in Chapter 15 or see an introductory statistics text.
4. While this might intuitively seem to be the case, consider that one might prefer *a* over *b*, *b* over *c*, yet *c* over *a*. These results cannot be scaled as ordinal data because there is apparently more than one dimension involved.
5. L. L. Thurstone, *The Measurement of Values* (Chicago: University of Chicago Press, 1959).
6. Parametric tests are appropriate when the measurement is interval or ratio and when we can accept certain assumptions about the underlying distributions of the data with which we are working. Nonparametric tests usually involve much weaker assumptions about measurement scales (nominal and ordinal), and the assumptions about the underlying distribution of the population are fewer and less restrictive. More on these tests is found in Chapters 17–19 and Appendix E.
7. Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences* (New York: McGraw-Hill, 1956), p. 32.
8. Norman A. Anderson, "Scales and Statistics: Parametric and Nonparametric," *Psychological Bulletin* 58, no. 4, pp. 315–16.

9. Kerlinger, *Foundations*, p. 403.
10. See Chapter 9 for a discussion of the differential scale.
11. See Chapters 17 and 18 for a discussion of these procedures.
12. To learn more about Swatch's BeatTime, visit: http://www.swatch.com/internettime/internettime.php3.
13. The exception involves the creation of a dummy variable for use in a regression or discriminant equation. A nonmetric variable is transformed into a metric variable through the assignment of a 0 or 1 and used in a predictive equation.
14. Claire Selltiz, Lawrence S. Wrightsman, and Stuart W. Cook, *Research Methods in Social Relations*, 3rd ed. (New York: Holt, Rinehart & Winston, 1976), pp. 164–69.
15. Robert L. Thorndike and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education*, 3rd ed. (New York: Wiley, 1969), p. 5.
16. Examples of other conceptualizations of validity are factorial validity, job-analytic validity, synthetic validity, rational validity, and statistical conclusion validity.
17. Thomas D. Cook and Donald T. Campbell, "The Design and Conduct of Quasi Experiments and True Experiments in Field Settings," in *Handbook of Industrial and Organizational Psychology*, ed. Marvin D. Dunnette (Chicago: Rand McNally, 1976), p. 223.
18. *Standards for Educational and Psychological Tests and Manuals* (Washington, DC: American Psychological Association, 1974), p. 26.
19. Wayne F. Cascio, *Applied Psychology in Personnel Management* (Reston, VA: Reston Publishing, 1982), p. 149.
20. Thorndike and Hagen, *Measurement and Evaluation*, p. 168.
21. See, for example, Cascio, *Applied Psychology*, pp. 146–47; and Edward G. Carmines and Richard A. Zeller, *Reliability and Validity Assessment* (Beverly Hills, CA: Sage Publications, 1979), pp. 48–50.
22. Emanuel J. Mason and William J. Bramble, *Understanding and Conducting Research* (New York: McGraw-Hill, 1989), pp. 260–63.
23. Cascio, *Applied Psychology*, pp. 135–36.
24. Mason and Bramble, *Understanding and Conducting Research*, p. 268.
25. Thorndike and Hagen, *Measurement and Evaluation*, p. 199.
26. Martin Patchen, *Some Questionnaire Measures of Employee Motivation and Morale*, Monograph No. 41 (Ann Arbor: Institute for Social Research, The University of Michigan, 1965), p. 17.
27. Ibid., p. 25.

REFERENCES FOR SNAPSHOTS AND CAPTIONS

Internet Advertising

"Interactive Advertising Bureau (IAB), DoubleClick, MSN, and CNET Networks Release Groundbreaking Online Brand Research Findings," IAB press release, July 18, 2001 (http://www.iab.net/news/content/brand_research.html).

Christopher Saunders, "Industry Players Seek to Distance Themselves from Click-Throughs," *InternetNews*, July 9, 2001 (http://www.internetnews.com/IAR/article/0,,12_797851,00.html).

Rob Walker, "System for Measuring Clicks Is Under Assault," *The New York Times on the Web*, August 27, 2001 (http://www.nytimes/2001/08/027/technology/27NECO.html?ex=1000139238&ei=1&en=c315615dca93ee07).

Interest in Unions

"Education, Collective Action and New Rules Can Make Things Better," Peter D. Hart Research Associates (AFL-CIO) to SHRM, May 1999 (http://www.aflcio.org/articles/high_hopes/4.htm).

Copyright Infringement

John Fetto, "Americans Voice Their Opinions on Intellectual Property Rights Violations," *American Demographics*, September 2000, p. 8.

Measurement instrument prepared by TaylorNelson Sofres Intersearch.

Data tabulation generated by TaylorNelson Sofres Intersearch

CLASSIC AND CONTEMPORARY READINGS

Cascio, Wayne F. *Applied Psychology in Personnel Management*. 4th ed. Englewood Cliffs, NJ: Prentice-Hall, 1990.

Cook, Thomas D., and Donald T. Campbell. "The Design and Conduct of Quasi Experiments and True Experiments in Field Settings." In *Handbook of Industrial and Organizational Psychology*, ed. Marvin D. Dunnette. Chicago: Rand McNally, 1976, Chapter 7.

Embretson, Susan E., and Scott L. Hershberger. *The New Rules of Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates, 1999. Bridges the gap between theoretical and practical measurement.

Guilford, J. P. *Psychometric Methods*. 2nd ed. New York: McGraw-Hill, 1954.

Kelley, D. Lynn. *Measurement Made Accessible: A Research Approach Using Qualitative, Quantitative, and TQM Methods*. Thousand Oaks, CA: Sage Publications 1999. Sections on bias, reliability, and validity are appropriate for this chapter.

Kerlinger, Fred N., and Howard B. Lee. *Foundations of Behavioral Research*. 4th ed. New York: HBJ College & School Division, 1999.

Newmark, Charles S. *Major Psychological Assessment Instruments*. 2nd ed. Boston: Allyn and Bacon, 1996.

Nunnally, J. C., and Ira Bernstein. *Psychometric Theory*. 3rd ed. New York: McGraw-Hill, 1994.

Thorndike, Robert M. *Measurement and Evaluation in Psychology and Education*. 6th ed. Upper Saddle River, NJ: Prentice-Hall, 1996.

CHAPTER 9

Measurement Scales

Learning Objectives

After reading this chapter, you should understand . . .

1. **The six critical decisions involved in selecting an appropriate measurement scale.**
2. **The various scale formats for measurement and how to construct each.**
3. **The five ways that measurement scales are constructed.**