PART ONE

INTRODUCTION

The main purpose of Part One (a single chapter in this case) is to define "measurement" in terms of two fairly simple concepts: Measurement consists of rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification). Most of the book is concerned with the first of these meanings. The topics of levels of scaling and the general standards by which measurement rules are evaluated are focal issues.

CHAPTER 1

INTRODUCTION

CHAPTER OVERVIEW

This opening chapter begins with a definition of measurement which we break down into two subtopics: scaling and classification. Some general properties of good measurement are introduced, and the importance of standardization is discussed. The separate roles of measurement and pure mathematics are contrasted. One major, and still controversial, topic in measurement concerns what are known as levels of measurement. According to some, the appropriate level of a measure must be established before employing mathematical and statistical procedures associated with that level. Many look for ostensive (visualizable) properties of measures like the yardsticks and clocks of physics. They view present scales as imperfect correlates of unknown "true" scales. We attempt to show that these strategies easily lead to unreasonable outcomes. One should demonstrate that a measure has the properties ascribed to it, establish scales by convention, but be prepared to change these conventions as better measures become available. The chapter concludes by noting some of the changes brought to the study of measurement that result from the availability of computers.

MEASUREMENT IN SCIENCE

Although tomes have been written on the nature of measurement, in the end it boils down to two fairly simple concepts: "measurement" consists of rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification). Most of what is historically called measurement involves scaling, and therefore properties of numbers, but classification can be equally

important. The objects in psychology are usually people, but they may be lower animals as in some areas of psychology and biology or physical objects as in some market research. The term "rules" indicates that the assignment of numbers must be explicitly stated. Some rules are so obvious that detailed definition is unnecessary, as in measuring height with a tape measure. Unfortunately, these obvious cases are exceptional in science. For instance, assaying a chemical compound usually requires extremely complex procedures. Certainly the rules for measuring most attributes such as intelligence, shyness, or priming are not intuitively obvious.

Rules, in turn, are an important aspect of standardization. A measure is standardized to the extent that (1) its rules are clear, (2) it is practical to apply, (3) it does not demand great skill of administrators beyond that necessary for their initial training, and (4) its results do not depend upon the specific administrator. The basic point about standardization is that users of a given instrument should obtain similar results. The results must therefore be reliable in a sense to be discussed at several points in this book. Thus, measuring the surface temperature of a planet is well standardized if different astronomers obtain very similar estimates. Similarly, an intelligence test is well standardized if different examiners obtain similar scores from testing a particular child at a given time.

The term "attribute" in the definition indicates that measurement always concerns some *particular* feature of objects. One cannot measure objects—one measures their attributes. One does not measure a child per se, but rather his or her intelligence, height, or socialization. The distinction between an object and its attributes may sound like mere hairsplitting, but it is important. First, it demonstrates that measurement requires a process of abstraction. An attribute concerns relations among objects on a particular dimension, e.g., weight or intelligence. A red rock and a white rock may weigh the same, and two white rocks may have different weights. The attributes of weight and color must not be confounded with each other nor with any other attributes. It is quite easy to confuse a particular attribute of objects with other attributes. For example, some people find it difficult to understand that a criminal and a law-abiding citizen can both be equally smart. Failing to abstract a particular attribute from the whole makes the concept of measurement difficult to grasp.

A second reason for emphasizing that one measures attributes and not objects is that it makes us consider the nature of an attribute carefully before attempting measurement. An attribute we believe in may not exist in the form proposed. For example, the many negative results obtained in the efforts to measure an overall attribute of rigidity make it debatable that such an attribute exists. Even highly popular terms used to describe people may not correspond to measurable attributes, e.g., clairvoyance. It is also common for an assumed unitary attribute to confound several more specific attributes. For example, "adjustment" may include satisfaction with one's life, positive mood, skills in coping with stress, and other meanings of the term. Although such conglomerate measures may be partly justifiable on practical grounds, their use can undermine psychological science. As this book will show in detail, a measure should generally concern some one *thing*—some distinct, unitary attribute. To the extent that unitary attributes need be combined in an overall appraisal, e.g., of adjustment, they should usually be rationally combined from different measures rather than being confounded within one measure.

The first part of the definition of measurement stresses the use of numbers to represent quantities in scaling Technically, quantification concerns how much of an attribute is present in an object, and numbers communicate the amount. Quantification is so intimately intertwined with measurement that the two terms are often used interchangeably. This is unfortunate, as the second part, classification, is at least as important to science.

Although the definition emphasizes that rules are at the heart of measurement, it does not specify the nature of these rules or place any limit on the allowable kinds of rules. This is because a clear distinction must be made between measurement as a process and the standards for validating measures. The measurement process involves such considerations as the levels-of-measurement issue that is discussed later in this chapter. Validation involves issues that are discussed in Chapter 3. Numerous standards can be applied to obtain the usefulness of a measurement method, including the extent to which data obtained from the method (1) fit a mathematical model, (2) measure a single attribute, (3) are repeatable over time if necessary, (4) are valid in various senses, and (5) produce interesting relationships with other scientific measures. Such standards will be discussed throughout this book. Thus, a psychologist might establish rules to measure, say, dogmatism, in a manner that seems quite illogical to other psychologists, but the measure's usefulness cannot be dismissed beforehand.

The rules employed to define a particular measure must be unambiguous. They may be developed from an elaborate deductive model, based on previous experience, flow from common sense, or simply spring from hunches, but the crucial point is how consistently users agree on the measure and ultimately how well the measurement method explains important phenomena. Consequently any set of rules that unambiguously quantifies properties of objects constitutes a legitimate measurement method and has a right to compete with other measures for scientific usefulness. Keep in mind, however, that clarity does not guarantee explanatory power.

What Is "Meaningful" and "Useful"?

There is both agreement and disagreement among scientists about what is a meaningful and/or useful result. It is fair to say that there is a high degree of agreement on two points. One is that any result should be *repeatable* under similar circumstances. It is quite possible that a finding obtained on April 8, 1991, from a particular group of psychology students at the University of Texas at Arlington was a real effect descriptive of that group of people. However, unless that effect also applied to some other group, e.g., students at the University of Texas at Arlington tested on another day or at some other university on the same day, there is no need for a scientist to be concerned with it.

The second point of agreement that all scientists have learned is that any set of results can be understood after the fact even if it is a chance occurrence or even systematically wrong. Perhaps every investigator has analyzed a set of results and formulated an explanation only to discover that there was a "bug" in the analysis. That bug probably did not hamper a "creative" explanation of the wrong results. In a like manner, some of the more sadistic instructors we have known assign randomly generated results to students for explanation. Students often find the exercise creative until they are let on.

The keys to meaningfulness are to proceed from some position that *anticipates* results. This is where scientists differ. Some are strongly biased toward testing hypotheses derived from highly formalized theories; others are more informal and/or results-oriented in their approach. For a debate on this issue, see Greenwald, Pratkanis, Leippe, and Baumgardner (1986) and a series of commentaries that appeared in the October 1988 issue of *Psychological Review.* As of this writing, the pendulum seems to have swung in a more formal direction, at least in cognitive psychology, but it probably will swing back. Whatever the level of formality preferred, meaningfulness depends upon context. One of the most common phrases one hears about results is "So what?" The answer lies in placing findings in a relevant context.

This is not to rule out unanticipated findings, which are always an exciting part of science. However, before one becomes too enraptured by an interpretation given a set of findings, one should be prepared to replicate them, preferably in some way that broadens their generality.

ADVANTAGES OF STANDARDIZED MEASURES

Although you may already have a healthy respect for the importance of measurement in science, it is useful to look at some particular advantages that measurement provides. To note these advantages, consider what would be left if no measures were available, e.g., if there were no thermometers or intelligence tests. Measures based upon well-developed rules, usually including some form of norms that describe the scores obtained in populations of interest, are called "standardized." Despite criticisms of standardized psychological tests, the decisions that these are used for would still be made. What would be left would consist of subjective appraisals, personal judgments, etc. Some of the advantages of standardized measures over personal judgments are as follows:

Objectivity

The major advantage of measurement is in taking the guesswork out of scientific observation. A key principle of science is that any statement of fact made by one scientist should be independently verifiable by other scientists. The principle is violated if scientists can disagree about the measure. For example, since there is no standardized measure of "libidinal energy," two psychologists could disagree widely about a patient's libidinal energy. It is obviously difficult to test theories of libidinal energy until it can be measured.

One could well argue that measurement is *the* major problem in psychology. There are many theories, but a theory can be tested only to the extent that its hypothesized attributes can be adequately measured. This has historically been the problem with Freudian theory: There are no agreed-on procedures for observing and quantifying such attributes as libidinal energy, etc. Major advances in psychology, if not all sciences, are often based upon breakthroughs in measurement. Consider, for example, the flood of research stimulated by the development of intelligence tests and of personality tests like the Minnesota Multiphasic Personality Inventory (MMPI), or, in a very

different area, the development of techniques to record from single neurons (Hartline, 1940; Kuffler, 1953). Scientific results inevitably involve functional relations among measured variables, and the science of psychology can progress no faster than the measurement of its key variables.

Quantification

The numerical results provided by standardized measures have two advantages. First, numerical indices can be reported in finer detail than personal judgments, allowing more subtle effects to be noted. Thus the availability of thermometers makes it possible to report the exact increase in temperature when two chemicals are mixed, rather than for the investigator to intuitively judge only that "the temperature increases." Similarly, teachers may be able to reliably assign children to broad categories of intelligence such as bright, average, and below normal, but intelligence tests provide finer differentiations.

Second, quantification permits the use of more powerful methods of mathematical analysis that are often essential to the elaboration of theories and the analysis of experiments. Although important psychological theories need not be highly quantitative, the trend is and will continue to be clearly in that direction. Mathematically statable theories make precise deductions possible for empirical investigation. Also, other mathematical models and tools, such as factor analysis and the analysis of variance (ANOVA), may be used to analyze various results even when the study does not test any formal theory.

Communication

Science is a highly public enterprise requiring efficient communication among scientists. Scientists build on the past, and their findings must be compared with results of other scientists working on the same problem. Communication is greatly facilitated when standardized measures are available. Suppose, for example, it is reported that a particular treatment made the subjects "appear anxious" in an experiment concerning the effects of stress on anxiety reaction. This leaves many questions as to what the experimenter meant by "appear anxious," and makes it difficult for other experimenters to investigate the same effect. Much better communication could be achieved if the anxiety measure were standardized, as the means and standard deviations of these scores could be compared across treatment groups. Even very careful subjective evaluations are much more difficult to communicate than statistical analyses of standardized measures.

Economy

Although standardized measures frequently require a great deal of work to develop, they generally are much more economical of time and money than are subjective evaluations after they have been developed. For example, even the best judges of intelligence need to observe a child for some time. At least as good an appraisal can usually

be obtained in less than an hour with any of several inexpensively administered group measures of intelligence. Similarly, one can use a standardized activity measure such as rate of bar pressing in a Skinner box to evaluate the effect of a proposed stimulant on animals.

Besides saving time and money, standardized measures often free professionals for more important work. Progress generally favors measures that either require relatively little effort to employ or allow less highly trained technicians to do the administration and scoring. The time saved allows practitioners and scientists more time for the more scholarly and creative aspects of their work.

It is sometimes difficult to disentangle the measurer from the measurement process, as in individually administered intelligence tests. Although individual intelligence tests are highly standardized, they still require much time to administer and score. Context determines whether there are sufficient advantages to compensate for these disadvantages over even more highly standardized pencil-and-paper tests.

Scientific Generalization

Scientific generalization is at the very heart of scientific work. Most observations involve particular events—a "falling" star, a baby crying, a feeling of pain from a pin scratch, or a friend remarking about the weather. Science seeks to find underlying order in these particular events by formulating and testing hypotheses of a more general nature. The most widely known examples are the principles of gravitation, heat, and states of gases in physics. Theories, including those in the behavioral sciences, are intended to be general and thereby explain a large number of phenomena with a small, simple set of principles.

Many scientific generalizations, particularly in the behavioral sciences, must be stated in statistical terms. They deal with the probability of an event occurring and cannot be specified with more exactness. The development and use of standardized measurement methods are just as essential to probabilistic relationships as they are for deterministic ones. Figure 1-1 illustrates a simple probabilistic relationship noted by the first author between the complexity of randomly generated geometric forms and the amount of time that subjects looked at the forms. The data are group averages and are much more regular than individual subject data. However, the principle seems clear: People look longer at more complex figures than at simpler figures; but this would have been much less apparent in the data of individual subjects.

MEASUREMENT AND MATHEMATICS

A clear distinction needs be made between measurement, which is directly concerned with the real world, and mathematics, which, as an abstract enterprise, needs have nothing to do with the real world. Perhaps the two would not be so readily confused if both did not frequently involve numbers. Measurement always concerns numbers relatable to the physical world, and the legitimacy of any measurement is determined by data (facts about the physical world). In particular, scaling, but not classification, always concerns some form of numerical statement of *how much* of an attribute is

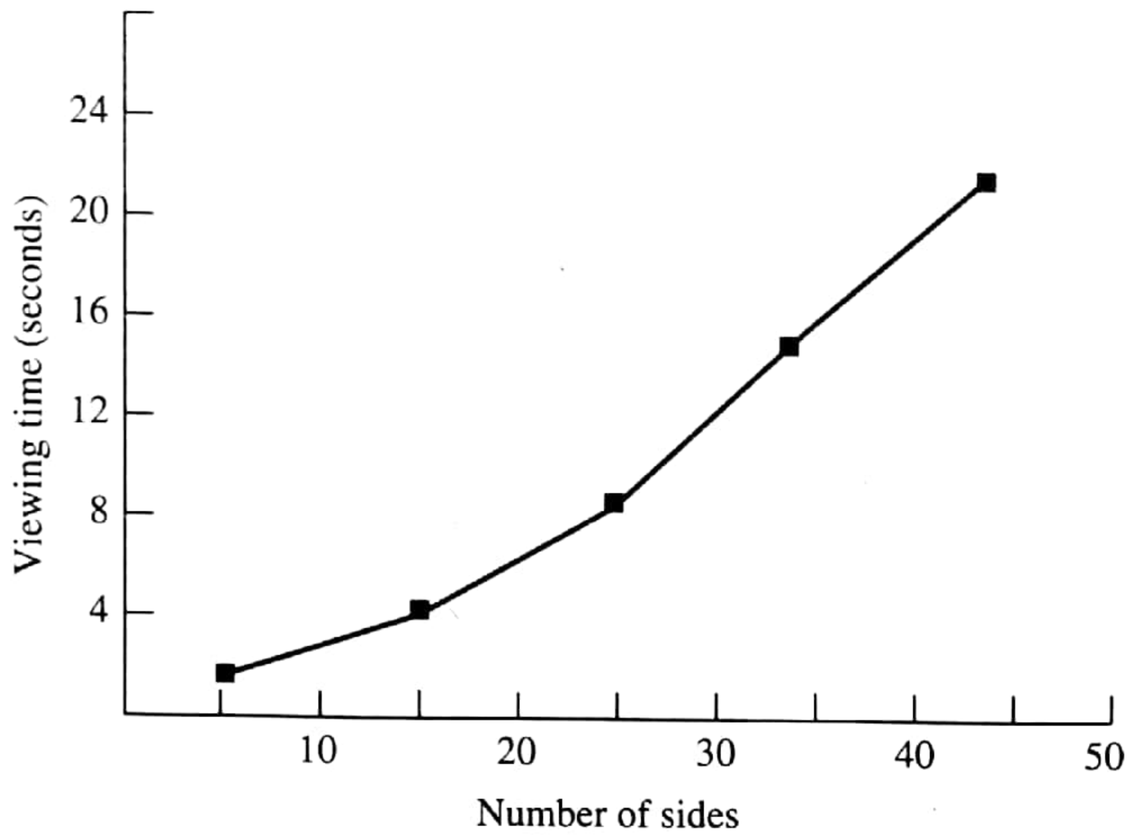

FIGURE 1-1 Viewing time as a function of stimulus complexity (number of sides on randomly generated geometric forms).

present, as its purpose is to quantify the attributes of real objects. A measure may be intended to fit a set of measurement axioms (a model), but its fit to the model can be determined only by seeing how well the data fit the model's predictions. Even if there is no formal model, the eventual and crucial test of any measure (scale or classification) is how well it explains relations among variables. As will be discussed in Chapter 3, the various types of validity for psychological measures all require data rather than purely mathematical deductions.

In contrast to measurement, pure mathematics is limited to deductive sets of rules for the manipulation of symbols, of which those used to denote quantities and categories are only one type. Many deductive systems in modern mathematics do not involve numbers, though they may involve classification. Any internally consistent set of rules for manipulating a set of symbols can be a legitimate branch of mathematics. Thus the statement "iggle wug drang flous" could be a legitimate mathematical statement in a set of rules stating that when any iggle is wugged it drang a flous. Mathematical systems could be constructed in which both the objects and the operations are symbolized by nonsense words. This system might not and need not be of practical use, as its legitimacy depends entirely on the internal consistency of its rules.

As a result, scientists *develop* measures by stating rules to quantify attributes of real objects, but *borrow* mathematical systems to examine the structure of the data. Fortunately scientifically useful measurement methods can usually be associated with appropriate mathematical systems.

Measurement and Statistics

Because the term "statistics" is used broadly, some distinctions among different uses of the term are necessary in order to see their implications for psychometric theory.

There is a basic distinction between descriptive and inferential statistics. "Descriptive statistics" concerns quantitative statements about an attribute of a particular group

of observations and does not necessarily imply generalization. Thus, one may compute the arithmetic mean of the scores on a classroom test, the correlation between two presumed measures of anxiety, or the scores of two job applicants without making any broader statements about those not taking the tests. In contrast, "inferential statistics" concerns generalizing from observed sample values (statistics) to their counterparts in a population (parameters), nearly always in the form of probability statements. A common example is to estimate the probability that the observed mean difference between an experimental group and a control group is a chance departure from 0, the expected result if the treatment had no effect.

We will say less in this book about inference than description, as most of the traditional quantitative methods to be presented are primarily designed for description rather than inference. Thus correlational analysis, factor analysis, discriminant analysis, and other procedures can be discussed and employed with minimal use of inference. This is not to say that inferential statistics are unimportant or that they will be totally neglected. We will consider some advances in inferential statistics that have become prominent since the last revision, particularly maximum likelihood estimation. There are three reasons to emphasize description. First, classical psychometric theory and some newer models are large-sample theories that assume that many subjects are studied. Second, even some investigators who have been very concerned with developing these newer inferential measurement models stress the importance of description (Bentler & Bonnett, 1980). Finally, we have enough material to present without going too far into a somewhat ancillary topic. There are excellent books on the relevant inferential statistics for psychometric theory that will be referenced where appropriate.

A second important statistical distinction is that between the sampling of objects (in this context, usually people) and the sampling of content (items). After a measure has been developed, it is often important to make statements about objects as in developing test norms. Before measures are developed, however, measurement is much more closely related to the sampling of content, as in deciding which test items to include. We will later stress how it is useful to think of particular test items as a sample from a hypothetical infinite population or universe of items measuring the same trait. Thus a spelling test for fourth-grade students can be thought of as a sample of all possible appropriate words. Part of measurement theory thus concerns statistical relations between the actual test scores and the hypothetical scores that would be made if all items in the universe had been administered.

There is a two-way problem in all psychology concerned with the sampling of objects to be measured and the sampling of content. The former usually concerns the generality of findings over objects, and the latter concerns the generality of findings over test items. Some item response theory models (Chapters 2 and 10) simultaneously take objects and items into account. However, most analyses take only one of these dimensions into account explicitly and keep the other in mind or, worse, simply ignore it. Thus, a study comparing different approaches to teaching mathematics upon a particular achievement test may explicitly concern gender differences. However, it might have to acknowledge that different results might have been obtained with different achievement measures.

The frequent necessity of considering only one of these two dimensions is not ideal, but it is not necessarily fatal. Subsequent studies can deal with generalizing over

the other dimension. The most desirable situation is when one samples so extensively on one dimension that the only sampling error present is on the other dimension. This normally requires an extremely large sample of subjects. At least hundreds, if not thousands, of subjects should be used in the development process. Except as noted, we will assume that all mathematical analyses are based on large numbers of subjects so that issues will be limited to the sampling of content. Studies conducted on relatively small numbers of subjects are usually not sufficient. Thus, even though a few dozen subjects may suffice to establish that the test reliability is greater than zero, a more precise statement of the magnitude is nearly always required.

The idea that sampling content is more important than sampling objects in developing a measure is not easy to grasp. Many students fall into the trap of assuming that a test's reliability increases with the number of objects (subjects) used in the study of reliability, when in fact it is directly related to the number of items on the test and independent of the number of objects.

MEASUREMENT SCALES

A series of articles by Stevens (1946, 1951, 1958, 1960) evoked considerable discussion and soul searching about the different possible types of measurement scales. Stevens proposed that measurements fall into four major classes (some extensions of these basic types will be noted below): nominal, ordinal, interval, and ratio. The levels allow progressively more sophisticated quantitative procedures to be performed on the measures but in turn demand progressively more of the measurement operations. In addition, the levels restrict the transformations possible upon the data. Table 1-1 provides an illustration of this proposed classification which we will embellish on in the succeeding pages.

Stevens' work evoked a great deal of controversy at the time, some of which continues. One major effect was that it led to a healthy self-consciousness about

TABLE 1-1 STEVENS' LEVELS OF MEASUREMENT, BASIC DEFINING OPERATIONS, PERMISSIBLE TRANSFORMATIONS, EXAMPLES OF PERMISSIBLE STATISTICS, AND EXAMPLES

Scale	Basic operation	Permissible transformations	Permissible statistics	Examples
Nominal	= vs ≠ (equality vs. inequality)	Any one-to-one	Numbers of cases, mode	Telephone numbers
Ordinal	> vs. < (greater than vs. less than)	Monotonically increasing	Median, percentiles, order statistics	Hardness of minerals, class rank
Interval	Equality of intervals or differences	General linear $x' = bx + a$	Arithmetic mean, variance, Pearson correlation	Temperature (Celsius), conventional test scores (?)
Ratio	Equality of ratios	Multiplicative (similarity) $x' = bx$	Geometric mean	Temperature (Kelvin)

Source: Adapted from Stevens (1951) by permission of John Wiley, Inc.

psychological measurement, but it also led to some unfortunate conclusions about the legitimacy of employing particular classes of mathematical procedures with measures of psychological attributes. Of these, the issue of whether or not it is meaningful to compute the mean of a series of test scores derived by summing individual items had the greatest implications. We will first present Stevens' position in a simplified, conventional manner, after which we will discuss the nature of psychological measurement in more general terms.

Nominal scales

Nominal scales contain rules for deciding whether two objects are = (equivalent) or ≠ (not equivalent), i.e., for categorizing. Equivalence means that two objects have a critical property in common, e.g., two people are both females. It does not imply identity or equality with respect to all relevant properties, and it will be discussed in a more formal sense below. The result of a nominal scale is a series of classes which *may* be given a numeric designation. The numbers are frequently used to keep track of things, without implying that they can be subjected to any mathematical analysis. Telephone and social security numbers are common examples of using numbers simply as labels that could just as well be expressed without numbers. These labels have no mathematical properties, and so it makes no sense to average a work and a home telephone number. However, it is important to distinguish between using the category "names" numerically, which is improper, and the category "frequencies," which is quite proper, e.g., to ask whether there are more Democrats, Independents, or Republicans in a political poll.

It is sometimes useful to distinguish between labels and categories even though both can be nominal scales. Labels, numeric or otherwise, are used to identify individual objects. These may be unique, as are the social security numbers given to U.S. citizens and residents, or there may be many duplications, as with given names. In contrast, categories are groupings of objects, in which it is usually desirable to have relatively few categories compared to the number of objects. Common categories are race, ethnicity, and gender.

Although categories and labels need not reflect any specific quantitative relationship, they may lead to the discovery of important correlates. For example, the finding that people of a certain ethnicity are more prone to a particular disease than people of a different ethnicity is vital to geneticists. However, this is an issue of classification, discussed below and in Chapter 15, and not scaling. Labels and categories are nominal scales, but nominal scales have thus far offered little to formal scaling models even though such models exist.

Nominal scales can be transformed in any manner that does not assign the same number to different categories. Thus, males and females could, respectively, be coded 1 and 0, 0 and 1 or even −257.3 and 534.8 without gain or loss of information. These one-to-one transformations are permissible because the names do not have numeric properties. The flexibility with which one can transform nominal scales reflects the limited mathematical operations that can be performed with them. For example, assume that a survey has coded potential voters as 1, 2, or 3 for Democrat, Republican, and Independent and that the frequencies of individuals in these three classes are 35,

25, and 40. One could compute a "mean" as $(35 \cdot 1 + 25 \cdot 2 + 40 \cdot 3)/100$ or 2.05. However, this figure would change capriciously if permissible transformations were made upon the categories. For example, it would change to 2.95 if Independents were coded 0, Democrats were coded 2, and Republicans were coded 9, and there is no logical connection between changes in the scale values and changes in this mean. One important exception to this principle is when there are two categories. This exception underlies much contemporary multiple regression theory, as we will see later in this book. In this case, statistics such as means do change predictably as categories are changed. We will show why this is the case when we consider interval scales.

Ordinal scales

Ordinal scaling involves rules for deciding whether one object that is $\neq$ to another is $>$ (greater than) or $<$ (less than) with respect to a given attribute (there may also be ties so $\leq$ and $\geq$ are also used). A *ordinal scale* for N persons (Ss) allows one to determine that $S_i \geq S_j \geq S_k \geq S_n$ with respect to an attribute (the = part of $\geq$ allows for ties). This implies that (1) a set of objects is ordered from "most" to "least" with respect to an attribute, (2) one does not know how much any of the objects possess of the attribute in an absolute sense, and (3) one does not know how far apart the objects are with respect to the attribute. An ordinal scale is obtained if a group of people are ranked from tallest to shortest. This scale gives no indication of the average height. The mean rank of the height of N jockeys and N professional basketball players will be $(N + 1)/2$. In both cases, the mean of five ranked observations will thus be $(5 + 1)/2$ or 3. Likewise, the variance of the ranks will equal $(N^2 - 1)/12$ regardless of whether the measures are very similar or very dissimilar. If there are five ranked observations, the result will be $(5^2 - 1)/12$ or 2.

Dichotomous (pass-fail) scoring is a special and, indeed, the simplest case of ordering. It is commonly present in true-false or multiple-choice ability tests. A pass is commonly designated 1, and a failure is designated 0. Items using an agree-disagree format in personality or attitude measurement logically also yield pass-fail orderings, since agreeing with the key is a form of passing.

Ordered categories arise when a measure yields relatively precise information, but the investigator lumps scores into a smaller number of successive categories. For example, an economist may categorize family income measures into a small number of levels. This can sacrifice a great deal of information, but it may be needed for data presentation. In contrast, data may be gathered as ranks. Likert scale items are a common example used in personality and attitude measurement in which subjects describe their intensity of feeling toward the item. For example, subjects might be asked whether they "strongly agree," "agree," "are indifferent," "disagree," or "strongly disagree" with the statement "I feel uncomfortable asking professors questions in class." The subject is then assigned a score from 1 to 5, and the total scale score is the sum of individual item scores. This format generates more information than dichotomous scoring, as it may increase the range of scores substantially over dichotomous items scoring, a benefit to the statistical analysis as it more faithfully reflects the individual differences on the attribute.

Rank ordering is basic to higher forms of measurement. Most of the information contained in higher level scales is contained simply in the rank orderings (Coombs, 1964; Parker, Casey, Ziriax, & Silberberg, 1988). Thus, if two sets of measures obtained from higher level scales are correlated and converted to ranks, and the ranked data also correlated (see Spearman's rank order correlation in Chapter 4), the correlation between the original numbers and the correlation between the ranks are usually quite similar in magnitude. In contrast, considerable information is lost if both sets of observations and correlations become much smaller when data are dichotomized. Consequently, methods based upon rank ordering, such as rank order multidimensional scaling considered in Chapter 14, often do justice to the relations contained in higher-level data, but the common practice of dichotomizing variables when the underlying data are of a stronger form should be avoided (Cohen, 1990).

The class of transformations permissible for ordinal scales is more limited than it is for nominal scales. The transformation must preserve the rank-order properties of the data. Thus, category names 1, 2, and 3 may be transformed to 4, 5, and 23 or −1.3, 2.05, and 5.33, but not 3, 1, and 2. These permissible transformations are called "monotonic" and are illustrated in Fig. 1-2. A set of statistical operations has been designed for use with ordinal data. The central tendency may be described in terms of the median or the mode (which is also meaningful with nominal data) rather than the arithmetic mean. The median and mode will change predictably with permissible transformations, whereas the mean will not. For example, if the median and mode are in the second of four ordinal categories coded from 1 to 4, they will remain so under any permissible transformation, which is not true of the arithmetic mean. A considerably different mean will obtain if the categories are recoded as 2, 4, 17, and 39, for example, but the median and mode simply change to the second category, 4.

Interval scales

Interval scales reflect operations that define a unit of measurement as well as >, =, and <. They are often referred to as "equal interval scales" for this reason. Consequently (1) the rank ordering of objects on an attribute is known, (2) the distances among objects on the attribute are also known, but (3) the *absolute* magnitudes of the attribute are unknown. Expressing the height of each of a series of children relative to their mean height would yield an interval scale of their height. Thus a child 2 inches taller than average would receive a score of +2, a child 3 inches shorter than average would receive a score of −3, etc. Deviations from any mean can be calculated without actually knowing how far anyone is from a true zero point, e.g., zero height. The absolute magnitudes of the attribute are potentially important but unknown since the tallest child is probably short in a more general sense. However, psychological measures are commonly described as deviations from the mean.

Interval scales do not require an equal number of objects (people) at each point, i.e., a rectangular distribution of scores. The term "equal" describes the intervals on the scale, not the number of people between equally spaced points on the scale. Thus, the difference between intelligence measures of 100 and 105 are assumed equal to the difference between intelligence measures of 120 and 125 even though many more people fall between 100 to 105 than 120 to 125.

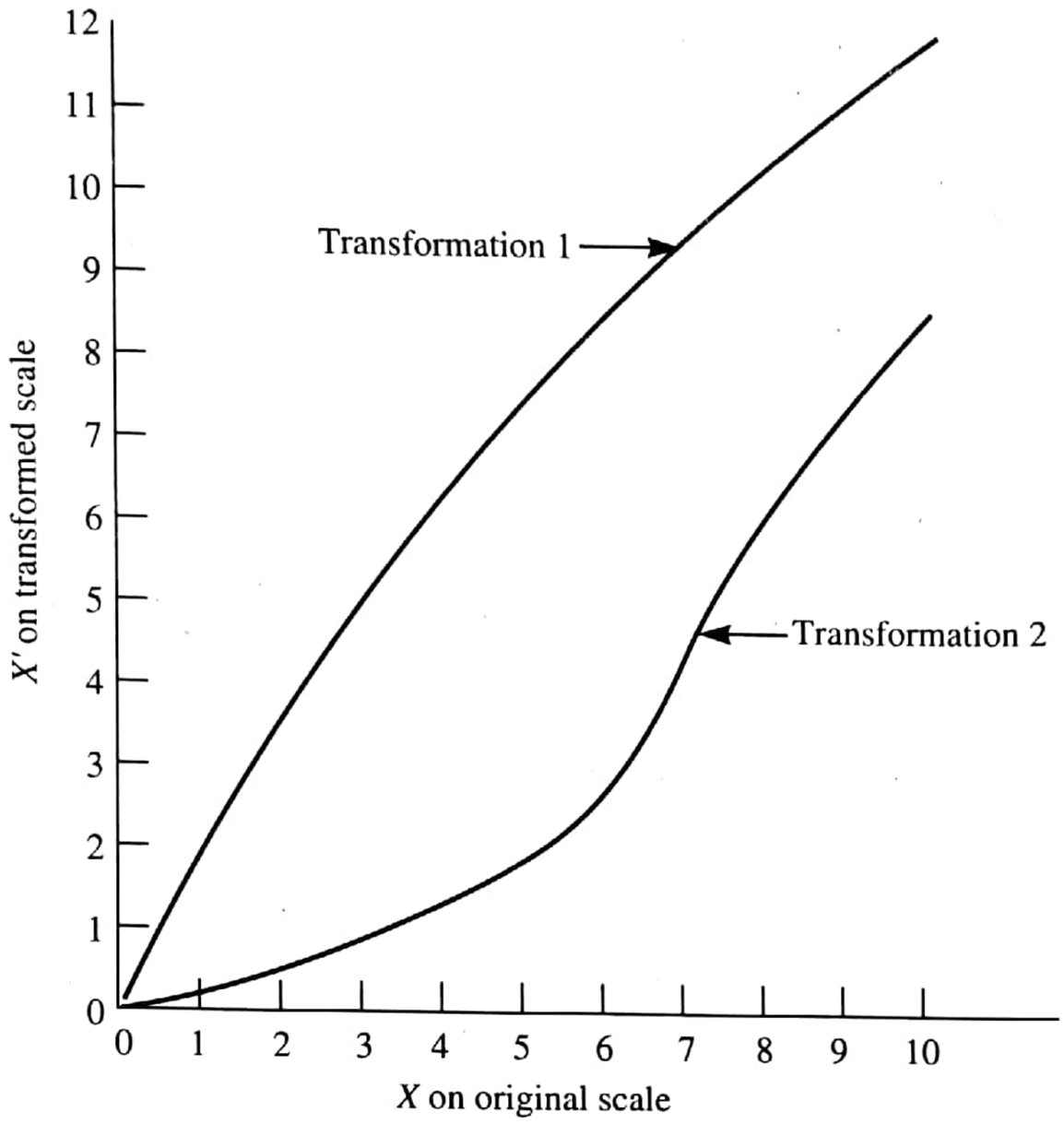

FIGURE 1-2 Two examples of monotonic transformations permissible on an ordinal scale. The general form of these transformations is difficult to define algebraically.

Interval properties imply that if $a, b, c, \cdots, k$ are equally spaced points on the scale, the scale is defined by two statements:

1 $a > b > c > \cdots > k$
2 $a - b = b - c = c - d = \cdots = j - k$

An interval scale is defined by algebraic differences between points, and so addition and subtraction of the scale points are permissible operations. Since $a - b = b - c$, the sum of the two intervals equals $(a - b) + (b - c) = a - c$.

The difference between the two intervals equals zero:

$$(a - b) - (b - c) = a - 2b + c$$

The expression equals zero because $a + c = 2b$:

$$a - b = b - c$$
$$a + c = 2b$$

Since points are assumed to be equidistant on an interval scale,

$$\frac{a - b}{b - c} = 1$$

Similarly, the distance from a to c equals twice the distance from a to b.

Whereas there is usually little dispute over whether nominal or ordinal properties have been established, there is often great dispute over whether or not a scale possesses a meaningful unit of measurement. Formal scaling methods designed to this end are discussed in Chapters 2, 10, and 15. For now, it suffices to note that many measures are sums of item responses, such as conventionally scored multiple-choice, true-false, and Likert scale items. Data from individual items are clearly ordinal. However, the total score is usually treated as interval, as when the arithmetic mean score, which assumes equality of intervals, is computed. Those who perform such operations thus implicitly use a scaling model to convert data from a lower (ordinal) to a higher (interval) level of measurement when they sum over items to obtain a total score. Some adherents of Stevens' position have argued that these statistical operations are improper and advocate, among other things, that medians, rather than arithmetic means should be used to describe conventional test data. We strongly disagree with this point of view for reasons we will note throughout this book, not the least of which is that the results of summing item responses are usually indistinguishable from using more formal methods. However, some situations clearly do provide only ordinal data, and the results of using statistics that assume an interval can be misleading. One example would be the responses to individual items scored on multi-category (Likert-type) scales.

The only transformation that preserves the properties of an interval scale is called the general linear transformation and is of the form $X' = bX + a$, where X' is the transformed measure, X is the original measure, and a and b are, respectively, additive and multiplicative constants involved in the transformation. Transforming temperatures from Celsius (C) to Fahrenheit (F), both of which are interval scales, by the relation $F = \frac{9}{5}C + 32$ is a common example. Figure 1-3 illustrates three general linear transformations. Ratios of *individual* values are not meaningful on an interval scale because the zero of an interval scale may be legitimately changed through changes in the additive constant a. The ratios, in degrees Fahrenheit of 64 to 32 and of 100 to 50 are both numerically computable in degrees as 2:1. However, these no longer remain equal, and indeed the first of them becomes undefined, if these temperatures are expressed in degrees Celsius. On the other hand, ratios of *differences* in interval scale values are meaningful. For example, assume the summer mean temperature (in degrees Fahrenheit), of a particular city is 90 during the day and 75 at night. These respectively change to 50 and 40 in the winter. The ratio of the difference in summer and winter temperatures is (90 – 75)/(50 – 40) or 1.5. The corresponding ratio in degrees Celsius is (32.2 – 23.9)/(10 – 4.4) or (within rounding error) also 1.5. This is because the effects of changes in b and a cancel in the process of forming ratios of differences.

When there are only two categories, there is only one interval to consider, so that one interval may be considered an "equal" interval. That is why binary (dichotomous) variables may be considered to form interval scales, the point noted above as being so important to modern regression theory and elsewhere in statistics.

Ratio Scales

A ratio scale is an interval scale with a rational (true) zero rather than an arbitrary zero. A rational zero for children's height in the above example would be physical

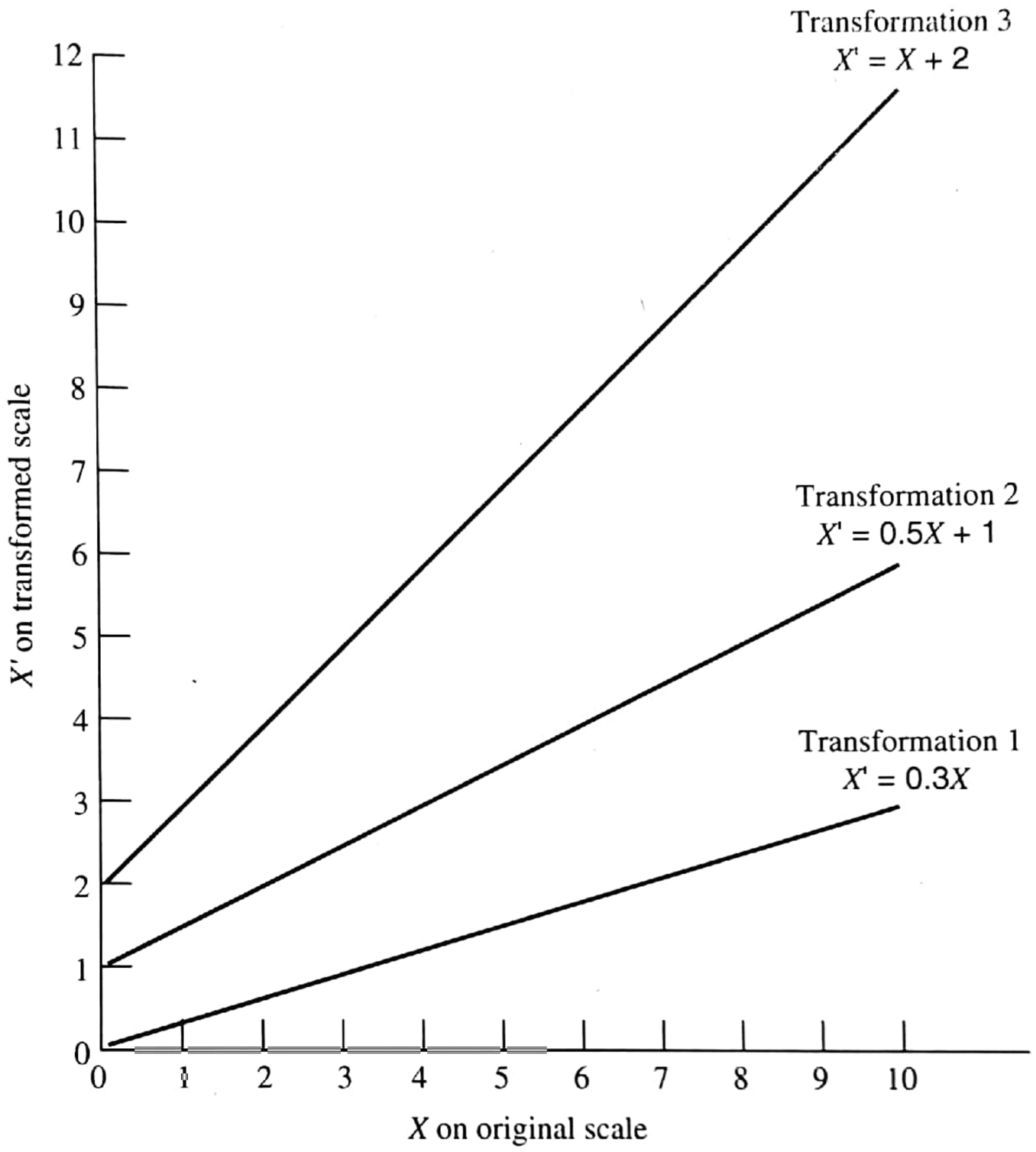

FIGURE 1-3 Three examples of general linear transformations permissible on an interval scale: $X' = X + 2$, $X' = 0.5X + 1$, and $X' = 0.3X$. The general form of the transformation is $X' = bX + a$.

zero rather than the mean height. The presence of a meaningful zero makes ratios of any two measures meaningful. Unlike the three lower types of scales, all four fundamental operations of algebra—addition, subtraction, division, and multiplication—may be used with individual values defined on ratio scales.

A rational zero means absence of the attribute and not simply "reasonable," e.g., zero height or weight. It is often reasonable to reference scores to the mean, but the mean clearly does not denote absence of the attribute, and so it is not a rational zero in the present sense. If there is no rational zero, it does not make sense to form ratios since ratios change as the arbitrary zero changes, another way of saying that ratios of individual values on an interval scale are not meaningful. For example, suppose the class average on a test is 30 and two particular students obtain scores of 50 and 40. Relative to a score of zero, the ratios of these two scores is 1.25:1. However, zero correct is not a rational zero because a student obtaining a score of zero might be able to answer some simpler items correctly. Relative to the mean, the ratio becomes (50 – 30)/(40 – 30) or 2:1, but this ratio is just as arbitrary as the 1.25:1 ratio relative to zero.

There are many examples of rational zeros in physics—zero time and absolute zero (Kelvin) temperature being two others. However, it has proven difficult to define

absolute zeros for most psychological attributes like intelligence. Zero reaction time is based upon physical time, and so it is a rational zero. This means that it is sensible to form such ratios as the mean reaction time obtained from a more versus a less intense stimulus. The major example of ratio scales comes from the fact that differences between observations on an interval scale form a ratio scale. Thus, if pre- and posttest scores on a measure are obtained, the resulting change score can be assumed to form a ratio scale with 0 representing no change. However, Chapter 5 will discuss why change scores may have other problems—it is difficult to compare two change scores based upon different pretest scores.

Actually, ratio scales are rarely needed to address the most common needs of scaling. Defining an interval is very important, but ordering is the most crucial concept. In contrast, nominal measurement rules suffice for most classification problems. It is not proper to employ the general linear transformation permissible with interval scales, only the more restricted form $X' = bX$ is allowable. This more specific form of linear transformation, depicted in Fig. 1-4, is also called a multiplicative transformation. Employing an additive constant (a) implies that the zero point is not fixed, which it is in a ratio scale, by definition. Changing from feet *(F)* to inches *(I)* by the relation $I = 12F$ is a frequently used multiplicative transformation.

Ratios of height, weight, etc., as measured from their true zero points are meaningful. These ratios do not change with permissible transformations since these permissible transformations do not allow a change in the zero point. This is why the term "ratio scale" is used. Someone who weighs twice as much as another person in pounds will also weigh twice as much in kilograms.

Other Scales

Those within the tradition exemplified by Stevens have proposed scale types other than these basic four, and it is important not to think that all scales are divided into four levels. Coombs (1964), Coombs, Dawes, and Tversky, (1970) and Stine (1989a) have discussed these in some detail. One additional type is an ordered metric in which (1) the rank order of objects is known, (2) the rank order of intervals between objects is known, but (3) the magnitudes of the intervals are unknown. Such a scale allows one to say that a and b differ more than c and d but does not allow more precise statements about the relative magnitudes of difference. Stevens (1958) proposed a logarithmic interval scale where the ratios of magnitudes corresponding to successive points a, b, c, d are $a/b = b/c = c/d$, etc. Then $\log a - \log b = \log b - \log c = \log c - \log d$, etc. The decibel scale that is familiar to physicists is a logarithmic interval scale (it is not limited to the measurement of sound intensity), since it involves transforming stimulus energies to their logs.

The absolute scale formed from counts is the strongest type of measurement because it has the interesting property of being its own invariant scale of measurement: When one says "There are three people in the room," the meaning of "three" is inherent in the real number system. In contrast, if you were told a room is three units wide, this might refer to yards, meters, or some other unit of measurement. As interesting as some of these other scales are, though, the four basic ones listed above are far and away the most important to psychometric theory and application.

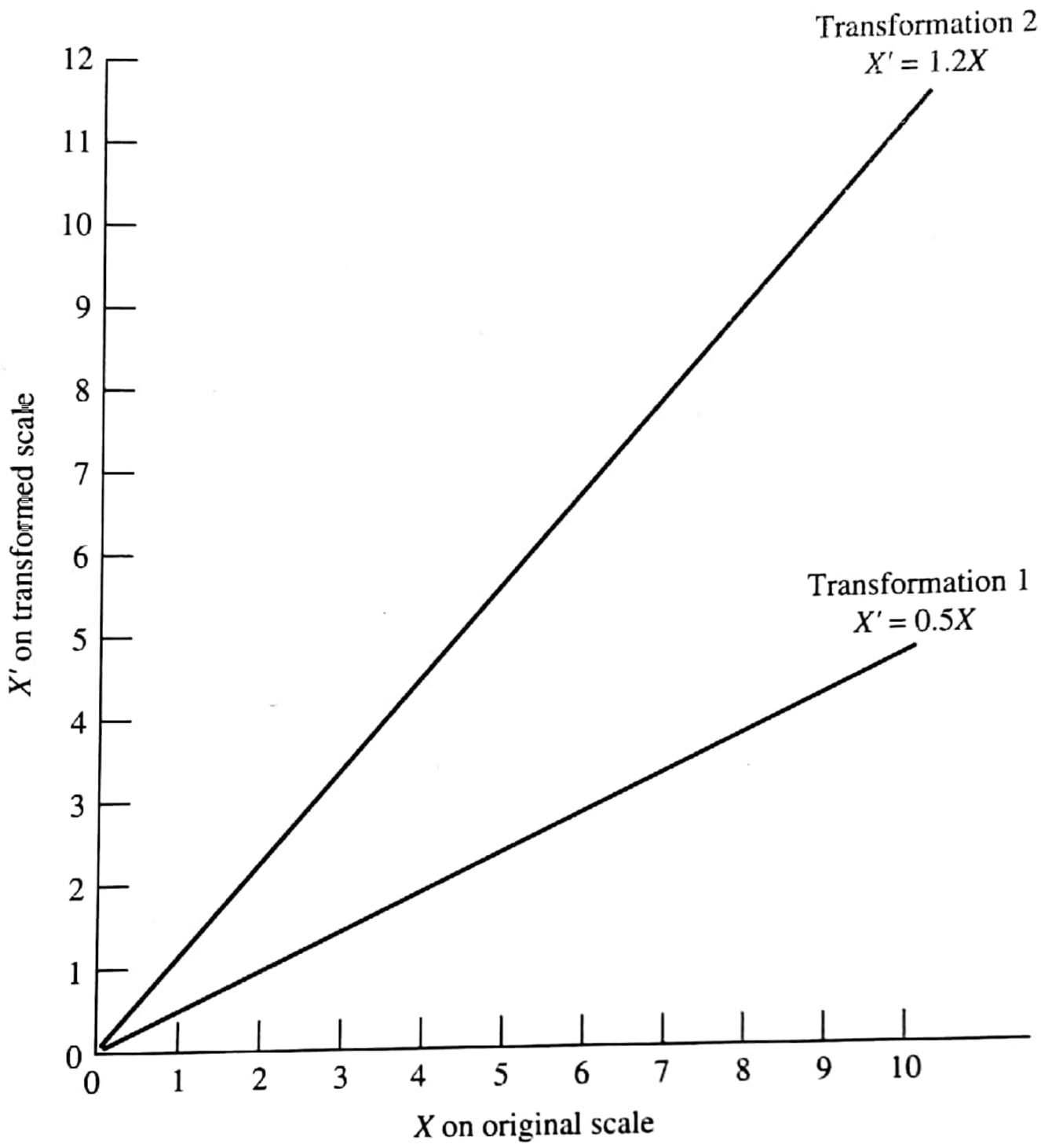

FIGURE 1-4 Two examples of multiplicative transformations permissible on a ratio scale: $X' = 1.2X$ and $X' = 0.5X$. The general form of the transformation is $X' = bX$.

Invariance

It is important to consider the circumstances under which a particular type of scale remains invariant, i.e., maintains its properties when the unit of measurement is changed. As we have seen, the more powerful the mathematical operations meaningful with a given scale, the less free one is to change it. Thus, nominal scale labels may be changed in an almost unlimited manner as long as no two categories are given the same label, but at the other extreme absolute scales lose their absolute properties when changed in any way.

Invariance is basic to the generality of scientific statements derived from a scale (Luce, 1959b, 1990). It is easy to imagine the chaos that would result if some physical measures lacked the invariance of ratio scales. Without invariance, a stick that is twice as long as another measured in feet might be three times as long when measured in inches. The range of invariance of a scale determines the extent to which principles remain unaffected by expressing the scale in different units, e.g., feet rather than inches. This does not mean that the results of using the scale will not change. A mean temperature in degrees Fahrenheit will be numerically different than a mean temperature in degrees Celsius even though both are permissible operations. The point is that the means will change in an orderly fashion: Specifically, the same equation will relate the

means as the one presented above that relate the individual values. Any permissible transformation of a scale produces an equivalent scale—one that maintains the same general form of relationship. Similar statements apply about operations meaningful on other scales.

DECISIONS ABOUT MEASUREMENT SCALES

A strong view of measurement is called the representational position (or the "fundamentalist" position in the previous edition of this book) about measurement scales because it states that scale values represent empirical relations among objects (Michell, 1986, also see Stine, 1989b). Its main assertions are that (1) measurement scales have empirical reality in addition to being theoretical constructs, (2) a possible measure of an attribute can be classified into one of a small number of distinct levels, and (3) investigators must document the scale properties of particular measures before analyzing data obtained from the scale because the scale's level limits the permissible mathematical operations. Besides Stevens, the tradition includes Krantz, Luce, Suppes, and Tversky (1971), Luce (1959b), Suppes and Zinnes (1963), and Townsend and Ashby (1984; also see Ashby & Perrin, 1988; Davison & Sharma, 1988, 1990).

Representational theory had great impact in the 1950s. Investigators tended to avoid parametric tests (*t*, the *F* ratio of the ANOVA, etc.) that required an interval scale (at least according to representational theory) and used nonparametric tests (Siegel & Castellan, 1988) that required only ordinal or nominal assumptions instead. Representational proponents of nonparametric tests argued that these tests were only slightly weaker (less able to detect differences) than their parametric counterparts, a difference that could generally be overcome by gathering slightly more data. However, they largely ignored the greater flexibility of parametric methods in evaluating interactions (combined effects of two or more variables that are not predictable from the individual variables). Starting in the 1960s, investigators returned to the use of parametric tests.

As a simple example of the representational approach, consider this approach to defining the equivalence ("=") of two objects (the presence of a property in common, e.g., being enrolled in the same college course). Equivalence requires transitivity, symmetry, and reflexivity. "Transitivity" means that the relation passes across objects—if John and Richard are enrolled in the course and if Richard and Mary are enrolled in the course, then John and Mary must be enrolled in the course. "Symmetry" means that the relationship extends the same way in both directions—if John is enrolled in the same course as Mary, then Mary must be enrolled in the same course as John. "Reflexivity" states that the relation extends to the object itself—every object is equivalent to itself (if John is enrolled in the course, then John is enrolled in the course but not all examples are that obvious, as we will see in Chapter 15). Parallel considerations yield definitions of the ">" and "<" relationships used to define ordinal scales, the unit used to define interval scales, and the zero point used to define ratio scales. These latter relations are *not* symmetrical, among other things. If Mary is ">", e.g., taller than, Susan, then Susan cannot be ">" Mary. Representationalists have been most concerned with whether a particular measurement achieves interval status so that computing the mean is permissible. We have already stressed the issue of whether scores on a conventionally scored test form an interval scale, and they have often argued that they

do not. We strongly suggest that this position can easily become too narrow and counterproductive. Michell (1986) describes two other traditions that he terms operational theory (Gaito, 1980; Bridgman, 1928) and classical theory (Rozeboom, 1966). Neither accepts Stevens' view that one must have achieved a particular level of measurement to perform a particular statistical operation. Operational theory views a concept as synonymous with the operations that define it. In other words, a score on a test does not represent (stand for something beyond) a measure. It *is* the measure, but it may contain an error so that it merely estimates the trait; operationalism does not require the measure to be the trait itself. Finally, classical theory views measurement as the determination of quantity or how much of an attribute is present in an object (as noted above, we assume that measurement also includes classification).

Gaito (1980; also see Baker, Hardyck, & Petrinovich, 1966) termed his position "statistical theory" and was highly critical of representational theory (which he called "measurement theory"). His tone was very clearly pejorative, but his view nonetheless strikes a sympathetic chord with many investigators who have had to defend what they considered to be obvious aspects of their statistical analyses. Perhaps his major point is that using presumably impermissible transformations usually makes little, if any, difference to the results of the most common analyses. For a counterexample, see Townsend and Ashby (1984).

Ostensive Characteristics

The physical characteristics of the measurement operations provide one way to judge the scale characteristics of a particular measure, e.g., length with some form of yardstick. To prove that the attribute in question is measured on a ratio scale requires proof of both (1) equal intervals and (2) an axiomatically unquestionable zero point. Anyone can *see* the zero point where the yardstick starts. The beginning of the measuring instrument is the front of the yardstick, and open space is behind that point. Who could argue for a more meaningful zero point? The equality of intervals is also easy to demonstrate, e.g., saw the yardstick inch by inch and compare the inch long pieces to ensure equality.

To a lesser or greater extent, all other measures employ correlates of the attribute rather than the attribute itself and are therefore indirect. We can establish equality of time intervals but, strictly speaking, we observe the effects of time and not time itself—ticks, pendulum swings, and the earth's rotation are only consequences of time. Nearly all measures of interest to behavioral scientists are indirect. We cannot observe intelligence per se but only its by-products. Likewise, a subject's perception can only be inferred from subjects' ability to discriminate and/or report what they experience (Eriksen, 1960).

Many investigators, who may not even consider themselves representationalists in a formal sense, tend to evaluate scale properties in terms of ostensive characteristics and think of actual measures as imperfect correlates of "real" ones even though the scales may have been developed from a formal scaling model. We suggest that if the data obtained from applying a measurement scale fit the axioms of the particular model under consideration and the axioms (assumptions) of the model are appropriate, then the measure has scale properties specified by the model. For example, Chapter 2 contains

a model proposed by Louis Guttman for the construction of ordinal scales. It is based upon assumptions about patterns of responses to test items. Relevant data can be analyzed to determine how well the actual score patterns relate to the patterns predicted by the model. A good fit implies that an appropriate scale exists.

Since, for example, there are no ostensive properties to guarantee the equality of intervals measuring intelligence, some have argued that intelligence tests, for example, provide ordinal scales at best. We hope the above discussion illustrates that few measures in all sciences would be considered more than ordinal scales by these standards; the following sections will show that proper standards for judging the scale properties of a measure do not require observing the ostensive characteristics of an attribute. In particular:

1 Standards can be based on data rather than ostensive characteristics. One studies the results of applying a measure to real objects when using a scaling model, or one studies the measurement tool directly when using ostensive characteristics. Thus, instead of relying upon the ostensive properties of yardsticks, one could test a model concerning properties of ratio scales and then see if it fits data obtained from yardstick measurements. One could therefore derive the scale properties of the yardstick from the model before seeing a yardstick. People have done this, and the data fit a variety of scaling models beautifully, e.g., produce transitivity. This is what psychological scaling is about: *It is an attempt to work backward from data to test the fit to a model.* In this way, ratio, interval, ordinal, or perhaps nominal scales for psychological attributes which cannot be seen directly may be constructed.

2 Using scaling models is a healthy trend in the development of measurement methods. Many models are intuitively quite appealing. Because they specify the characteristics that should be found in data, they are subject to refutation (can be falsified, Popper, 1959). Some models have produced scales that have led to interesting scientific findings.

3 A model is no better and no worse than its assumptions (axioms). There is ample room for disagreement, and there is plenty of it, about the fruitfulness of different models. For example, we have argued that measures like multiple-choice test scores should be viewed as having interval properties. However, if psychologists disagree about the correctness of different scaling models, how are scale characteristics ever determined? If, for example, several interval scaling models are being tried on a particular type of data, a failure of the data to fit one model does not automatically prevent the measure from being considered as an interval scale. Conversely, even if the data fit *all* the models, the measures should not automatically be thought of as constituting an interval scale. A more final decision should be made with respect to standards to be discussed in the following sections.

Consequences of Assumptions

Even if one believes that there is a real scale for each attribute that is either directly present in a particular measure or mirrored in a monotonic transformation, an important question is What difference does it make if the measure does not have the same

zero point or proportionally equal intervals as the real scale? If the scientist assumes, for example, that the scale is an interval scale when it really is not, something will go wrong in the daily work of the scientist. What could go wrong? How could the difficulty be detected? The scientist could misstate the specific form of the relationship between the attribute and other variables. For example, a power function might be found between two measures using an imperfect interval scale, whereas the right scale may produce a linear relationship.

How seriously would such a misstatement affect the progress of the behavioral sciences? At present, the usual answer is "very little." Most results are reported as either correlations or mean differences. We have stressed and will stress that correlations are little affected by monotonic transformations on variables. These correlations are the basis of still more powerful methods like factor analysis. However, we also stress that justifying the rank order is vital. Even if one accepted the representational point of view about measurement scales, what sense does it make to sacrifice powerful methods of correlational analysis just because there is no way of proving the claimed scale properties of the measures?

There is also often major concern about the ratios of variances among different sources of variation in analyzing mean differences among groups, e.g., F, the variance among means relative to the variance within groups. This ratio and related statistics are also little affected by monotonic transformations of the dependent measure. If it is granted that the measure used in the experiment is at least monotonically related to the real scale, it usually makes little difference which is used in the analysis. There are some exceptions of import. Two of these are (1) in examining details of functional relationships, such as whether a particular monotonic relation is linear, logarithmic, a power function, or some other form, and (2) for some goodness-of-fit tests used in structural modeling (see Chapters 5, 10, and 15).

A simple rule of thumb is that transformations become more important as the level of sophistication of the research hypotheses increases. Thus, tests simply concerned with looking for group differences and rank orderings of groups typically involve statistical procedures that are little affected by transformations. Numerically, these perhaps account for the vast majority of research. Interval assumptions are therefore not crucial when interest centers on ordinal relations among group means, etc. However, more refined tests of highly quantitative models are very sensitive to the interval properties of the scale, virtually by definition.

After analyzing the results of investigations, as in correlations and/or ratios of variance components, it often is important to make probability statements about the results after applying inferential statistics. Thus, it may be important to set confidence zones for a correlation coefficient or to test the significance of a particular ratio among components of variance. Such statistical methods are completely indifferent to the zero point on a scale and consequently do not require ratio scales. However, they do assume interval properties, but since they are based on ratios of variation and covariation, they are also little affected by monotonic deviations from any true interval scale. Moreover, statistical methods are completely blind to any meaning in the real world of the numbers involved. These methods require only a definable population of numbers that meets the assumptions in the particular statistical method, such as normality of the

population error distribution. We suggest that it is perfectly permissible to employ the ANOVA to test hypotheses about the average size of the numbers on the backs of football players on different teams. What use you may make of the result is, of course, a different story, since there is no meaning to a theory of football numbers beyond identifying the position individuals play, rather than how well they play it (see Lord, 1953).

Chapters 14 and 15 will consider some extremely useful consequences of the representational point of view. We merely note that it is easily misused when the usual intent is to compute correlations or infer the ordering among groups means. Moreover, even when the intent is to study specifics of functional relations, one may discover that two perfectly good definitions of attributes are not linearly related to one another so that the "true" relation to other measures depends upon how the attribute is defined.

Convention

We have thus far considered the representational point of view that scientists normally think in terms of "real" scales and obtain measures as approximations to such "real" scales. Our opinion is that (1) this point of view frequently leads to unanswerable questions and (2) violations of even relatively important assumptions are not harmful in *most* settings. The authors oppose the concept of "real" scales in most settings and deplore the confusion that this conception has wrought to the average investigator. It is much more appropriate to think of measurement scales as conventions or agreements among scientists about a "good" scaling.

In saying that scales are established by convention and not God-given, we do not mean that such conventions should be arbitrary. Before measuring an attribute, all manner of wisdom should be sought as to the nature of the attribute—one cannot measure something unless one has some general conception about what is to be measured. The nature of a "good" scaling of certain measures can be so readily agreed that a convention is easily established, e.g., length, weight, and time. Exasperation about theories of measurement has tempted some to wish that there were no yardsticks and no balances for the measurement of weight so that all scientists could see that measurement always involves convention rather than discovery of the "real" measure.

Sometimes, one person establishes a measurement convention and other scientists often neglect to participate in establishing the particular convention. Consequently, the particular scale becomes accepted as *the* scale. The Fahrenheit thermometer was once taken as *the* scale of temperature. Later, the discovery of absolute zero led to a new and more useful scaling. In psychology, intelligence was once defined as the ratio of mental age to chronological age, i.e., as an intelligence quotient (IQ), but intelligence is now measured relative to performance within a given age distribution. Both these instances illustrate why it is wrong to think that "real" scales had been discovered. It is better to say that conventions changed because better conventions were developed. The key is continued *validation* of measures.

After applying all available wisdom to the problem, it is good to apply some type of formal scaling model when actually constructing measurement scales. Although any set of rules for the assignment of numbers constitutes measurement, silly and/or ad

hoc rules probably will not result in a useful measure. It is useful to think of a scaling model as an internally consistent plan for scaling an attribute. When the plan is put to use, the measure may eventually prove unsatisfactory to the scientific community, but having a plan increases the probability that it will be acceptable. Sometimes, useful measures are simply stumbled upon. However, explicit plans based on common sense and past experience improve the probabilities of a useful measurement scale.

A convention establishes the scale properties of a measure. If it is established as a ratio scale, then the zero point can be taken seriously and the intervals may be treated as equal in any form of analysis. If it is established as an interval scale, the intervals may be treated as equal in all forms of analysis. This is not meant to imply that such conventions are, or should be, established quickly or until much evidence is in, but in the end they are conventions, not discoveries of "real" scales.

Certain conventions are not employed because they make no sense or do not lead to useful results. For example, the Celsius scale's use of the freezing point of water to define temperature's zero point has limited scientific utility. Water is an important substance, but it not the only important substance. On the other hand, the absolute zero of the Kelvin scale based upon the absence of molecular activity is useful to a wide range of physical laws. It similarly makes little sense to establish zero points on scales of many, but not all, psychological attributes. Zero intelligence might be defined as the problem-solving ability of a dead person, but the utility of this convention in establishing a ratio scale of intelligence remains to be determined. Psychologists seek to develop interval scales for many attributes because it is reasonable to ask how far apart people are on the scale and not simply their ordering. For example, we frequently need to determine if a is closer to b than to c.

Scaling procedures that make sense may still not produce scales that *work well in practice.* These last four words are the key to establishing a measurement convention—a good measure is one that mathematically fits well in a system of lawful relationships. Chapter 3 will emphasize that the usefulness (validity) of a measure is the extent to which it relates to other variables in a domain of interest. The "best" scaling of any particular attribute is that producing the simplest forms of relationship with other variables. An increasing hierarchy of simplicity is (1) a random relationship, (2) a nonrandom pattern fitting no particular line of relationship, (3) an unevenly ascending or descending monotonic relationship, (4) a smooth monotonic relationship, (5) a straight line, and (6) a straight line passing through the origin. The only way to describe a random relationship completely is to describe every point. However, a straight line passing through the origin is completely described by $Y = bX$, and the b (slope) parameter is usually arbitrary. Since the scientist's task is to translate and simplify the complexity of events in the universe through lawful relationships, the simpler these relationships, the better.

One way to make relationships simpler is to change the scaling of one or more of the variables. Thus, an irregular monotonic relationship can be smoothed by stretching some of the intervals, a procedure widely used by Anderson (1981, 1982) under the name "functional measurement." Any monotonic curve can be transformed to a straight line by this device. A straight line can be made to pass through the origin by changing the origin (zero point) on one of the scales. Of course, conventions about

a particular attribute should not be altered because of the relationships found with only one or two other measures. One should consider the effects upon several measures. Nonetheless, if many relationships are simplified by a particular transformation, the new scale is logically a better scale. Such transformations are made actually quite frequently. For example, logarithmic transformations are quite common, especially in sensory psychology.

Following this point of view to the extreme, there is no reason why all variables known to science could not be rescaled to simplify all relationships. This would be a wise move if it could be done—a big "if." The new scales are as "real" as the old ones, and there might be every reason to take the zero points and the intervals on the new scales seriously.

There are two major problems with considering scaling merely as a matter of convention. First, it is disquieting to those who think of real scales and futilely wish for infallible tests of the relationships among real scales. Looking at measurement scaling as convention also seems to make the problem "messy." How well a particular scaling of an attribute fits in with other variables is vague. Which variables? How good is a particular fit? To avoid such questions, however, is to blind oneself to the realities of scientific enterprise. To seek shelter in the apparent neatness of conceptions regarding real scales is not to provide answers about the properties of measurement scales but to ask logically unanswerable questions.

A second, and more serious, problem with considering scaling as a matter of convention is that two or more conventions often compete with one another. For example, there has been much dispute about whether Thurstone's law of comparative judgment or Stevens' magnitude-estimation methods better describe the results of measuring sensations (see Chapter 2). As it turns out, Thurstone's procedures are more useful in describing lawful relations involving confusion among stimuli, and Stevens' methods are more useful in predicting how stimuli will appear (the two are also simply related through a logarithmic transformation). More appropriate than asking which is correct would be to ask whether confusion among stimuli or their appearance is at issue in the particular situation.

Having competing conventions regarding the scaling of attributes is not as bad as it sounds for two reasons. First, if the two scalings are monotonically related to each other, as is usually the case, and if one has a monotonic relationship with a third variable, so will the other. Thus the principles established with the two scalings will produce the same general functional relations, even though the specifics may differ. The specific form of relationship is rarely the major issue in contemporary psychology even though it can be. The more common question is the strength of relationship between the two variables. Correlations greater than .60 are the exception rather than the rule, and, as was said previously, such correlations are largely insensitive to monotonic transformations. Consequently if there are two competing, monotonically related conventions for scaling that are equally reliable in the sense to be describe later, both will produce about the same correlation with any other variable. In sum, the specific forms of relationship can be settled only when there are firm conventions for scaling. The specific form of a relationship is relative to the measurement convention. To hope to find the relationship is either to continue to search vainly for real scales or to assume that one measurement convention eventually will win out over others.

Classification as Measurement

We have devoted nearly all of this chapter to the first part of the definition of measurement, measurement as scaling. This is because measurement as scaling has led to more issues of dispute than has measurement as classification, and because until recently there were few sophisticated techniques to use with categorical (nominal) data, the usual fruits of classification. This has changed, especially since the last edition of this book, and Chapter 15 will focus on some of these new developments.

Classification demands a nominal scale (rules to define "=" and "≠") at a minimum and, conversely, illustrates that a nominal scale, which was considered "lowly" in terms of scaling, can be extremely important. Consider two common statements: (1) "Everyone is unique; no two people are the same" and (2) "People are pretty much alike." Although these two statement appear totally contradictory, both share the characteristic that they lead one away from some useful, if not obvious, results. For example, people who describe themselves as Republicans are quite likely to answer a variety of politically related questions differently from people who describe themselves as Democrats, e.g., "Should prayer be allowed in public schools?" Similarly, the relation between political affiliation and response to the political issue may jointly vary with additional variables such as whether the person lives in a rural, suburban, or urban area. Note that this analysis does not necessarily ignore individuality. Two people who fall within the same "cell" of the analysis (e.g., who are both Democrats, live in a suburban area, and oppose school prayer) may differ in countless ways (e.g., gender, religion, height, or weight). As with scaling, classification assumes *equivalence* and not *identity*.

Although classification is relatively simple conceptually, it can be quite difficult empirically. Useful classification along one dimension implies that the dimension in question will relate to another dimension (which in turn could be at any of the previously mentioned levels). There is no reason to classify people as type alpha versus type beta unless these categories have a useful external correlate. Even such obvious categories as Catholic, Protestant, Jewish, and Muslim may not be widely useful (though religiously orthodox versus religiously nonorthodox, disregarding the specific religion, may be). Moreover, apparent relations between a categorical variable (or any other) and a given criterion may be an artifact of a third variable; religious differences may, for example, be an artifact of differences in education and/or income. Thus, one may obtain apparent differences between Catholics and Protestants on an issue that involves liberal versus conservative attitudes because more affluent individuals also tend to be more conservative and the two groups differ in affluence. Likewise, empirical disputes often arise between "lumpers" (people who favor a small and therefore more parsimonious number of broad categories) and "splitters" (people who favor a larger number of more finely defined categories).

RECENT TRENDS IN MEASUREMENT

The Impact of Computers

It is very easy to think that the main role of a computer is to expedite analyses that one would have performed anyway. This is certainly important. Anyone who has used computers for a long time appreciates the increasing flexibility and user-friendliness of

major computer packages such as BMDP, SAS, SPSSX, SYSTAT, and UniMult. One likewise appreciates the related factors of greater power, increased reliability, and lower cost in the personal computers that are now beginning to dominate statistical analyses and the availability of supercomputers for massive undertakings. However, one additional point must be stressed—computers now allow fundamentally different kinds of analyses to be performed, i.e., open form analyses that are effectively impossible to do by hand.

Closed versus Open-Form Solutions

Many of the techniques, concepts, and measurement theories that have recently become popular actually have long histories. However, they were essentially interesting statistical curiosities before computers became generally available. The distinction between closed- and open-form solutions helps make this point more understandable. Consider your first statistics class where you were taught to compute the arithmetic mean of a sample by adding up the scores and dividing by the number of scores and given the associated equation $X = \Sigma X/N$. This is a closed form solution because all you need do is plug the numbers into the formula to obtain the result. You might wish to use a computer if N were very large, but the principle would be the same.

On the other hand, suppose you did not know the formula but for some bizarre reason you remembered that the mean minimizes the sum of squared deviations. This too can be expressed by a kind of formula: $\Sigma(X - C)^2 =$ a minimum when $C = X$, but the formula does not tell you how to obtain X. You might use this information to compute X by plugging in different values of C, computing the sum of squared deviations for each value, and accepting the one producing the smallest sum. If you performed enough calculations, you could in fact obtain an open-form estimate of X.

Many statistical quantities of interest, particularly those of recent prominence, require an open form of estimation because they lack a closed-form solution. This is often true of maximum likelihood estimates discussed at several points in this book. For all intents and purposes, such estimates require a computer and, even then, can be very time-consuming. The process involves repeated calculations or iterations. Numerical analysts often specialize in developing better algorithms to obtain the necessary successive approximations. Iterative proportional fitting and Newton-Raphson algorithms are two such common computational processes. You will not need to know how to use either one yourself, but they are widely employed in programs you may use.

Computer Simulation

Computers are also invaluable in simulating processes. A particular form of simulation that is widely performed on computers is the Monte Carlo method in which an estimate of a parameter is obtained by random sampling. If you were asked to verify that the probability of obtaining heads on a coin flip is .5, you might actually flip a coin a large number of times and count the actual number of heads, hoping the coin was fair. This would illustrate the Monte Carlo method but would not be a computer simulation. The experiment may be done more efficiently on a computer where the program would

conduct a series of trials. On each trial, the program generates a random number from 0 to 1 and adds one to the count of heads if the random number is greater than 0.5. When finished, it prints the proportion of times heads occurred. Computer simulations are often performed when it is difficult to obtain a solution analytically (algebraically) or if no solution is known to exist.

SUMMARY

Measurement consists of rules for assigning symbols to objects to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification). Both scaling and classification involve the formulation and evaluation of rules. These rules are used to measure attributes of objects, usually, but not exclusively, people. It is important to remember that we can measure only attributes of objects, not the objects themselves. Among the characteristics of good rules are repeatability (reliability) and, more importantly, validity in senses to be described. Standardization is an important goal of measurement because it facilitates objectivity, quantification, communication, economy, and scientific generalization.

Measurement uses mathematics, but the two serve separate roles. Measurement needs to relate to the physical world, but *pure* mathematics is solely concerned with logical consistency. One traditionally important, but controversial, aspect of scaling that involves mathematics is the concept of levels of measurement: Scales generally fall at one of four levels (others have been suggested): nominal, ordinal, interval, and ratio. These four levels represent progressively better articulated rules. For example, nominal scales simply define whether or not two objects are equivalent to one another with respect to a critical attribute, but ordinal scales determine whether one object that is not equivalent to another is greater than or less than the other. Stronger results are possible from higher levels of measurement. Basic to these levels of measurement is the concept of invariance, which concerns what remains the same as permissible changes are made in the scale (e.g., in its unit of measurement); higher-level scales are more restricted as to how they may be transformed and still preserve key invariances.

Focal to the debate about levels of measurement is what statistical operations are permissible on a given set of measures. The representational position asserts that scale properties must be established before performing relevant operations; e.g., a scale must demonstrably have interval properties before it is proper to compute an arithmetic mean. Alternative positions, classical and operational, do not share this view. Many, who need not be formally aligned with a specific position, look for scales to have ostensive (visualizable) properties like yardsticks or clocks have before accepting a scale as real; they view existing measures as highly imperfect correlates of true scales. We suggest that very few measures in science are ostensive. A much better criterion is the extent to which the results of using the scale fit a scaling model. All measurement use is essentially based upon convention, and progress is made when better conventions are agreed upon. In general, the more well elaborated a hypothesis is stated quantitatively, the more important formal scaling issues are.

The most important single factor in the recent progress in measurement has been the computer. Although computers obviously allow analyses that could be done by hand to be done more easily and accurately, they allow fundamentally different analyses to be performed. Many of these use open-form solutions, so named because the results cannot be defined directly by a formula (closed-form solution). In addition, computers allow simulation of processes that are difficult to study directly.

SUGGESTED ADDITIONAL READINGS

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 12,* 1304–1312.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87,* 564–567.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms, *Psychological Bulletin, 100,* 398–397.

Stevens, S. S. (1958). Problems and methods of psychophysics. *Psychological Bulletin, 55,* 177–196.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived, *Psychological Bulletin, 96,* 394–401.

Note: Sage Publications offers many short monographs on an extremely wide variety of relevant topics aimed at scholars who are not quantitative specialists. Although they should not be used as the sole guide to a given problem because of the innumerable complications that may be present, they are highly recommended as starting points. We will not cite these works individually.