PART TWO

STATISTICAL FOUNDATIONS

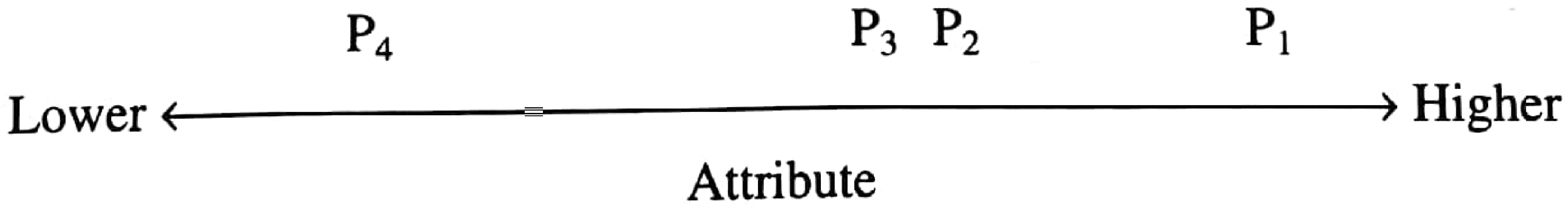
Part Two contains four chapters that deal with statistical concepts basic to measurement. First, we look at some models used to construct scales. One central concept is that of the item trace line (item-characteristic curve) which relates the magnitude along a dimension (trait) to the magnitude of response to a particular item. The next chapter deals with the three basic meanings of test validity: content validity, construct validity, and predictive validity. Many have debated whether these are ultimately the same or not. Though they share important similarities, there are also important differences among them. The third chapter considers statistical description and estimation. Much of this involves traditional issues in correlation and regression that you may have been previously exposed to. However, two additional topics may be less familiar: structural relations and alternative forms of statistical estimation. The latter is important because statistical inference plays a much larger role in psychometric theory than it did in the previous edition. The method of maximum likelihood is especially important. Finally, we discuss properties of linear combinations which are central to psychometric theory.

CHAPTER 2

TRADITIONAL APPROACHES TO SCALING

CHAPTER OVERVIEW

Scaling was defined in Chapter 1 as the assignment of numbers to objects to represent quantities of attributes. Although any relevant set of rules can be spoken of as measurement, it helps to have some internally consistent plan when developing a new measure. The plan is a "scaling model," and the resulting measure is a "scale" or a "measurement method." The simplest example is a ruler used as a scale of length. The methods for constructing and applying rulers constitute the scaling models. Scaling models are designed to generate one or more dimensions (continua) to locate people or objects. In the following example, persons P_1, P_2, P_3, and P_4 fall along one such dimension, which could be social anxiety, spelling ability, attitude toward abortion, etc.



Because this is an interval scale, the distances between people are meaningful. Thus P_1 is considerably higher in the attribute than P_2, P_2 and P_3 are close together, and P_4 is far below the others.

We begin this chapter with an introduction to the concept of a data matrix, which is central to nearly all measurement data, and some differences between scaling stimuli and scaling people. Next, we present a brief history of "psychophysics," which is the study of the relation between variation in physical dimensions of stimuli and their associated responses—as it forms the foundation for "psychometric" theory. In contrast, "psychometrics" in general may or may not study the effects or variation in a single physical dimension, and so it includes psychophysics as a topic. Then, some distinctions among different types of stimuli and, especially, responses are made. We then consider some general principles underlying the development of ordinal, interval, and ratio scales. Following this, we present what is probably the historically most important scaling model for stimuli, Thurstone scaling. The ensuing section considers some models used to scale people. In particular, we introduce the linear model (also called the summative or centroid model), which simply involves the familiar process of defining a score as the ordinary sum, perhaps weighted, of responses to individual items.

DATA MATRICES

Most measurement problems begin with a data matrix or two-way array or table (we will describe some other matrices from time to time). Rows typically represent *N* different objects (usually people), and columns represent *K* different stimuli (content), e.g., questionnaire items (see Table 2-1). It is convention to denote the entire matrix by an uppercase letter in boldface, e.g., **X.** The data are responses, e.g., 0 = incorrect versus 1 = correct, Likert scales, etc. Individual elements appear in lowercase italics. The first subscript conventionally denotes the row (usually the object being measured, e.g., a person), and the second subscript denotes the column (stimulus, as a questionnaire item number), so that x_{ij} denotes the response of subject i to stimulus j. However, the stimuli and responses can represent anything that the experimenter does to the subjects and anything the subjects do in return. Consequently, we need not limit the discussion to people and test items in the ordinary sense. Subjects might estimate the weights of various objects, for example. It is possible, though rare, that the matrix is a single person's response to a series of stimuli studied over occasions (e.g., Nunnally, 1955), among other variants.

Most classical psychometric models treat scale items as replicates of one another in the sense that differences among the items are ignored in scaling. Thus, a patient's anxiety is typically defined by counting the number of anxiety-related symptoms that are endorsed regardless of which specific items these are. Alternative models, mainly of recent origin, derive scale scores from the pattern of responses. These latter models

TABLE 2-1 A BASIC TWO-WAY DATA MATRIX (**X**) CONTAINING RESPONSES OF *N* PERSONS (ROWS) BY *K* STIMULI (COLUMNS)

				Stimuli			
		1	2	3		*j*	*K*
	1	x_{11}	x_{12}	x_{13}	...	x_{1j}	x_{1k}
	2	x_{21}	x_{22}	x_{23}	...	x_{2j}	x_{2k}
Objects	3	x_{31}	x_{32}	x_{33}	...	x_{3j}	x_{3k}
(People, usually)	...	...	...	...	...	...	...
	i	X_{i1}	x_{i2}	x_{i3}	...	x_{ij}	x_{ik}
	n	x_{n1}	x_{n2}	x_{n3}	...	x_{nj}	x_{nk}

will be introduced here but are discussed in more detail in Chapter 10. Likewise, methods of scaling objects, as in market research studies, often assume that people are replicates of one another. For example, the percentage of persons in a group that prefer one brand of cereal to another is assumed to be the same as the percentage of times a typical (modal) individual would have this preference over occasions. These classical methods, by definition, treat individual differences among items and people as random error. In contrast, newer methods incorporate individual differences in a more systematic manner.

It is only meaningful to obtain a single measure by counting the number of positive responses if the stimuli measure a single attribute. This in turn implies that differences in response to the various stimuli are highly correlated; e.g., if people who admit to one anxiety-related symptom also tend to admit to others, and vice versa for people who deny these symptoms. Various correlational methods are used to evaluate the extent to which people or stimuli can be viewed as replicates. If responses correlate poorly with one another, two or more scales would have to be formed from the items. These involve methods discussed throughout the book, especially in Chapters 11 through 14. This chapter will be limited to models that assume the stimuli measure a single attribute (unidimensional scaling)—situations in which the data under consideration can be summarized satisfactorily with only one "yardstick."

More Complex Organizations

The two-way organization of Table 2-1 contains the minimal elements of interest to a measurement problem. If there were but a single column (stimulus), there would be no way to evaluate the structure of the stimuli, which is basic to psychometric theory. The only results possible would be descriptive statistics on the single measure (e.g., the mean and standard deviation) for the single group of subjects. These data are rarely of interest to the psychometrician because nothing can be said about the structure. Likewise, data from a single row (subject) in isolation are unlikely to be informative. At a minimum, we need to compare that person's data to normative data.

More complex arrangements of the data are extremely common. First, the two-way matrix may be repeated over occasions, as when a pre- and a posttest are administered. This gives rise to a three-dimensional arrangement in which there are rows and columns, as before, plus "slices" that represent the two or more occasions. Another possibility is that subjects are sampled from two or more groups; e.g., one studies gender differences in response to items measuring depression. A third possibility is that two or more attributes are investigated simultaneously, as when one series of items measures job satisfaction and another series of items reflects job performance. This design involves methods of multidimensional (multivariate) analysis considered later in the book.

Scaling objects often involves a three-dimensional array, as when a market researcher conducts a taste test and has people judge multiple attributes of several brands of cola, e.g., sweetness and intensity of flavor. (As an incidental point, the application of measurement methods to quantify the perceived appearance, including taste, of consumer product preferences is known as "sensory evaluation" to market researchers.) These possibilities may be combined in still higher-order ways, e.g., by obtaining pre- and posttest measures that compare two or more groups.

We have frequently used the phrase "people or objects," but the vast majority of studies examine people's responses to different stimuli. In fact, objects (which may be abstract concepts) play the same role as people in some studies and as stimuli in others.

"Holes" in the Matrix (Missing Data)

In an ideal situation, there is an outcome at each location in the matrix; e.g., each person is administered each stimulus. Sometimes this is not possible or even meaningful. For example, the number of stimuli may be too large to allow a given person to respond to each one. Similarly, the effects of administering one stimulus may influence subsequent behavior, known as "carryover" effects. Subjects are then often deliberately given a subset of the stimuli chosen according to a predesignated plan usually involving random assignment of stimuli to a given subject. This is part of the experimental design. Perhaps the most comprehensive text dealing with these problems is Winer, Brown, and Michels (1991). Although some statistical power is lost when subjects do not respond to all stimuli, this loss of power can be offset by increasing the sample size. The problem will not be considered further since it poses no additional complications.

Far more serious problems emerge when the resulting holes in the data matrix are nonrandom. For example, the second author once was given neuropsychological test data. The data involved many scales (subtests) that were normally not all administered to each patient. Thus, patients with frontal lobe damage were given one set of subtests, patients with temporal lobe damage were given a different set of subtests, etc. Such limitations on data gathering caused the missing data to be nonrandom. Type of injury was confounded with the particular scales that were administered. The results obtained from analyzing these data might well differ substantially from a study in which all subjects responded to all measures or the pattern of administration was random. Good design dictates minimizing the impact of missing data. If all measures are equally important, randomize the order of administration or administer random subsets if all cannot be administered to each subject. Conversely, if some are relatively unimportant because they are being used for more exploratory purposes, administer these at the end.

EVALUATION OF MODELS

Often different models can be applied to a given set of data to develop alternative scales. These models and their associated scales sometimes lead to different substantive conclusions. Two different models might produce scales that are not linearly related. One model might suggest that the data do not even possess ordinal properties, whereas another might indicate they clearly form an interval scale. How, then, does one know which model to choose? Chapter 1 noted why this cannot be known in advance. We suggest that the most crucial test is how well the scale provides meaningful repeatable relations with other variables. Before time and effort are spent on such investigations, however, some additional criteria can be applied.

1 The intuitive appeal of a scaling model provides one criterion for "reasonable." Although the data of science must be public, a scientist's intuition plays an indisputable role in the gathering and analysis of data. Looked at in one way, a measurement model is nothing more than an explicitly defined hunch that particular operations on data will be useful. In particular, we suggest that psychologists lean toward measurement models that are most analogous to the measurement of simple physical attributes, e.g., length.

2 Another aspect of "reasonable" is that one should exploit what is already known about similar data. For example, power functions are well known to describe relations between physical and perceived intensity (see below). On the negative side, some models assume that individual test item responses are highly reliable; yet, a wealth of evidence shows that such responses usually are highly *un*reliable.

3 Preliminary analyses often provide cues about the usefulness of a scale. If the scale values for objects or persons are markedly affected by slight procedural differences, the scale will probably not work well in practice. There are, for example, numerous ways in which subjects can judge weight. If two similar appearing approaches yield very different intervals of judged weight, either or both methods are suspect. Conversely, different models that yield similar results provide converging operations (Garner, Hake, & Eriksen, 1956) that mutually strengthen the confidence one may have about any given method. "Triangulation" is another common term used to describe this.

4 Another important type of evidence is the magnitude of measurement error in using a particular scale, which we will discuss in detail in Chapters 6 to 10. A scale that yields a great deal of measurement error cannot possibly be useful.

Beyond the standards of good sense, however, the ultimate test of any model is the extent to which it yields useful empirical results.

Scaling Stimuli versus Scaling People

Although psychometric methods can be used to scale people, stimuli, or both, different methods are often used when the focus is on scaling people than when the focus is on scaling objects. As Cronbach (1957) pointed out in a classical article, clinical, counseling, and school psychologists are more inclined to think in terms of individual differences among people, e.g., in measuring such attributes as intelligence and level of adjustment. These individual differences are a nuisance to experimental psychologists and market researchers who largely ignore individual differences, though both may be interested in group differences. Their problems typically involve scaling stimuli, e.g., measuring which words or advertisements are most readily recalled. Regardless of the focus of the research, the basic data are representable as a two-dimensional array, perhaps extended into other dimensions because of additional considerations.

Unidimensional scaling of people is probably the easiest situation to describe. For example, a spelling test contains words as stimuli and students as subjects. The data are simply 1 = correct and 0 = incorrect. The simplest model for scaling subjects (see the linear model below) collapses the stimulus dimension of words by adding the number of 1s for each person. Although additional analyses are usually conducted to determine the interrelations among responses to different words, these simple sums of correct responses scale students on their spelling ability. Consequently, Dina may obtain a score of 48 and Ralph may obtain a score of 45 out of 50 words. It is quite possible that a simple ranking of the students will suffice so that an ordinal scale may be all that is necessary for such purposes as grading. The major requirement in scaling people is that alternative scalings be monotonically related to one another, i.e., that they rank-order people in the same way. Thus if two different methods for scaling anxiety have a strong monotonic relationship, research results will be much the same regardless of which scale is employed.

The roles of people and stimuli are often reversed to scale objects. Specifically, sums over students for each word describe differences in the difficulty of the words, e.g., if 50 students spell "abacus" correctly but only 35 spell "mnemonic" correctly, "mnemonic" is considered more difficult than "abacus." In fact, these data are usually a standard part of a test analysis, even when interest is directed toward scaling people. However, studies directed toward scaling stimuli are also more likely to be concerned with establishing functional relationships to various attributes, in which case ordinal scales are quite likely to be insufficient. Assume, for example, that the stimuli are tones of different intensity which subjects rate for loudness. Everyone knows that more intense tones will be rated louder; the key to the study is whether the relationship is logarithmic, linear, or of some other form. A unidimensional scale of stimuli should also fit a typical (modal) individual. Such a scale should be typical of a group even if it imperfectly represents the data from any one individual.

Because of the thornier problems in stimulus scaling, most of the issues and more complex scaling models have arisen from scaling stimuli. This difference has influenced the language used to describe psychological research. "Scaling" and "scaling methods" usually denote the scaling of stimuli. Problems of scaling people are more likely to evoke the terms "measurement" and "test construction." Those who are interested in the details of stimulus scaling could well consult the classical works of Guilford (1954), Torgerson (1958), and Woodworth and Schlossberg (1954). Despite their age, all three of these books describe the major models in unique step-by-step detail; more recent books have tended to concentrate on newer models.

Perhaps the main consideration in measurement is what kind of response is to be obtained from the subject, because this has profound effects on what subsequent analyses may be performed—one cannot analyze data that one has not obtained. There are two broad approaches, and both derive from psychophysics. In one, which originated with Gustav Fechner, subjects make only ordinal judgments as to whether a stimulus was seen or not and whether a comparison stimulus is more or less intense than a standard stimulus. The methods require very little of subjects. Indeed, animals can be trained to make requisite responses by means of such devices as bar pressing. In the other approach, most strongly associated with S. S. Stevens (see Chapter 1), subjects are required to use properties of the real-number system to make interval or ratio judgments, as by saying how much more intense a comparison stimulus was than a standard. Such methods normally require adults or older children.

A BRIEF INTRODUCTION TO PSYCHOPHYSICS

The overview defined psychophysics as the study of the relation between variation in physical dimensions of stimuli, which we will symbolize as Φ (for physical), and their associated responses, historically called "sensations," which we will symbolize as Ψ (for psychological). The physical dimension need not be intensity, but it will be for all examples in this chapter, and the associated responses will describe apparent intensity. We have already noted the obvious ordinal relation between the physical and apparent intensities of weights, flashes of light, and tones. A 5-pound weight obviously feels heavier than a 1-pound weight. In particular, the probability that a weak event will be detected also increases as the intensity increases. Psychophysics is concerned with making more detailed statements about the relations between Φ and Ψ which, as was also noted, are usually required by the problem under study. Three particular questions are historically important yet relevant to many contemporary problems:

1 What is the minimal energy needed for a particular event to be perceived under particular conditions, i.e., the absolute threshold or limen? For reasons to be noted below, this normally involves determining the stimulus event that is perceptible 50 percent of the time.

2 How different must two stimuli be in order to detect a difference between them or to determine which is of greater intensity? This involves what is variously called the "difference threshold," "difference limen," or "just noticeable difference" (JND) between a standard and a comparison stimulus.

3 How may the relation between physical intensity and its associated sensation be described in the interval or ratio terms of Chapter 1? This is known as the problem of psychophysical scaling.

The history of these questions is covered in several excellent books on the general history of experimental psychology (Boring, 1950; Robinson, 1981) because early experimental psychology was psychophysics. Simple but useful discussions of current applications may be found in any standard undergraduate textbook on perception such as Coren and Ward (1989). For a more detailed treatment, see Engen (1972a, 1972b) or Woodworth and Schlossberg (1954). Psychophysics is important for its own sake as exemplified by its use in such areas as communications engineering and photography. Audiologists perform psychophysical scaling on individuals in testing for hearing loss when they compare absolute thresholds they obtain with norms. An abnormally high threshold implies hearing loss. Psychophysics is limited to the study of relationships that hold when stimuli vary along a specified physical dimension such as sound intensity. Measuring intelligence, psychopathology, etc., is not psychophysical because no physical dimension underlies these attributes. Nonetheless, concepts like the threshold are applicable to psychometrics in general.

Psychophysical Methods

Methods used to gather psychophysical data were first developed by Fechner (1860/1966) to study the relation between mind and body. Later, J. M. Cattell, Fullerton (Fullerton & Cattell, 1892), Thurstone (1928), and others expanded upon their use.

Several psychophysical methods developed by Fechner are still widely used. One is called the method of constant stimuli. Assume that a tone whose physical intensity is 185 units is essentially never reported as being heard, but a tone whose physical intensity is 215 units is nearly always reported as being heard. The experimenter might choose to use intensities of 185, 190, 195, ..., 215 units. On each trial, one level (magnitude) is chosen at random for presentation. There is no limit upon the number of levels the experimenter may use. The levels need not be equally spaced and they need not occur equally often, but it is typical to use from 5 to 10 equally spaced and equally probable levels. The results are the probabilities of an affirmative response (e.g, saying the tone was heard) for each level.

Two related procedures are the method of adjustment and the method of limits. In the "method of adjustment," a standard is varied until it is barely sensed to determine an absolute threshold, or a comparison is made to barely differ from a standard to provide a difference threshold (JND). The method of limits takes two forms. The "ascending method" as used to determine an absolute threshold starts with a stimulus that is not sensed. The stimulus is progressively increased until it is sensed. The "descending method" starts with a stimulus that is sensed and decreases the intensity. The modification made to determine difference thresholds is straightforward. The comparison stimulus is presented either below (ascending method) or above (descending method) and incremented or decremented.

Absolute Thresholds

The original idea of an absolute threshold goes very far back in philosophy. It implied a "cut" in Φ—the subject never sensed the stimulus below the cut (threshold) and always detected it above the cut. Imagine that the method of constant stimuli is used to present a series of weights. This predicts a step function relating Φ to Ψ (in this case, the probability of reporting that the stimulus was sensed or detected), as illustrated in Figure. 2-1*a*. The general name given to any relation between Φ and Ψ is a "psychometric" (mind/measuring) function. This particular function describes local psychophysics, because Ψ is defined in terms of sensations in the location of the threshold. However, it is extremely unusual for data to provide a step function, which we will later show is of general importance to psychometric theory. The data will more likely resemble panel (b) of Figure 2-1*b*, known as an ogive or *S*-curve.

Figure 2-2*a* illustrates an ogive and its associated data points as simulated by methods defined below. Although several mathematical functions produce ogives and there are many explicit curve fitting methods (see Chapter 15), curve fitting can often be done by inspection. The point at which the curve crosses the .50 level for Ψ defines the absolute threshold. This is approximately 200 units in the present case.

In order to explain this lack of a step function, the original threshold hypothesis was modified to incorporate sensory noise. "Sensory noise" refers to random error in perceiving an event, causing a fixed stimulus to have variable effects on different trials. The process may be thought of as physiological in origin, but it need not be so viewed. The most popular specific conception of sensory noise is the phi-gamma hypothesis—numerous independent factors contribute to the error, and so it varies

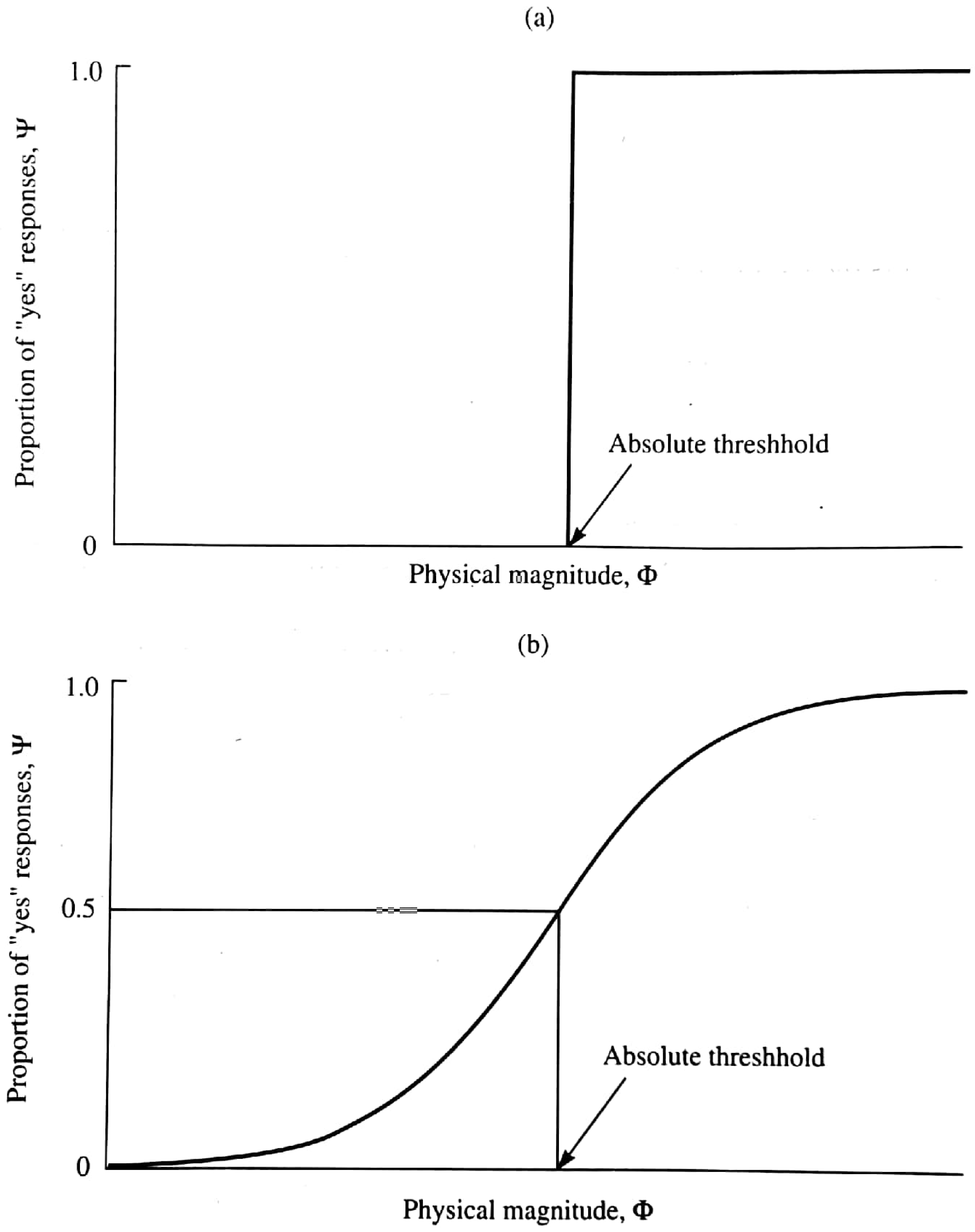


FIGURE 2-1 (*a*) A step function representing the initial concept of the threshold and (b) an ogive (*S* curve) representing a more realistic outcome.

normally over trials because of the central limit theorem (see Chapter 5). The specific form of the ogive (psychometric function) is the cumulative normal. An alternative model (Luce, 1959a, 1963) leads to a logistic function, defined below. Cumulative normals and logistic ogives are closely related mathematically and cannot be differentiated by eye. A third, but thus far less fruitful, possibility is neural quantum theory (Stevens, Morgan, & Volkmann, 1941). It leads to a linear function which will not be considered further. The 0.5 point that describes the absolute threshold is therefore arbitrary.

The location of the psychometric function is one of its two basic parameters. If auditory stimuli are used, the function of a subject with more acute hearing and consequently a lower threshold will fall to the left of the function of a subject with less

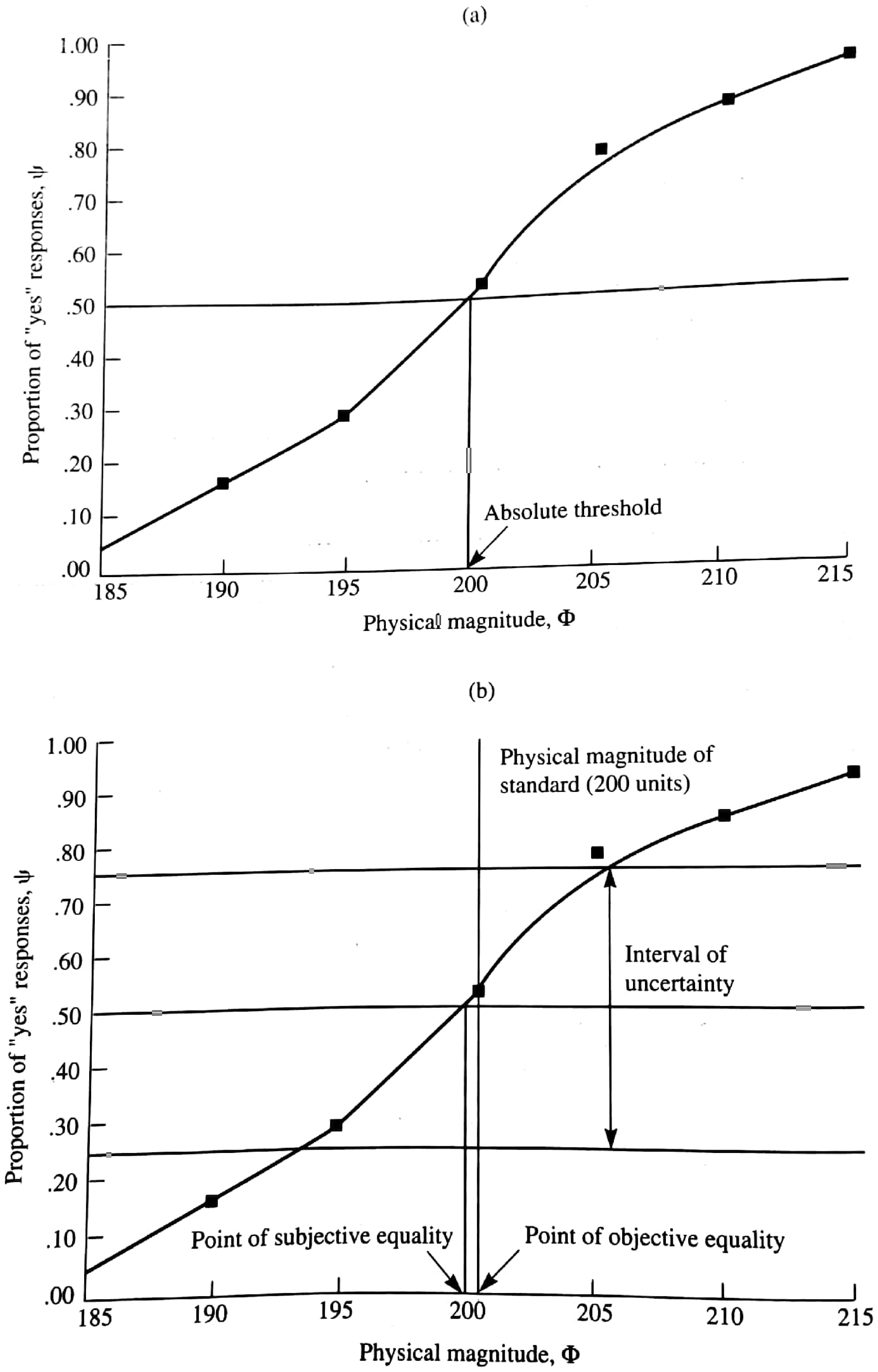


FIGURE 2-2 Psychometric functions derived from applying the method of constant stimuli (simulated data) to (*a*) absolute responses (detection) and (*b*) comparative responses.

acute hearing. Likewise, we hear tones at middle frequencies better than lower- or higher-frequency tones, holding intensity constant, so that middle-frequency tones produce psychometric functions to the left of higher- and lower-frequency tones. Location thus defines task difficulty. The second parameter of importance is the slope of the function or the extent to which it resembles a step function. The steeper the slope, the more discriminating the responses are. Quantities related to these two parameters play a crucial role in psychometric theory, as we will show later in this chapter.

Now consider a question like "Are you unhappy at life" on a depression inventory. The probability that this question will be answered in the affirmative should be quite low for people who are low in the attribute (not depressed) and increase with the level of depression until it reaches 1.0. This implies that there should be a level of depression for which the probability of endorsing the item is .5, and so it is meaningful to think of an absolute threshold associated with the item. Similar considerations hold for items for which there is a correct answer and the underlying dimension is course knowledge or general intelligence. We will exploit the generality of the threshold and psychometric function concepts, especially in this chapter and in Chapter 10. The fact that there are physical dimensions of weight, sound intensity, and light intensity, but none of depression, course knowledge, or general intelligence, might appear to reflect a major difference between psychophysical and other applications. However, as we noted in Chapter 1, such ostensive characteristics are not needed to provide a scale. The scaling models considered in this book allow dimensions that are not defined physically to be inferred.

Simulating a Threshold

The data in Fig. 2-2 were actually derived from a very simple computer simulation to illustrate the absolute threshold and sensory noise. We defined the absolute threshold as 200 units. Sensory noise was produced by choosing a random number from a normal distribution with a mean of 0 and a standard deviation of 10 in accord with the phi-gamma hypothesis. The mean of any given physical magnitude (Φ) was its physical value (185 to 215 in 5-unit steps), but it varied normally about this mean on any given trial. The sensory effect for a stimulus on any given trial equaled Φ plus the random number. We ran 100 trials per stimulus.

For example, the two random numbers obtained for the first two trials using the 195-gram stimulus were +20.6, and +2.8. These produce sensory effects of 215.6 and 197.8. If the effect equaled or exceeded the threshold value of 200, the subject said yes (the stimulus was felt); otherwise the subject said no. Consequently, the subject said yes in the first case and no in the second. Note that the sensory effect of any comparison stimulus *can* exceed 200, but the probability of this happening increases as its physical magnitude increases. The resulting proportions of yes responses (Ψ) for the seven stimuli were 0.07, 0.17, 0.35, 0.53, 0.66, 0.87, and 0.94, as plotted. The important point to remember is how sensory noise can cause physically unchanging stimuli to vary over trials.

Difference Thresholds

Defining a difference threshold (JND) is a bit trickier when the subject compares two stimuli in order to determine which is of the greater magnitude. The corresponding point at which the psychometric function is .5 describes the comparison stimulus perceived as equal to a standard half the time, not the threshold. This is called the point of subjective equality. Its value need not match the physical magnitude of the standard (the point of objective equality). For example, suppose the standard and comparison stimuli in Fig. 2-2*b* were weights of different density, e.g., were lead versus wood. A 200-gram lead standard stimulus would obviously be much smaller than a 200-gram wood comparison, and so there might be an illusory difference in weight. The two weights might have to differ in physical magnitude to appear equal.

The "interval of uncertainty" is that range of stimulus differences for which judgments can "go either way" and is usually taken from .25 to .75 on the function, as illustrated in Fig. 2-2*b*. The concept also applies to absolute thresholds, even though that is not depicted here. The difference threshold (not presented in the figure) is usually defined as half this interval of uncertainty, again by convention. The key to both types of threshold is the varied psychological effect of a fixed physical stimulus due to sensory noise.

It is possible to simulate a difference threshold in a manner similar to the absolute threshold. However, sensory noise would affect both the standard and the comparison. Although this might seem to decrease subjects' ability to make judgments, this need not be the case. The covariance (or correlation) between the two noise sources is also important for reasons that will become clear when we consider the logic Thurstone (1928) used to develop his discriminant model.

The Weber Fraction, Fechner's Law, and Psychophysical Scaling

E. H. Weber noted an important property of the JND which was the main stimulus to Fechner's subsequent ideas—its magnitude is proportional to the standard against which it is derived. Subsequent research indicates that his findings are a good first approximation for a wide variety of sensory dimensions as long as the standard is not extremely weak or strong. Thus, suppose he found that a 1.05-gram weight was just noticeably different from a 1-gram standard weight so that the JND was 0.05(1.05 – 1) grams. The Weber fraction is the JND divided by the magnitude of the standard (Φ), or 0.05/1 or 0.05 in this particular case. Weber's results were that a 10.5-gram comparison stimulus was just noticeably heavier than a 10-gram standard, a 105-gram comparison was just noticeably heavier than a 100-gram standard, etc. His results may be generally stated as $\Delta\Phi/\Phi$ equals a constant where $\Delta\Phi$ is the physical magnitude of the JND associated with a given Φ.

Suppose that Weber's law had held exactly, a 1.0-unit standard was also the absolute threshold, and the fraction was 0.05. A 1.05-unit comparison will be 1 JND more intense than this standard. Now, let the resulting 1.05-unit stimulus become a new standard. A 1.10, i.e., 1.05(1 + 0.05) unit comparison will be just noticeably more intense. Keep repeating the process of obtaining a stimulus that is 1 JND more intense by multiplying by 1.05 and use it as the next standard. The resulting values will be 1.16, 1.22, 1.28, 1.34, ..., to two decimal places. It does not matter what type of stimulus is being judged.

Fechner made what in essence is a simple yet dramatic (and controversial) proposal: Let each of these steps, separated by a JND, define equal units on an interval scale of sensation. A corollary is that one can speak of two stimuli in terms of how many JNDs separate them—2.3, 0.5, or whatever. Mathematically, this relationship can be expressed as Eq. (2-1), which is called Fechner's law:

$$\Psi = b \log(\Phi) + a \tag{2-1}$$

where Ψ = scale value of the sensation (apparent magnitude)
Φ = physical magnitude
b, a = scaling constants

Neither scaling constant is important to our discussion; a is commonly chosen to make $\Psi = 0$ when Φ is at threshold, but this is usually not viewed as a rational zero in the ratio scale sense. Figure 2-3*a* depicts Fechner's law. Unlike Figs. 2-1 and 2-2, values of Ψ need not fall near threshold. The relation applies to the entire physical dimension (Φ) and is known as global psychophysics.

Logarithmic functions have several important characteristics. The one particularly important for our purpose is that equal physical ratios yield equal sensory differences. Suppose stimuli *a, b, c,* and *d* are, respectively, 10, 20, 100, and 200 grams. Since *a/b* = *c/d, a* and *b* are just as many JNDs apart from each other as are *c* and *d.*

Fechner's methods are called indirect methods because subjects do not define sensory magnitudes directly, and discriminant methods because they concern the subject's ability to discriminate. They are also called confusion methods because scale values require that stimuli generally be confusable with one another in magnitude.

Direct Psychophysics and the Plateau/Stevens Tradition

Coren and Ward (1989) described a test of Fechner's law made by Plateau in 1872. He had artists mix black and white pigments to make a gray appear midway between the two. Fechner's law predicts that the gray's intensity should be the average of the black's intensity and the white's intensity. Plateau obtained a systematic departure in that the grays fell near the cube roots of the two other intensities. Four important things about Plateau's research and Stevens' (1951, 1956, 1975) subsequent extensions are that (1) unlike Fechner's approach, subjects respond directly through subjective estimates; (2) equal physical ratios provide equal sensory ratios and not differences with these subjective estimates; (3) equal numbers of JNDs between different pairs of stimuli are not equal appearing, the emphasis is upon global and not local psy-

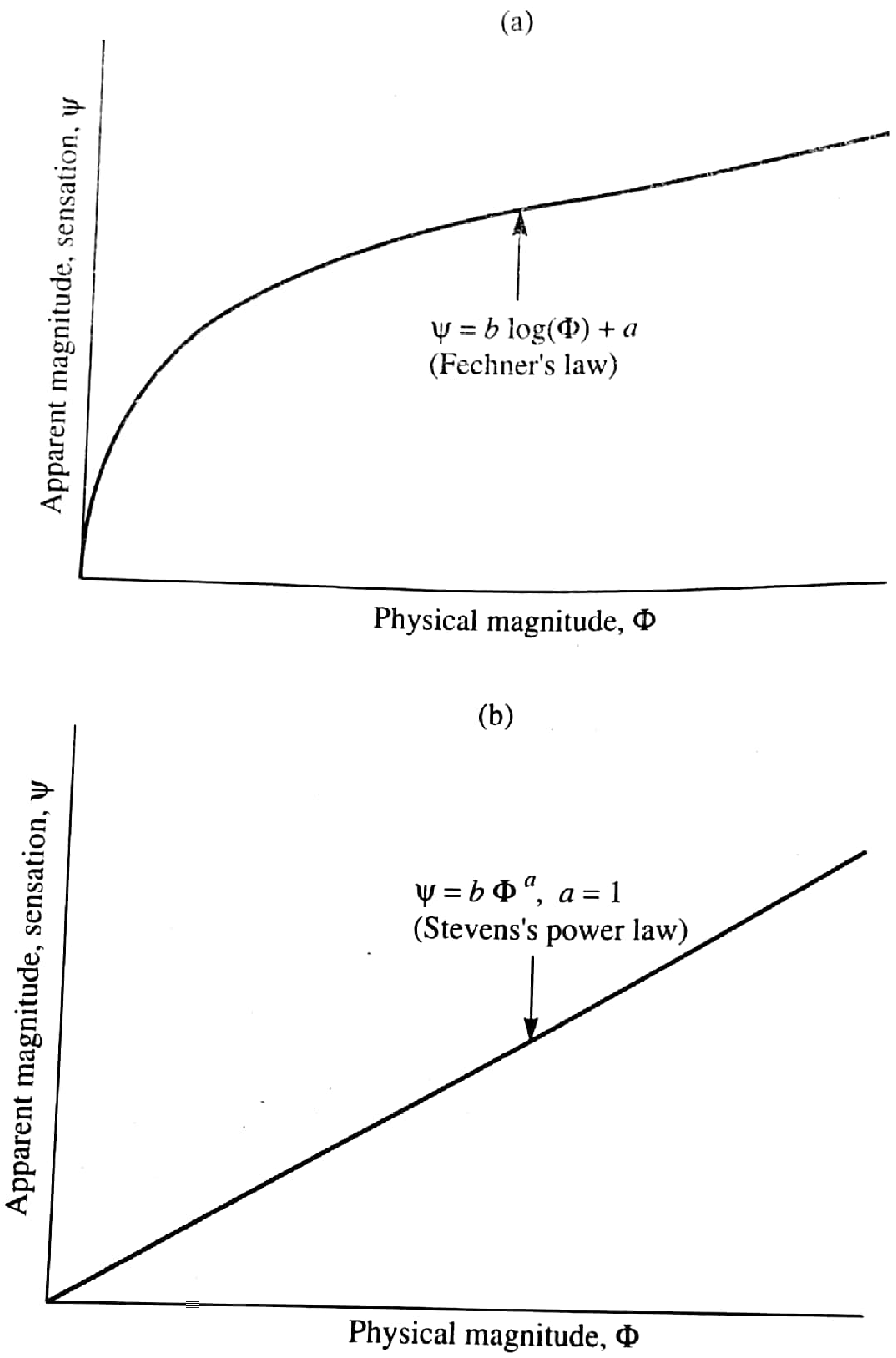


FIGURE 2-3 (*a*) Fechner's logarithmic law for indirect psychophysics, (*b*) Stevens' power law for direct psychophysics with an exponent $a = 1$.

chophysics. Point 2 may be stated as Eq. 2-2, called Stevens' law, since he examined it so thoroughly, or the power law from its mathematical form:

$$\Psi = b\Phi^a \tag{2-2}$$

where Ψ = scale value of the sensation (apparent magnitude)
Φ = physical magnitude
b = scaling constant

The *a* parameter is more complex. It describes the sensory ratio associated with the physical ratio of two stimuli that differ along the physical dimension in question, Φ. Let the two stimuli be *x* and *y*, their associated sensory ratio be Ψ_x/Ψ_y, and their phys-

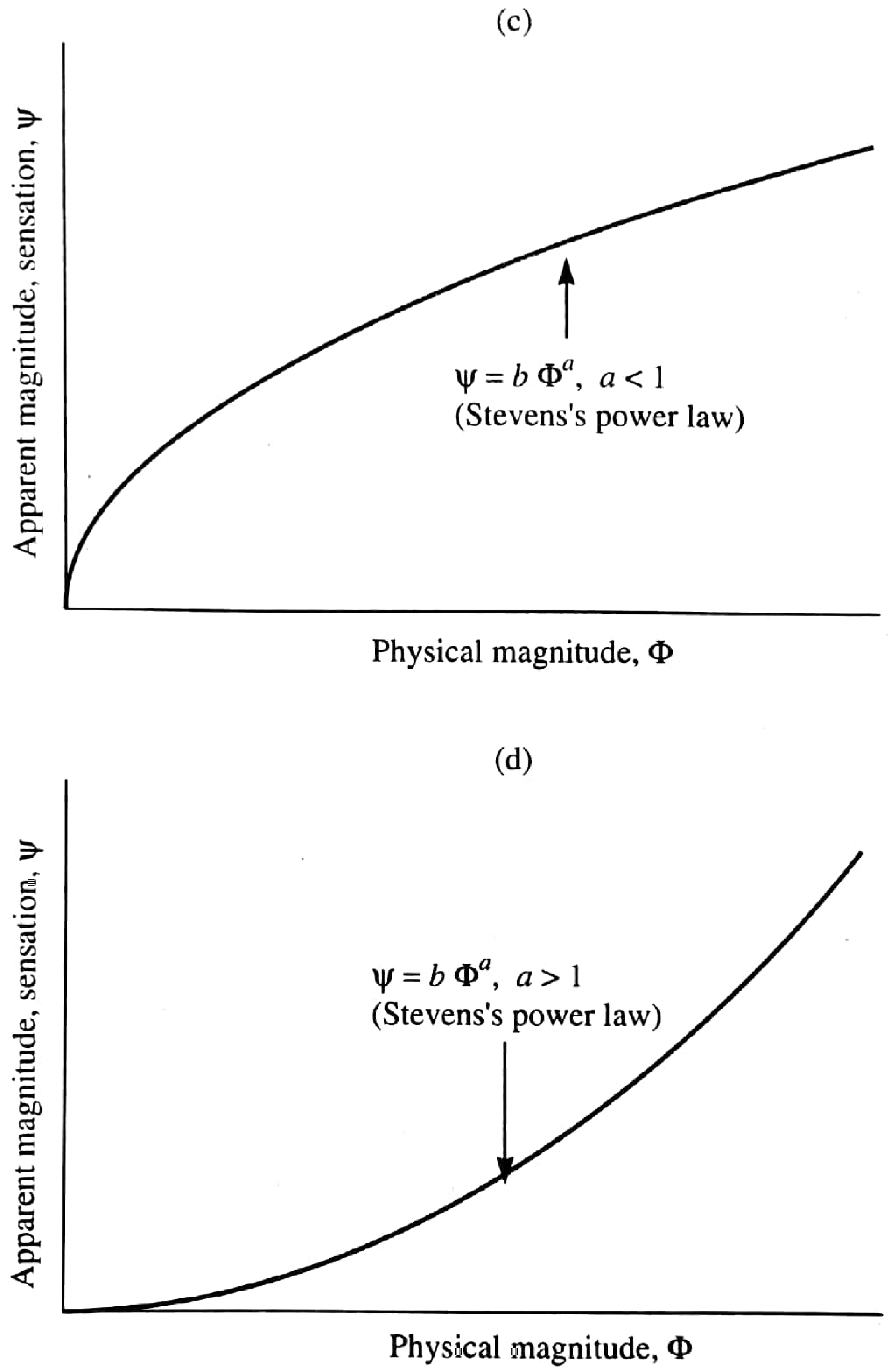


FIGURE 2-3 (*c*) Stevens' law with an exponent $a < 1$, and (*d*) Stevens' law with an exponent $a > 1$.

ical ratio be Φ_x/Φ_y. If the two are the same ($\Psi_x/\Psi_y = \Phi_x/\Phi_y$), the relation is linear; $a = 1$. For example, doubling the duration of a noise also makes it appear to last twice as long. However, the sensory ratio is smaller than the associated physical ratio for most dimensions ($\Psi_x/\Psi_y \; \Phi_x/\Phi_y$), and so $a < 1$. The brightness (apparent intensity) of many light sources increases only as the cube root of the change in physical intensity. This means the physical intensity of two lights must be in an 8:1 ratio for the more intense light to appear twice as bright. Finally, a few sensory ratios are larger than their associated physical ratios ($\Psi_x/\Psi_y > \Phi_x/\Phi_y$), and so $a > 1$. If one electric shock is physically twice as powerful as another, it will actually appear more than 10 times as intense. Stevens and his associates devoted many years to a thorough study of different sensory modalities. In particular, Stevens (1961) "cataloged" the exponents of various dimensions. Figure 2-3*b* through 2-3*d* depict these three outcomes ($a = 1$, $a < 1$, and $a > 1$). Note that even though the function for $a < 1$ resembles Fechner's law in being concave downward, the two are quite different. Data fitting Fechner's law become linear when the abscissa, but not the ordinate, is logarithmic (semilog graph paper), and data fitting a power law become linear when both axes are logarithmic (log-log graph paper), regardless of the magnitude of the exponent. The slope of the line in the latter case defines the magnitude of the exponent. Although Fechner and Stevens' laws were once regarded as competitors (investigators commonly asked which one was "right"), it is now generally recognized that the Ψ of Fechner's law for discrimination need not be the same as the Ψ of Stevens' law for subjective estimates, and so there need be no incompatibility. Indeed, the two would be completely compatible if Stevens' Ψ were the logarithm of Fechner's (Luce, 1963).

Stevens also developed several methods for inferring the exponents and showing that any given estimate was not an artifact of a single method; i.e., he used converging operations as defined above. The most commonly used of these methods are the following:

1 *Ratio production.* A subject is shown a standard stimulus and is then asked to adjust a comparison so that it appears in a specified ratio to the standard. The simplest and most common ratio is 2:1, so that the subject is asked to make the second stimulus appear twice as intense. If, for example, the comparison has to be physically four times as intense, the ratio *(a)* will be .5. However, the subject might also be asked to make the second stimulus three times as intense.

2 *Ratio estimation.* The subject is shown standard and comparison stimuli and asked to define the ratio of their apparent intensities. Thus, they might report that a comparison tone is 1.5 times louder than a standard tone.

3 *Magnitude estimation.* The subject is shown a single stimulus and simply asked to define its magnitude numerically. Usually, subjects are also shown a different stimulus, called the modulus, which is given an assigned value to fix the units of the scale, making it somewhat similar to ratio estimation.

4 *Bisection.* As in Plateau's experiment, subjects are shown two stimuli and asked to adjust a third so that it appears midway between the first two. Unlike other subjective estimates, bisection requires interval rather than ratio judgments.

5 *Cross-modal matching.* The subject is presented a stimulus in one modality and asked to adjust a stimulus in another modality to apparent equality. For example, the task might be to make a tone appear as loud as a light is bright. As bizarre as the task may seem, the exponent relating the two modalities is predictable from the exponents inferred from the other tasks. For example, the sweetness of a sucrose solution and the apparent thickness of wood blocks both have exponents of about 1.3. Suppose a given sucrose solution is matched with a given thickness. Then the concentration of the sucrose is then doubled. According to Stevens' power law, the matching wood block should seem twice as thick, which it does.

In all methods, the procedure is repeated with different stimuli in order to determine the consistency of the exponent.

Although it is not associated as strongly with the Stevens tradition as the above, the method of equal-appearing intervals (category scaling) also tends to fit Stevens' power law (Marks, 1974; Ward, 1974). Subjects simply sort stimuli into categories so that the intervals between category boundaries appear equal. In particular, the sensory difference between the upper and lower boundaries of each category should be the same.

The Fullerton-Cattell Law

The Fullerton-Cattell (Fullerton & Cattell, 1892) law is a basic link between Fechnerian indirect psychophysics and psychometrics in general. It states, simply and euphoneously, that equally often noticed differences are equal unless always or never noticed. This is certainly true in the psychophysical case since the unit (the JND) is defined by equally often noticed differences. The significance of the Fullerton-Cattell law is that it does not depend upon how the stimuli differ or on the basis of the judgment. In particular, the ">" relationship that meant brighter, heavier, or louder above can also mean "is more preferred," among other things. If you prefer bananas to apples 75 percent of the time and apples to pears 75 percent of the time, the distance between apples and bananas and the distance between apples and pears may be assumed equal; i.e., apples are at the midpoint of a scale defined by these three stimuli. The "always or never" part is simply a caveat that one cannot draw inferences when there is no confusion over trials: If you always prefer bananas to apples and always prefer apples to pears, their relative distances cannot be inferred from these data alone. However, if you sometimes prefer plums over each and sometimes not, a scale can be constructed.

Signal Detection Theory and Modern Psychophysics

In early studies of the absolute threshold, a stimulus was always presented. Subjects, who were often also the investigators, typically knew this but were trained at analytic introspection to report their sensations and to ignore this knowledge. Sometimes, however, the equipment would malfunction and fail to produce a stimulus, but subjects might say "Yes, I saw (heard, felt, etc.) it," thus committing the stimulus error by responding on the basis of their conceptions of the stimulus rather than the sensation itself. Gradually, "catch" trials were regularly used to "keep subjects on their toes," but no systematic use was made of the data obtained on these trials since the purpose of the experiments was to measure sensations.

Measuring sensations was the exclusive goal of nineteenth-century psychophysical research and is often a valid goal today, but it is not the only goal. Reflecting a variety of factors such as the behavioristic rejection of mental states like sensations, much of psychophysics eventually became concerned with subjects' ability to discriminate the presence of stimulation from its absence. A particular tradition emerged known as the theory of signal detection (TSD) (Egan, 1975; Green & Swets, 1967; Macmillan & Creelman, 1991; Swets, 1986a, 1986b; Swets, Tanner, & Birdsall, 1961; Tanner & Swets, 1954). It bears a close kinship to Thurstone scaling, and we will consider it in more detail in Chapter 15. For the present, it is most important in helping to illustrate thc difference between the classical psychophysics of judging sensations and the more modern emphasis upon accuracy of discrimination.

TSD has proven particularly important because of its emphasis upon assessing response bias or differential willingness to use the response alternatives independently of sensitivity or accuracy at discrimination. Threshold measures using psychophysical procedures derived from Fechner are particularly influenced by a subject's willingness to report having sensed the stimulus. A practical example of a response bias involves the diagnostic accuracy of two clinicians who see the same set of patients. Clinician A correctly diagnoses 90 percent of the patients determined to have a given disorder on the basis of some appropriate method, but clinician B diagnoses only 80 percent of the patients correctly. Does this mean that clinician A is the better diagnostician? The data are insufficient since only their hit (true positive) rates in identifying those who have the disease are known. We also need to know the false alarm (false positive) rates of diagnosing normals as having the disease. Perhaps clinician A has a false alarm rate of 90 percent, in which case he or she is just blindly guessing the presence of the disease in 90 percent of the population. If this is true and if clinician B's false alarm rate is less than 80 percent, clinician B could be the better.

TYPES OF STIMULI AND RESPONSES

Endless distinctions could be made about stimuli and responses that are important to psychometrics, but we will consider only the most important. Most are derived from psychophysics.

Judgments versus Sentiments

Although no two words perfectly symbolize the distinction, the distinctions between what we call "judgments," where there is a correct response, and "sentiments," which involve preferences, is very basic. There are correct (veridical) versus incorrect answers to "How much is two plus two?" and "Which of the two weights is heavier?" There may also be degrees of correctness, as in line-length judgments of visual illusions. In contrast, sentiments cover personal reactions, preferences, interests, attitudes, values, and likes and dislikes. Some examples of sentiments include (1) rating how much you like boiled cabbage on a seven-category Likert scale, (2) answering the question, "Which would you rather do, organize a club or work on a stamp collection?" and (3) rank-ordering 10 celebrities in terms of preference. Veridicality does not apply to sentiments—a subject is neither correct nor incorrect for preferring chocolate ice cream to vanilla ice cream. This distinction is very close to the difference between making discriminations in TSD and reporting sensations in classical psychophysics. Judgments also tend to be cognitive, involving "knowing," whereas sentiments tend to be affective, involving "feeling."

Ability tests nearly always employ judgments regardless of whether an essay, short-answer, multiple-choice, or true-false format is used. Coversely, tests of interests inherently concern sentiments as the subject identifies liked and disliked activities. Attitudes and personality measures can use either form. Items like "Do you like going to parties?" involve sentiments, but items like "How often do you go to parties?" are essentially judgments. The distinction may be obscured because the perceived frequency may reflect preference as well as actual frequency.

Social desirability may bias sentiments in the signal detection sense so that the popularity of socially endorsed behaviors may be overestimated. This is less likely to be a problem with judgments. However, the internal consistency or extent to which items measure the same thing is important to both. Temporal stability or the extent to which the measure tends to remain the same over time may or may not be important. Chapters 6 through 9 consider how these statistics are obtained. In general, the logic of using judgments is generally clearer than the logic of using sentiments because of advantages inherent in having a correct response. Other terms are frequently employed to describe these two categories. Goldiamond's (1958) distinction between what he called "objective" and "subjective" indicators of perception corresponds in essence to the judgment-sentiment distinction. The word "choice" is frequently used in place of the word "sentiment."

Absolute versus Comparative Responses

In general, an absolute response concerns a particular stimulus, whereas a comparative response relates two or more stimuli. The distinction applies to both judgments and sentiments. "How many concerts have you been to in the past year?" versus "Have you been to more concerts than movies in the past year?" illustrates this distinction for judgments. Likewise, "Do you like peas?" versus "Do you like peas more than you like corn?" involves sentiments.

One of psychology's truisms is that people are almost invariably better (more consistent and/or accurate) at making comparative responses than absolute responses. This is because there is a frame-of-reference problem present to at least some extent in absolute responses that is avoided in comparative responses. Asking a consumer "Is this cola sweet?" raises the question of how sweet is sweet that is avoided when one is asked to judge which of several colas is the sweetest since the criterion of sweetness can be applied equally to all colas. One possible application of this principle is in ability testing. If there are no "none of the above" or "all of the above" alternatives, multiple-choice tests are comparative judgments of the relative truth of the alternatives. We suggest (and some disagree) that these alternatives be avoided because they compromise the comparative nature of the test by asking whether none or all of the other alternatives are true in an absolute sense. Similarly, true-false tests are absolute judgments of the truth or falsity of a single item, and we suggest the use of multiple-choice questions for this and other reasons to be considered.

People rarely make absolute judgments in daily life, since most choices are inherently comparative. There are thus few instances in which it makes sense to employ absolute judgments. One important exception is when absolute level is important, as in attitudes toward various ethnic groups. A subject could, for example, rank various groups from most to least preferred. However, the subject may dislike all the national groups or like them all, which would not be apparent from the comparative rankings. Absolute responses are especially important when some indicator of neutrality is needed. For example, people who are more neutral with respect to candidates in an election are probably more susceptible to influence and change than those who have a clear preference. By requiring absolute responses from subjects, one is able to approximate a neutral point.

Another case in which it makes sense to phrase items in absolute terms is when the inherent ambiguity of absolute judgments is of interest. For example, the MMPI contains several items like "I often have headaches" and "I frequently have trouble falling asleep." A psychologist is probably not actually interested in the actual frequency of headaches or sleepless nights. If he or she were, more objective tests could be developed through clinical observation. The issue is how the patient interprets words like "often" and frequently." Absolute judgments are perfectly appropriate in that case.

Absolute responses are also useful because they are much easier and faster to obtain than comparative responses. For example, the method of paired comparisons is an extremely powerful way to gather data. A market research example could involve preferences among K brands of cola. The subject is given two brands in succession and asked to state a preference. This is repeated for all possible pairs of brands. Unfortunately, this requires anywhere from $K(K-1)/2$ pairs (if a given brand is presented in only one of the two possible positions in the pair) to K^2 pairs (if all brands appear in all orders and a given brand is paired with itself). The number of comparisons increases rapidly with K. For example, if there are 20 brands in the study, from 190 (20)(19/2) to 400 (20^2) trials are required per subject. However, it is much quicker to have subjects rate each brand individually. Any of several scaling models can be used to obtain interval estimates of preference from each cola's average rating over subjects. Conversely, paired comparison methods generally give much more reliable results when applicable.

To the extent that a person answering an item phrased absolutely has a criterion to define terms like "frequently," "seldom," or "hardly ever," the judgment becomes partly comparative. Individuals generally have feelings about their absolute liking for an object or activity, but such sentiments are influenced by the range of objects or activities available. An individual who rates how much they like boiled cabbage probably thinks "What else is there to eat?" Differences among subjects and/or time contribute to unreliability. However, temporal instabilities can be of interest in themselves (Spielberger, Gorsuch, Lushene, 1970).

If an absolute format is appropriate, anchoring by specifying the meaning of the response scale is generally important to reducing unwanted error due to differences in implicit bases of comparison. For example, instead of simply asking subjects to rate how often they go to the movies on a five-point scale, indicate that 1 means once a month or less, 2 means at least once a month, etc. (the actual anchors should be developed by pretesting). Similarly, if a pretest reveals that subjects nearly always answer the question "I absolutely adore rutabagas to the point that I must eat them daily" causes everyone to respond in the negative, change the anchor to favor a higher incidence of positive responses, such as "I would eat rutabagas if they were served to me." Not all situations demand anchors, as in the MMPI example where the ambiguity was intentional.

Preferences versus Similarity Responses

Different methods are required to study responses denoting which stimuli are preferred versus most similar. Preference responses are also known as dominance responses. Examples of these responses (which are nearly always sentiments) include which stimulus is most liked, tastes best, is least filling, would be most likely purchased, etc. Similarity responses denote which stimuli are most like one another. Preferences are clearly asymmetric; preferring A to B means not preferring B to A. In contrast, similarity responses are normally symmetric—saying A is similar to B implies that B is similar to A (Chapter 15 will consider an interesting exception). Thurstone scaling, described below, requires preferential data. However, the most common methods of analysis, (e.g., factor analysis, and multiple and partial correlation) require similarity data because they are based upon the ordinary Pearson correlation coefficient (Chapter 4), a measure of similarity rather than preference.

Specified versus Unspecified Attributes

By definition, psychophysical responses are obtained with respect to an attribute defined by the experimenter. This may also be the case when the attribute is not a single physical dimension. For example, a marketing study may ask which of several packages differing in height, width, and depth looks largest even though all contain the same volume. Conversely, subjects may be asked to evaluate similarities or preferences among stimuli without being told in what respect. If the stimuli clearly differ in a single, dominant respect, instructions may be unnecessary. However, if the stimuli are multidimensional, the goals of the experiment dictate whether or not some particular attribute should be specified. The study may concern how well subjects ignore a given attribute, so that it is important to tell him or her which attribute is critical. On the other hand, subjects should not be told, implicitly or explicitly, if the goal is to find out which actual attributes subjects actually use.

METHODS FOR CONVERTING RESPONSES TO STIMULUS SCALES

Fechnerian methods, which provide ordinal data, Stevens' methods, which provide interval or ratio data are applicable outside the confines of psychophysics. Keep in mind that the level at which data are gathered may well differ from the level of the resulting scale, particularly for Fechnerian methods. Scaling models often take data obtained at one level and transform it to a higher level, most specifically to produce an interval scale from ordinal data. Of course, data gathered at a ratio level need not be transformed. One part of successful scaling involves choosing an empirical procedure that is appropriate to the subjects' ability to respond; another part is to use a scaling model appropriate to the resulting data.

Ordinal Methods

In general, the simplest way to obtain ordinal data is the method of rank order in which subjects rank stimuli from "most" to "least" with respect to the specified attribute.

In the A-B-X method, subjects are presented with stimuli A and B followed by a third stimulus (X) which is either A or B. The subject is asked to say whether X is A or B. The process is repeated, comparing all pairs of stimuli. The probability of confusing any two stimuli is an ordinal measure of their similarity. This method is particularly useful in scaling stimuli that are difficult to describe. For example, suppose Alpha Cola and Beta Cola are fairly similar in taste, but both differ somewhat from Gamma Cola. Subjects' A-B-X judgments may be only 60 percent correct when Alpha and Beta are paired (50 percent is chance), but 80 percent correct when Alpha and Gamma are paired and 85 percent correct when Beta and Gamma are paired.

In contrast, the method of triads uses three different stimuli which may all be highly discriminable from one another and asks which two are most similar. For example, the subject might taste string beans, lima beans, and green peas. It is probable that lima beans and green peas would be found to be the most similar pairing. The data obtained from all possible triads in a larger set (the number of combinations of K things taken three at a time) provide similarity rankings.

In the method of successive categories, the subject sorts the stimuli into distinct piles or categories that are ordered with respect to a specified attribute. For example, subjects could sort the U.S. presidents into five piles ranging from "very effective" to "very ineffective." This information can be obtained most easily by having the subjects mark a printed rating scale. This method has many variants depending on the information sought by the experimenter. If the experimenter is seeking only ordinal information, the subject may be allowed free choice as to the number of stimuli per category and number of categories. In contrast, the categories may be constrained to appear equally spaced in the method of successive categories. Sometimes, subjects are required to place an equal number of stimuli in each category. Perhaps the most important variant is the Q sort where subjects sort the stimuli so that the distribution of stimuli in successive piles forms a normal distribution. These methods necessarily provide numerous tied ranks. Thus if stimuli are placed in a series of categories, those in the first category can be thought of as tied for the top rank. Averaging over subjects eliminates most of these ties.

Interval Methods

The primary methods used to obtain interval data from subjects are variations upon the method of successive categories and Stevens' methods of bisection. This involves instructing the subject to use the scale as though the distances between successive categories were the same; e.g., the difference between a rating of 2 and 4 is equal to the difference between a rating of 6 and 8. Frequency anchors are also employed. For example, pleasantness could be anchored with adjectives ranging from "extremely pleasant" to "extremely unpleasant." Rating anchors also may be expressed as percentages to further ensure the interval nature of the responses so that subjects can be asked what percent of the general population they feel agrees with each of a series of statements.

The method of bisection may be applied outside psychophysics as follows. Subjects may be given two statements differing in how favorable they are toward the President and asked to select another statement from a list that falls closest to halfway between them. Rather than bisecting the distance between the two stimuli, other ratios may be used, as in psychophysics. For example, subjects may be asked to select a stimulus X such that the interval between one of two fixed stimuli and X appears twice as great as the distance between the two standards. Another approach is to present subjects with two stimuli that are at the extremes of the attribute and have them judge the ratio of intervals formed when a third stimulus is inserted.

In all these methods, the subject evaluates *intervals* of judgment or sentiment. Even though he or she may describe 1:1 ratios in the method of bisection, these ratios are not formed with respect to the absolute magnitudes of the stimuli as in ratio scaling. The experimenter might eventually use a scaling model to obtain these absolute magnitudes, but it is important to maintain the distinction between what the subject is required to do and the experimenter's use of the data in a scaling model.

Ratio Methods

Ratio methods require subjects to evaluate the absolute magnitudes of stimuli. For example, subjects may be given the name of a food liked moderately well by most people and asked to name a food liked twice as much, half as much, etc. Note that in ratio production, the subject generates the actual stimulus, unlike in other ratio methods. This may be somewhat difficult outside psychophysical applications.

If a zero point can be taken seriously, previously described percentage scales can be employed for ratio estimation. For example, subjects might rate the complexity of 100 geometric forms. The stimulus rated as most complex in pilot research is used as a standard, and the other stimuli are rated in relation to this standard on a percentage scale. If the least complex form is rated at 20 percent, its scale value will be .20, where the standard is 1.0. These ratio scales closely resemble scales obtained from more direct ratio estimation methods (Stevens, 1951, 1958, 1960).

Interval and ratio estimation methods may appear superficially similar. For example, choosing a stimulus that is halfway between two others (bisection) seems similar to choosing a stimulus that is twice as great as another (ratio production). In both cases, the subject forms two equal-appearing intervals. The important difference between these two methods is that the lower interval is bounded by a *phenomenal* zero in ratio production. The subject is essentially required to form an interval between two stimuli that is equal to the interval between the less intense stimulus and zero. Moreover, if subjects are sophisticated enough to provide interval judgments, they can also usually provide ratio judgments, making interval methods somewhat unnecessary.

MODELS FOR SCALING STIMULI

The next step in scaling is to generate an ordinal, interval, or ratio scale as desired. The models considered in this chapter are considered classical primarily because they have been available for a long time. They may also be considered classical because they provide relatively simple closed-form solutions and therefore do not require a computer (in practice, computers would probably be used). In contrast, modern psychometrics, considered in Chapter 10, usually requires open-form estimation.

Ordinal scales do not require complex models, and the various methods of gathering data and scaling usually produce the same rank ordering. In general, simply average individual subjects' ranks and rank-order the average of these ranks. This final set of ranks is the desired ordinal scaling of a modal subject.

In paired comparison methods, the first step to determine the percentage of subjects that rate each stimulus as being higher on the particular response dimension than each of the other stimuli. Thus, each of 10 stimuli produce 9 percentages comparing that stimulus to the rest. The full data from the group of subjects are summarized by a square matrix containing all possible percentages of paired comparison preferences. These percentages are summed for each stimulus (column of the matrix), and these sums are then ranked from highest to lowest.

Formal scaling models are more important in constructing interval (the more common situation) or ratio scales. The remainder of this section will consider models used for these purposes. They fall into two broad classes of models paralleling the distinction between Fechnerian indirect (discriminant) methods and Stevens' direct (subjective estimate) methods. Stevens' approach will be discussed first because it is simpler.

Direct (Subjective Estimate) Models

Direct models are usually close to the data because the experimenter takes the subject's interval responses (e.g., bisections or ratio responses, magnitude estimations, ratio estimations, ratio productions) seriously. Often, the experimenter needs only to average responses over repeated measurements of one individual to obtain an individual scale or, more commonly, over subjects in a group to obtain a group scale. The Stevens tradition, like the Fechner tradition, recognizes variability from sensory noise but simply as error rather than as an intrinsic part of scaling.

One example is to use the aforementioned method of equal-appearing intervals. Subjects might sort 100 occupations into 10 successive categories ranging from least to most prestigious. The subjects are instructed to treat the 10 numbered categories as an interval scale. Error is minimized by averaging judgments over subjects or occasions. Thus "psychology professor" may be rated 9, 9, 8, and 8 by four subjects. This yields an average rating, and therefore a scale rating, of 8.5 on the interval scale. Measurements are obtained in a like manner for the 99 remaining occupations. This scale may then be used in any situation requiring an equal-appearing interval scale, e.g., to study the relation between job prestige and job satisfaction. A ratio scale can be formed in a like manner using ratio production. For example, one occupation (e.g., dentistry) can be anchored at 50 and subjects asked to rate others as ratios relative to this norm. See Stevens (1958, 1960) and the Suggested Additional Readings for further details. It is important to test the assumption that the subjects are behaving consistently. One important statistic is the internal consistency reliability (homogeneity) of the data. Chapters 6–8 will illustrate the process.

Indirect (Discriminant) Models

+ Although the logic traces back to Fechner, Fullerton, Cattell, and others, L. L. Thurstone's law of comparative judgment (Thurstone, 1928) is the foundation of modern discriminant models. This law takes on numerous forms depending upon more specific assumptions. We will consider only the basic ideas and stress the single most popular model. A more complete discussion may be found in Bock and Jones (1968), Guilford (1954), and Torgerson (1958). The law of comparative judgment led to signal detection theory and general recognition theory (Ashby & Townsend, 1986; Ashby & Perrin, 1988, see Chapter 15).

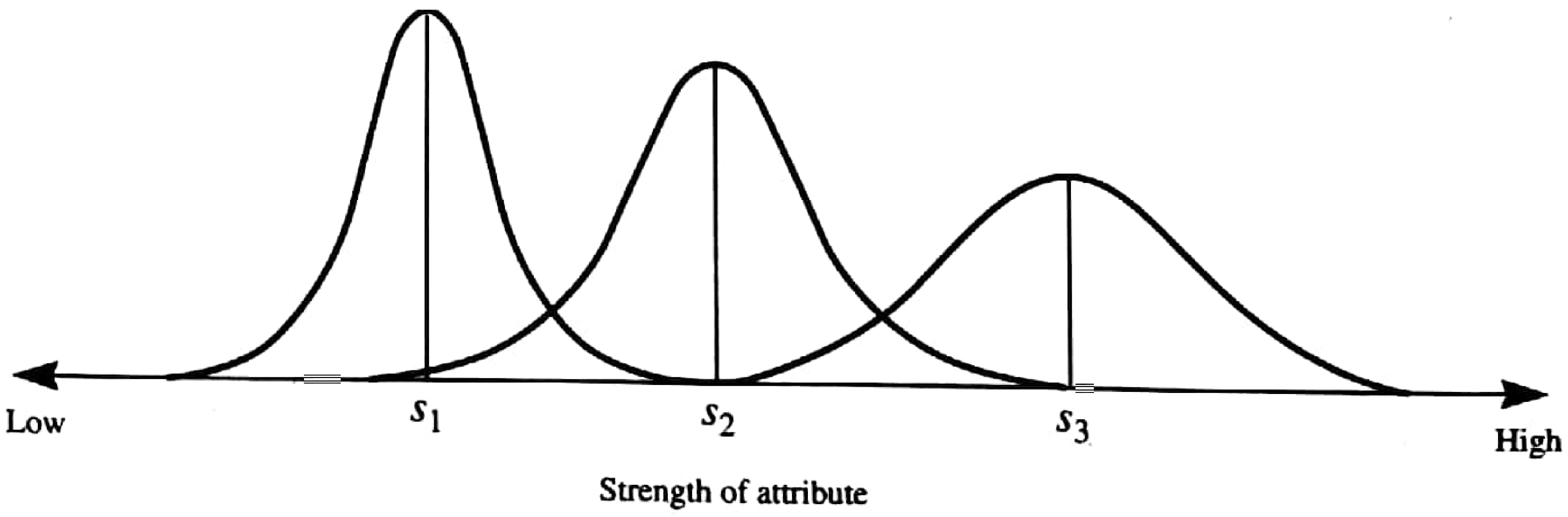
Although the same computational procedures can be applied to testing one individual repeatedly by pooling individual data, we will illustrate the logic here with the classic example of how one individual's subjective rank orderings can be "brought into the open" as an interval scale. Any stimulus is assumed to yield a discriminal process with respect to a specified attribute. The "discriminal process" is simply a broadly defined reaction which correlates with the intensity of the stimulus on an interval scale for an attribute. Because of what is equivalent to sensory noise, each stimulus has a discriminal distribution (discriminal dispersion) which reflects the variation in response to that stimulus. The model assumes the phi-gamma hypothesis by assuming reactions to a given stimulus are normally distributed, as shown in Fig. 2-4.

These distributions and the attribute continuum on which they fall, most simply called a "strength axis," are entirely hypothetical. Unlike psychophysics, the experimenter cannot locate the stimuli directly on the attribute—any model would be unnecessary if this could happen. Only after the experimenter makes a series of assumpions about what is going on in the subject's head and about the statistical relationship of such covert reactions to the hypothetical dimensions can a suitable model be formulated.

The mean discriminal process (reaction) to each stimulus is the best estimate of the scale value of that stimulus in several senses, such as most likely and least squares (see Chapter 4). If all stimulus means were known, an interval scale would complete the scaling problem, which is unfortunately not directly possible. They must be inferred from the subject's responses. Each of several variants upon the basic model make somewhat different assumptions about the nature of these discriminal processes. The standard deviations depicted in Fig. 2-4 are unequal, and so some stimuli are more variable than others. Because this is a discriminant model, the discriminal processes of at least some stimuli must overlap measurably. If the discriminal distribtion of any stimulus does not overlap with any of the others, its interval location cannot be determined. The major assumptions and deductions of the general model are as follows:

1 Denote the covert discriminal responses to stimulus j as $r_{j.}$ and the covert discriminal responses to stimulus k as r_k.

FIGURE 2-4 Discriminal distributions of three stimuli which fall at progressively higher points along the strength axis and are also progressively more variable.



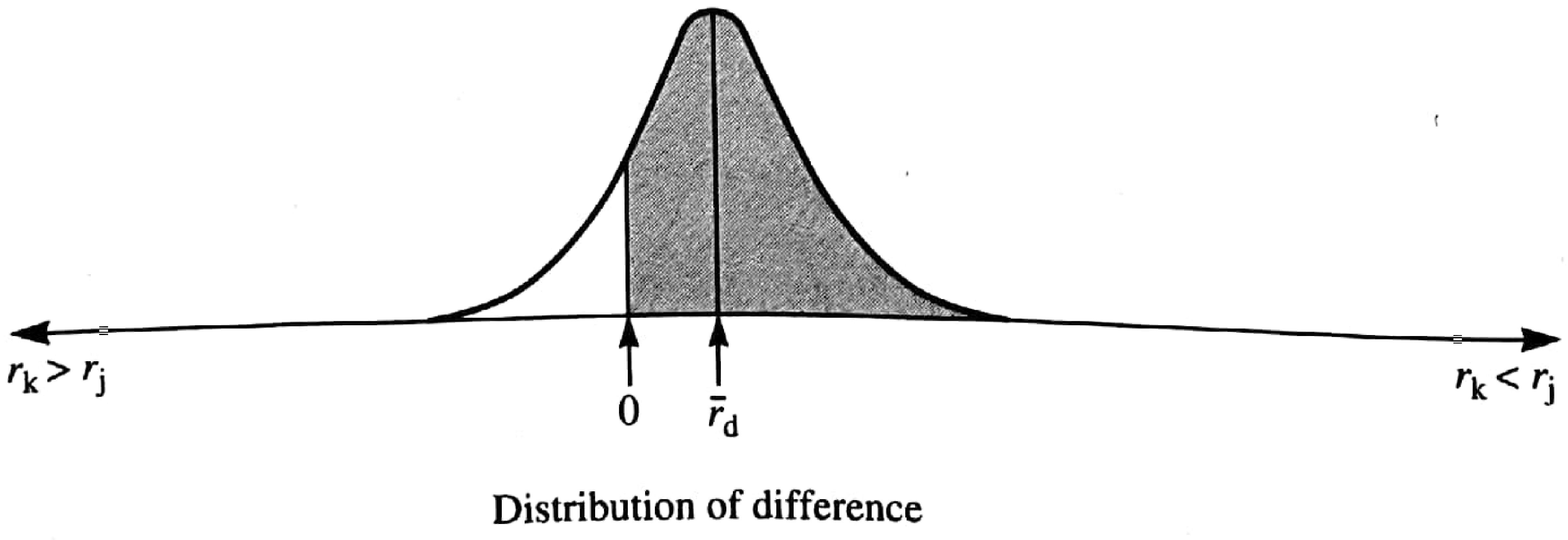
2 The means of these discriminal responses, $\bar{r}_j$ and $\bar{r}_k$, are the best estimates of their respective scale positions. That is, if each stimulus' discriminal processes could be determined directly, its mean (arithmetic average) would be the best estimate of a typical reaction and therefore its location on the interval scale of judgment or sentiment.

3 The overlap in discriminal distributions causes the difference in response to the two stimuli, $r_d = r_j - r_k$, to be positive on some trials and negative on others, producing the varied response to fixed stimuli that is necessary in discriminant models. In the present case, there is variation in the perception of difference. Understanding distributions of difference scores is absolutely crucial to understanding discriminant models used in comparisons. By analogy, two weight lifters each vary in their skill because of a variety of random factors. The varied amounts of weight they lift at a competition produce distributions analogous to those in Fig. 2-2. Heavier weights quite literally mean greater strength. One lifter may be better than the other on average. However, if their abilities are sufficiently similar, their distributions will overlap; the weaker athlete may sometimes lift a heavier weight than the better athlete. One could subtract the weight of the poorer lifter from the weight of the better lifter in any competition to obtain a difference score. Most of these differences will reflect the fact that the poorer lifter cannot lift as heavy a weight as the better lifter. It is perfectly proper to place these difference scores into a frequency distribution which summarize the overlap of the two separate distributions. In this case, the weights can actually be scaled directly, but this is the exception.

4 Because the individual discriminal processes r_j and r_k are assumed to be normally distributed, the distribution of their difference, $r_d = r_j - r_k$, will also be normally distributed. This distribution of differences is illustrated in Fig. 2-5. The shaded area is proportional to the percentage of times stimulus j is judged greater than stimulus k, and vice versa for the unshaded area. Note that the mean ($\bar{r}_d$) is positive. This is because the mean discriminal response to stimulus r_j ($\bar{r}_j$) is greater than the mean discriminal response to r_k ($\bar{r}_k$); consequently, the majority of the differences (the shaded portion) are positive rather than negative.

5 The mean of the differences between responses to the two stimuli on numerous occasions, $\bar{r}_d = \bar{r}_j - \bar{r}_k$, is the best estimate of the interval separating the two. Although this mean cannot be estimated directly because it is entirely hypothetical, Thurstone's

FIGURE 2-5 Distribution of discriminal differences for two stimuli, j and k, where j is ordinarily preferred to k.



law of comparative judgment allows it to be estimated from paired comparisons as follows.

6 Ask a subject to state whether stimulus j is greater or less than stimulus k with respect to an attribute. Denote the proportion of times j is judged greater as $p_{j>k}$.

7 Next, assume that discriminal differences are normally distributed with a mean of $\bar{r}_d$ and a standard deviation of 1.0. The zero point will fall to the left or to the right of the mean depending on which stimulus is more frequently judged greater with respect to the attribute. Convert $p_{j>k}$ into a corresponding number of standard deviation units from a table of the normal distribution. If, for example, j is judged greater than k 92 percent of the time ($p_{j>k} = .92$), the corresponding normal deviate (z_{jk}) is approximately 1.4. This implies that the zero point is 1.4 standard deviations below the mean. More importantly, $\bar{r}_d = \bar{r}_j - \bar{r}_k$ is 1.4 standard deviations above 0, which moves us close to a solution.

8 With $\bar{r}_d = \bar{r}_j - \bar{r}_k$ expressed in standard deviations units, all that needs to be done is to express $\bar{r}_d$ in terms of the actual standard deviation of the dispersion of discriminal differences. This is necessary because the standard deviations of discriminal differences might differ for different pairs of stimuli. In the above analogy to weight lifters, this would happen if some lifters are more consistent than others. If that occurs, two pairs of stimuli separated by the same *mean* distance could be separated by different *scale* distances. Thus even if z_{jl} and z_{jk} are the same, the standard deviations of the discriminal differences might require different intervals.

9 The standard deviation of the dispersion of discriminal differences can be expressed in the same way as the standard deviation of any set of difference scores. The formula is

$$\sigma_d = \sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_j\sigma_k} \tag{2-3}$$

This is the standard deviation of a difference score

where σ_d = the standard deviation of discriminal differences

σ_j and σ_k = the respective standard deviations of discriminal distributions for stimuli *j* and *k*

r_{jk} = the correlation between the discriminal distributions of the two stimuli

The standard deviation of the distribution of discriminal differences thus involves the standard deviations of the two discriminal distributions and the correlation between them. A correlation that differs from zero implies that the sensory noise components of the two discriminal processes are correlated over trials. Note that positive correlations reduce the magnitudes of discriminal dispersions. This is in fact the norm. For example, people vary in how highly they rate all the stimuli on the covert continuum. Thus, if people made absolute responses to the stimuli, one person might like all of the stimuli and rate them highly, and a second person might feel the converse. However, the process of comparison eliminates this difference. This is one reason why comparative judgments are more consistent (reliable) than absolute judgments.

10 The interval separating two stimuli is obtained from the standard deviation of the distribution of discriminal differences using Eqs. 2-4:

$$\bar{r}_d = \bar{r}_j - \bar{r}_k = z_{jk}\sigma_d \tag{2-4a}$$

$$\bar{r}_j - \bar{r}_k = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2 - 2r_{jk}\sigma_j\sigma_k} \tag{2-4b}$$

Equations (2-4) multiply the normal deviate by the standard deviation of the distribution of discriminal differences between the two stimuli. This allows the proper interval to be found on the underlying measurement scale. These equations define the "complete law of comparative judgments." Their use requires knowledge of (1) the proportion of times each stimulus is judged greater than another with respect to an attribute, (2) the standard deviation of discriminal dispersions for the two stimuli, and (3) the correlation between the two discriminal distributions.

Information is rarely obtained about all three of these statistics; consequently, some simplifying assumptions are usually made. These are discussed in Bock and Jones (1968), Guilford (1954), and Torgerson (1958). The two most common assumptions are (1) the correlations between discriminal dispersions are zero (i.e., responses are independent) and (2) the standard deviations of discriminal dispersions are all equal. Equation 2-4 then reduces to

$$\bar{r}_j - \bar{r}_k = z_{jk}\sqrt{\sigma_j^2 + \sigma_k^2} \tag{2-5a}$$

$$\bar{r}_j - \bar{r}_{jk} = z_{jk}\,\sigma\sqrt{2} \tag{2-5b}$$

Since all dispersions (standard deviations) of discriminal processes are assumed to be the same, the term under the radical reduces to $\sqrt{2}$ times any of the standard deviations. Since that term is constant for all pairs of stimuli and since the intervals on an interval scale are unaffected when all scale values are multiplied by a constant, the formula reduces to

$$\bar{r}_j - \bar{r}_k = z_{jk} \tag{2-6}$$

Thus, these assumptions allow the normal deviate representing the proportion of times one stimulus is preferred over another to define the interval separating two stimuli. Equation 2-6 is by far the most frequently used form of the law of comparative judgment. Further simplifying assumptions are made when the law of comparative judgment is actually applied. The most general form of the model is based on response distributions of one subject on numerous occasions. This is seldom done for three reasons. First, it is difficult to find subjects who will devote the time to the task. Second, most responses are not independent—subjects tend to remember their previous responses. Third, the usual goal of scaling stimuli is to obtain a scale that applies to a definable group of people. A scale that applies to only one person is usually of limited generality.

The law of comparative judgment can be applied to any form of ordinal data, such as the method of successive categories, but the method of paired comparisons is the most obvious approach. Consequently, each subject is presented with all possible pairs of stimuli in a set, which usually ranges from 10 to 20. The subjects indicate which member of each pair is preferred (greater) with respect to the attribute in question. The

TABLE 2-2 PROPORTIONS OF SUBJECTS PREFERRING EACH VEGETABLE (COLUMNS) COMPARED TO EACH OF THE OTHER VEGETABLES (ROWS)

	Vegetable								
Vegetable	1	2	3	4	5	6	7	8	9
1. Turnips	.500	.818	.770	.811	.878	.892	.899	.892	.926
2. Cabbage	.182	.500	.601	.723	.743	.736	.811	.845	.858
3. Beets	.230	.399	.500	.561	.736	.676	.845	.797	.818
4. Asparagus	.189	.277	.439	.500	.561	.588	.676	.601	.730
5. Carrots	.122	.257	.264	.439	.500	.493	.574	.709	.764
6. Spinach	.108	.264	.324	.412	.507	.500	.628	.682	.628
7. String beans	.101	.189	.155	.324	.426	.372	.500	.527	.642
8. Peas	.108	.155	.203	.399	.291	.318	.473	.500	.628
9. Corn	.074	.142	.182	.270	.236	.372	.358	.372	.500

Source: Adapted from Guilford (1954) by permission of the author and publisher.

result is a table containing the proportion of persons who prefer one stimulus to another ($p_{j>k}$). Table 2-2 lists typical results from a study of food preferences. Values of .5 are placed in each diagonal position in the table as each stimulus is assumed to be judged greater than itself half of the time. Each value of $p_{j>k}$ is then converted into a normal deviate z_{jk}, presented in Table 2-3.

If it is proper to assume Eq. 2-6, each normal deviate in Table 2-3 is an interval between the two stimuli. However, these normal deviates are likely to be affected by sampling error, which can be reduced as follows. The sum of the normal deviates for each column (stimulus) is obtained and then averaged. However, pairs that are widely

TABLE 2-3 TRANSFORMATIONS OF THE PROPORTIONS IN TABLE 2-1 TO NORMAL DEVIATES (z SCORES)

	Vegetable								
Vegetable	1	2	3	4	5	6	7	8	9
1. Turnips	.000	.908	.739	.882	1.165	1.237	1.276	1.237	1.447
2. Cabbage	–.908	.000	.256	.592	.653	.631	.882	1.015	1.071
3. Beets	–.739	.256	.000	.154	.631	.456	1.015	.831	.908
4. Asparagus	–.882	–.592	–.154	.000	.154	.222	.456	.256	.613
5. Carrots	–1.165	–.653	–.631	.154	.000	–.018	.187	.550	.719
6. Spinach	–1.237	–.631	–.456	–.222	.018	.000	.327	.473	.327
7. String beans	–1.276	–.882	–1.015	–.456	–.187	.327	.000	.068	.364
8. Peas	–1.237	–1.015	–.831	–.256	–.550	–.473	–.068	.000	.327
9. Corn	–1.447	–1.071	–.908	–.613	–.719	–.327	–.364	–.327	.000
Sum	–8.891	–4.192	–3.000	–.073	1.165	1.401	3.711	4.103	5.776
Average	–.988	–.465	–.333	–.008	+.129	+.156	+.412	+.456	+.642
Final scale	.000	.523	.655	.980	1.117	1.144	1.400	1.444	1.630

arbitrarily
put least
preferred at
0 to transfer
the rest

Source: Adapted from Guilford (1954) by permission of the author and publisher.

separated (e.g., $z_{jk} > 2.0$), are eliminated from this averaging process because the assumption that these stimuli overlap is not tenable (the "always or never" part of the Fullerton-Cattell law). The results are normal deviates expressed as deviations from the average stimulus in the set. Finally, the value of the lowest (most negative) stimulus is subtracted from each of the values to eliminate negative values in the final scale. This produces the final interval scale, e.g., of food preferences for the data in Table 2-3. Corn is the most liked vegetable, and turnips are the least liked. The latter is arbitrarily designated as zero on the scale. This zero is arbitrary, by definition, since this is an interval scale.

Simulating Thurstone Scaling

One can work backward from the scale values presented at the bottom of Table 2-3 to estimate the proportions found in Table 2-2 directly by applying Eqs. 2-3 through 2-6 in reverse order. Consequently, a Monte Carlo approach is unnecessary. However, it is instructive to perform one and compare it to our previous simulation. The first step is to multiply the scale values in the bottom line of Table 2-3 by $\sqrt{2}$ to conform to Eq. 2-5b. This provides values of .000, .740, .926, ..., which are the mean discriminal responses—r_j, r_k, ...

To compare turnips with cabbage, two numbers were chosen from a normal distribution having a mean of zero and a standard deviation of 1.0. The first number was added to the value associated with turnips (.000), and the second number was added to the value associated with cabbage (.740). These independent, normally distributed random numbers provided discriminal dispersions. When added to the scale values, they yielded the covert discriminal responses, r_j and r_k, of assumption 1. They were assumed normally distributed because of assumption 4. The subject preferred turnips over cabbage if $r_j - r_k$ was > 0 but preferred cabbage over turnips if $r_j - r_k$ was < 0. This was repeated 1000 times for each stimulus pair. The resulting probabilities appear in Table 2-4.

TABLE 2-4 ESTIMATED PROPORTIONS OF SUBJECTS PREFERRING EACH VEGETABLE BASED UPON COMPUTER SIMULATION

	Vegetable								
Vegetable	1	2	3	4	5	6	7	8	9
1. Turnips	.500	.710	.745	.845	.844	.870	.927	.923	.956
2. Cabbage	.290	.500	.550	.684	.735	.761	.813	.804	.864
3. Beets	.255	.450	.500	.614	.676	.697	.748	.784	.821
4. Asparagus	.155	.316	.386	.500	.525	.567	.649	.668	.745
5. Carrots	.156	.265	.324	.475	.500	.529	.601	.633	.700
6. Spinach	.130	.239	.303	.433	.471	.500	.575	.613	.675
7. String beans	.073	.187	.252	.351	.399	.425	.500	.523	.585
8. Peas	.077	.196	.216	.332	.367	.387	.477	.500	.548
9. Corn	.044	.136	.179	.255	.300	.325	.415	.452	.500

Source: Adapted from Guilford (1954) by permission of the author and publisher.

These probabilities are only a first approximation to those in Table 2-2. For example, turnips are preferred to cabbage .810 of the time, but the simulation only predicted a difference of .710. On the other hand, the observed preference for corn over cabbage (.858) is fairly close to the predicted preference (.864). Consider why the fit was not better. One major factor was that the simulation assumed the equal discriminal dispersions of Eqs. 2-5 and 2-6 instead of the more general Eqs. 2-3 and 2-4. Another possibility is that the stimuli vary along more than one axis, i.e., are multidimensional. Should one try a more general model with more parameters to estimate? Perhaps yes; perhaps no. This is a question of the tradeoff of completeness and goodness of fit against parsimony.

A Comparison of the Two Simulations

Two simulations have been presented in this chapter. The first involved absolute judgments along a single physical dimension, i.e., was psychophysical. The second involved a comparison of two sentiments with stimuli that did not vary along one physical dimension, i.e., was not psychophysical.

The law of comparative judgment has had both historical and continuing importance. The first author had the privilege of sitting in Thurstone's classroom when he indicated that the law of comparative judgment was his proudest achievement. This came from a man for whom the word "genius" is appropriate. Hundreds of journal articles and numerous books have been stimulated by the law of comparative judgment. Although the derivation of the law is not simple, the law itself is held in reverence by some psychometricians, and for good reason.

In the end, the law is very simple. It consists of transforming percentages of "greater than" responses for pairs of stimuli into z scores reflecting their difference. The process uses the inverse of the cumulative normal curve introduced in basic statistics. This inverse function is depicted in Fig. 2-6. The interval between any two stimuli is the z score that corresponds to the percentage of "greater than" responses. Intervals are computed for all pairs of stimuli. Although these z scores themselves can define intervals, they are usually averaged to increase the reliability of the estimates, and the lowest one is set to zero to simplify description.

The point basic to both simulations is that variability due to noise unified the two types of response. The additional factor of a correlation between the separate processes in a comparison is also important in reducing the magnitude of error. The simulations reasonably document what the subjects do.

The Logistic Distribution and Luce's Choice Theory

Although much statistical theory used in scaling employs the familiar normal distribution, more recent work tends to stress the logistic distribution. The ogival shape of the logistic distribution is visually indistinguishable from the cumulative normal distribution, but it is much more convenient mathematically. This will be especially important in Chapters 10 and 15. Equation 2-7 defines the logistic function:

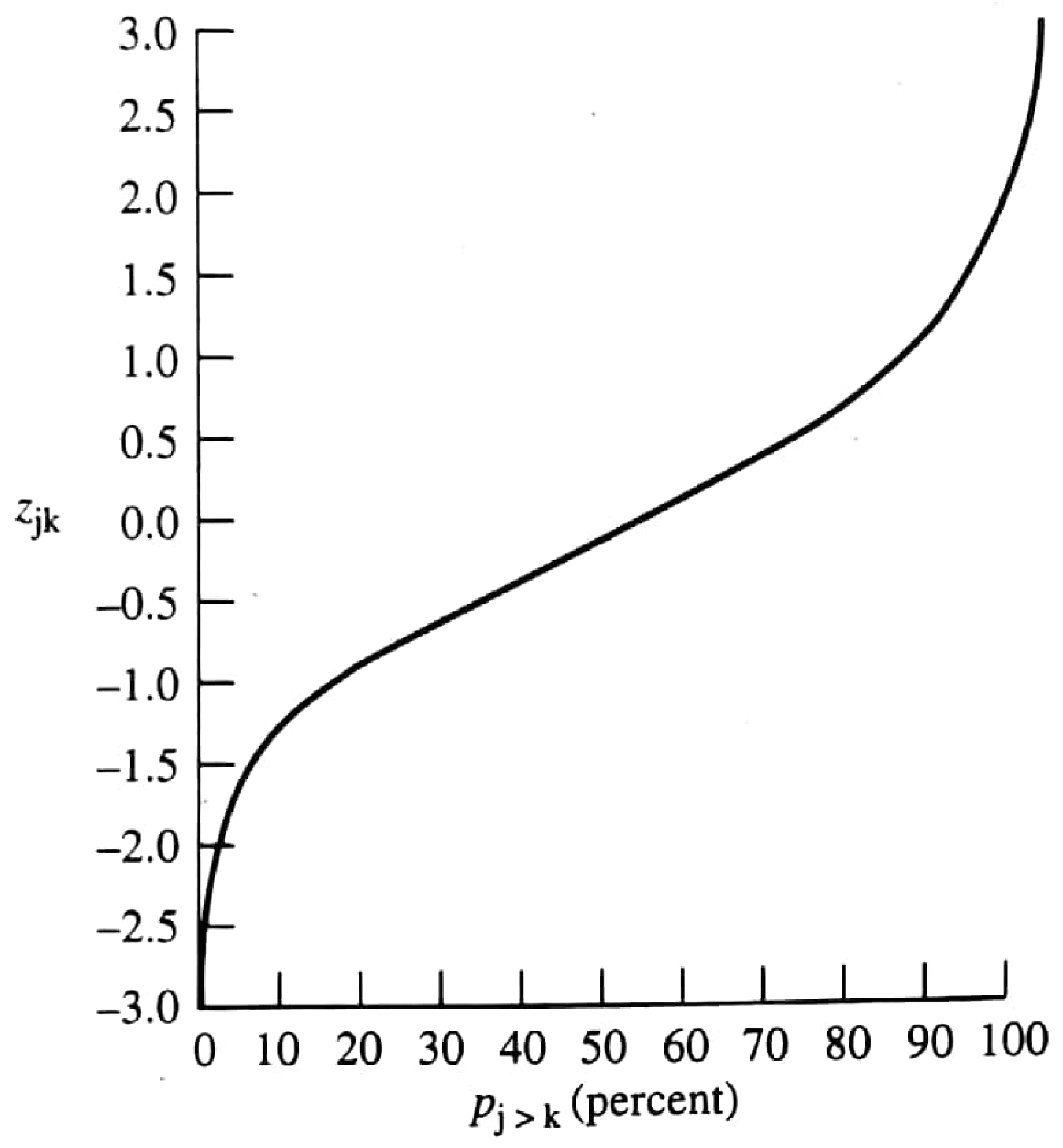


FIGURE 2-6 Interval scale values based upon the law of comparative judgment (z_{jk}) as a funtion of the percentage of "greater than" responses ($p_{j>k}$).

$$Y = \frac{e^{1.7X}}{1 + e^{1.7X}} \tag{2-7}$$

where $e = 2.718281828+$. The constant 1.7 causes Y to differ from the standard cumulative normal distribution by no more than 0.01 for any value of X. Had this distribution been used instead of the cumulative normal, the final scale values would have been indistinguishable.

Just as it is reasonable to use the cumulative normal distribution because the combined effects of independent sources of error tend to form a normal distribution, the logistic distribution may also be justified mathematically through Luce's choice theory (Luce, 1959a, 1963, 1977). Choice theory deals with preferences just as the law of comparative judgment does. Consider choosing one vegetable from menus on which (1) asparagus and beets are the only two choices and (2) there are other options. The probabilities of choosing asparagus and beets are obviously smaller when there are other options since one may prefer corn, cabbage, etc., over either.

The essence of choice theory is the constant ratio rule which predicts that the ratio of choosing asparagus over beets will be the same in both situations. Thus, Table 2-2 indicates that 56.1 percent of subjects chose asparagus over beets when they are the only options. This ratio is 56.1/(100 – 56.1) = 56.1/43.9 or 1.28. Now suppose that 10 percent of subjects choose beets from a larger menu. According to the model, asparagus should be chosen 12.8 percent of the time. These constant ratios in turn are also the ratios of their scale values. The logistic transformation relates scale values (X) to probabilities (Y). In contrast, Eqs. 2-3 through 2-6 show that Thurstone's law of comparative judgment is a constant-difference rule.

Averages as Scale Values

Both Thurstone's law of comparative judgment and choice theory are representational models of interval scaling in the sense of Chapter 1. The reason for choosing either the normal curve transformation that gave rise to Table 2-2 or the logistic transformation follows from the constant-difference and constant-ratio rules (assumptions). Consider what would happen if the scale were simply formed from the preference probabilities in Table 2-2 themselves. One first computes the column sums, which are 1.634, 3.001, 3.438, 4.439, 4.878, 4.947, 5.764, 5.925, and 6.414. Dividing each in turn by 9 to form averages gives 0.181, 0.333, 0.382, 0.493, 0.542, 0.554, 0.640, 0.658, and 0.712. Next, subtracting 0.181 from each average gives values of 0.000, 0.152, 0.201, 0.312, 0.361, 0.373, 0.459, 0.477, and 0.531.

In order to visualize the similarities between these values, based upon simple sums, and either Thurstone's or Luce's formal assumptions, multiply each value by the ratio of the highest scale value in Table 2-3 (1.630) to the highest scale value here (0.531) or 3.07. This makes the first and last values of the two scales the same. Both this and the subtraction of the smallest scale value (0.181) are permissible transformations of an interval scale. The resulting scale values are 0.000, 0.466, 0.615, 0.957, 1.108, 1.145, 1.409, 1.464, and 1.630. The similarities to the proper Thurstone values are apparent and important.

This similarity is one justification for the operationalist position (Gaito, 1980) discussed in Chapter 1. Were the table comprised of outcomes for nine baseball teams, the result would be familiar won-loss percentages. However, the operation and therefore the scale values are meaningless in a representational sense since, unlike the rationale provided by Thurstone and his predecessors, there is none for summing probabilities as opposed to *z* scores. The operationalist position is that it is difficult to see why one operation is meaningless when it gives results nearly identical to those of another that is meaningful.

Checks and Balances

So far in this chapter numerous assumptions have been discussed regarding the use of various models for scaling stimuli. How does one know if the assumptions are correct?

1 We have already noted the importance of internal consistency in developing subjective estimate scales. Similar considerations hold for discriminant models. Basically, an ordinal scale is developed by averaging individual subjects' rankings, and the data are internally consistent to the extent that different subjects give similar rankings. As previously noted, suitable methods for obtaining internal consistency measures are discussed later in Chapters 6 through 8.

2 As indicated in the simulations, one can work backward from Thurstone scale values to paired comparison probabilities. These estimated probabilities should be similar to the observed probabilities.

3 One should examine the transitivity of the response probabilities. If stimulus i is preferred to j and j is preferred to k, then i should be preferred to k. Violations of transitivity are an indication that the scale is not unidimensional. Of course, slight violations of transitivity may simply reflect measurement error.

4 A stronger (interval) criterion for unidimensionality is the additivity of the scale values, an issue we will consider in detail in Chap. 14. Tests of additivity depend upon which form of the model is being used, but the basic idea is simple. Suppose that Eq. 2-6, which assumes that the stimuli have equal variances, is used. Now suppose stimulus i is preferred to stimulus j 65 percent of the time ($p_{i>j} = .65$). A table of the normal curve indicates that stimuli i and j are separated by .39 units. Further, suppose that stimulus j is preferred to stimulus k 70 percent of the time ($p_{j>k} = 0$). This implies that stimuli j and k are separated by .52 *z*-score units. Additivity holds to the extent that the distance between i and k is close to .91 (.39 + .52) *z*-score units. Consequently, $p_{i>k}$ should be .82, the value of *p* associated with a *z* score of .91, within measurement error. A failure of additivity could imply that the data are multidimensional, but it could also imply correlated error or unequal variance. Guilford (1954) describes a formal chi-square test of significance.

5 As in any scientific endeavor, relative scale values should be replicable within the linear transformations permitted by an interval scale. As with any other criterion, the degree of replicability is a function of the sample size: The larger the sample, the more stable the expected results. However, other factors, particularly the care with which the data are gathered, are also important. Thus, the relative sizes of the intervals among the stimuli should remain much the same. If the relative sizes of these intervals change markedly across situations, scalings would be highly context-dependent. These findings would therefore ordinarily not be useful unless the changes occurred in a theoretically interesting way.

Multi-item Measures

This book stresses the need for multi-item measures, where "item" is broadly used to stand for any stimuli used in measurement. Thus items may be words on a spelling test, comparisons between weights, statements concerning attitudes toward the U.S. Congress, response latencies, etc. There are a number of important reasons for combining several items when measuring a psychological attribute.

1 Individual items usually correlate poorly with the particular attribute in question.

2 Each item tends to relate to attributes other than the one to be measured. For example, the ability of children to spell "umpire" correctly may partly depend on their interest in baseball.

3 Each item has a degree of specificity in the sense of not correlating with any general attribute or factor. The concept of specificity for individual test items will become clearer when factor analysis is discussed in Chapters 11 through 13.

4 Individual items have considerable random measurement error, i.e., are unreliable. This can be seen when people rerate stimuli. A person who initially rates stimulus A as 3 on one occasion may rerate it as 5. Some of this may reflect changes in the attribute over time, but it may occur even when one has every reason to believe the trait itself is stable. To the extent that some stimuli are rated higher and others are rated lower, measurement error averages out when individual scores are summed to obtain a total score.

5 An item can categorize people into only a relatively small number of groups. Specifically, a dichotomously scored item (one scored pass versus fail) can distinguish between only two levels of the attribute. Most measurement problems require much finer differentiations.

All of these difficulties are diminished by the use of multi-item measures. The tendency of items to relate to incidental factors usually averages out when they are combined because these different incidental factors apply to the various items. Combining items allows one to make finer distinctions among people. For reasons which will be discussed in Chapters 6 and 7, reliability increases (measurement error decreases) as the number of items increases. Thus, nearly all measures of psychological attributes are multi-item measures. This is true both for measures used to study individual differences and for measures used in experiments. The problem of scaling people with respect to attributes is then one of combining item responses to obtain one score (measurement) for each person.

Item Trace Lines (Item Characteristics Curves)

I.V. - general construct
D.V. - response to individual item

Nearly all models for scaling people can be described by different types of curves relating the attribute they measure to the probability of responding one way versus another. Functions of this form are called "item trace lines" or "item characteristic curves" (ICCs). For example, a trace line might denote the probability of recognizing Thurstone as the author of the law of comparative judgment as a function of overall knowledge of psychology. We will define response alpha as passing rather than failing an ability item scored as correct versus incorrect, answering an personality test item in the keyed direction, agreeing rather than disagreeing with an opinion statement, or remembering versus not remembering an item on a list. Response beta is the alternative outcome. More complex models can handle multicategory responses such as Likert scales and the nominal categories of a multiple-choice item. Figure 2-7 depicts four of the possible forms a trace line based upon dichotomously scored items may take: (*a*) a step function, (*b*) an ogive, (*c*) an irregular but monotonic function, and (*d*) a nonmonotonic function.

The point to note about all trace lines is their similarity to local psychometric functions like Figs. 2-1 and 2-2. The difference is that the abscissa of a psychometric function is a physical dimension (Φ) that can usually be described in ostensive terms. The abscissa of a trace line denotes an abstract attribute defined in terms of its strength as in Thurstone scaling and is commonly denoted "Θ". Different models make different assumptions about trace lines. Some are very specific as to form and require trace lines like those in Fig. 2-7*a* and 2-7*b*; others describe only a general form like that in Fig. 2-7*c*. Figure 2-7*d* generally represents what is normally an undesirable outcome. It is most likely to arise when a distractor on a multiple-choice test tends to be chosen by high-ability subjects, perhaps because it is correct in a way that the test constructor did not think of. However, there are some models that use nonmonotone items.

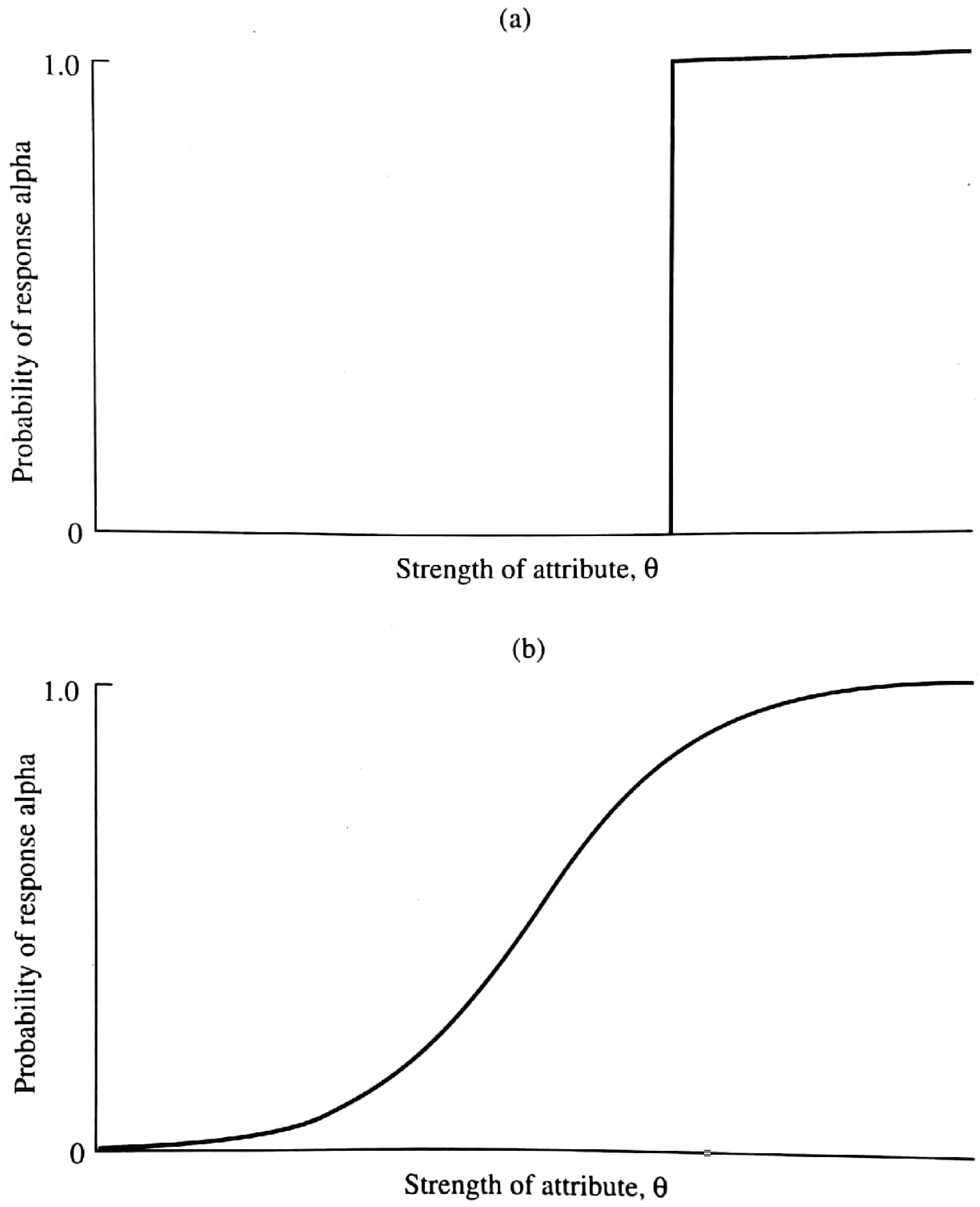


FIGURE 2-7 Trace lines (item-characteristic curves). (*a*) A step function, (*b*) an ogive.

In general, it is important to distinguish among (1) a single observation (test item), (2) a more general attribute measured by a finite number of items that may be spuriously influenced, an obtained or fallible score, and (3) a hypothetical, perfectly measured attribute or true score perhaps as measured on an infinite number of trials. A critical difference between the classical approach of Chapters 6 and 7 and the modern approaches of Chapter 10 (item response theories) is that classical approaches usually define Θ in terms of obtained measures (fallible scores), but item response theories always define Θ in terms of true scores. The ordinate in both cases is the probability or proportion of response alpha, and thus refers to a test item.

Attributes are also commonly called "constructs" or, in the narrower sense of personality theory, "traits." When an attribute is inherently categorical (e.g., political party or religious membership), the attribute is called a "class." Classes may vary complexly, but attributes are otherwise generally assumed to vary in only one way,

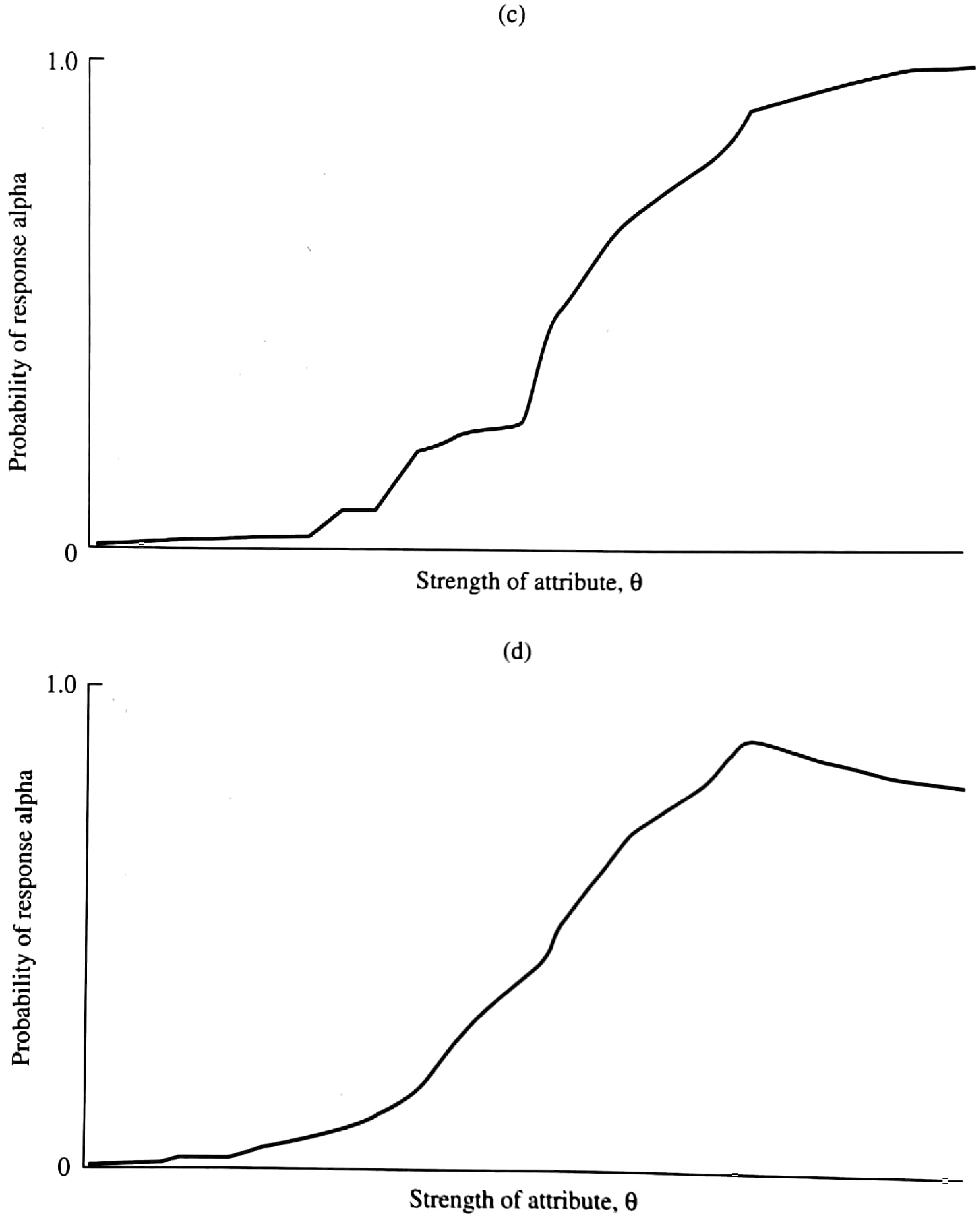


FIGURE 2-7 (*c*) A monotonic function with no well-defined form, and (d) a nonmonotonic function.

i.e., be unidimensional. The measurement of constructs is discussed in Chapter 3, and the principles involved in estimating fallible and true scores are considered in Chapters 6 through 10. The response of a subject in recalling a particular word from a list of words presumably relates to a more general attribute of memory. The principles apply to measuring any response, not just pencil-and-paper items.

An attribute is defined somewhat circularly in terms of whatever the items tend to measure in common. Chapter 3 considers the process of validation that is used to "break the circle." Appropriate methods also exist to infer how much the items have in common. Thus a list of spelling words are assumed to measure spelling ability, and the number of words correctly recalled are a (fallible) measure of memory for the particular material. The word "tend" indicates that no attribute is perfectly mirrored in any finite set of items. Perfectly reliable measurement demands that children be administered all words in the English language on a spelling test or subjects in a memory study be given an infinitely long list to recall.

Difficulty and Discrimination

Two basic properties of a trace line are its difficulty and its discrimination. "Difficulty" refers to how much of the attribute an individual must possess to achieve a given probability of response alpha. Increasing the difficulty of an item is equivalent to "sliding" its trace line to the right, as has been done with the item denoted *a* in Fig. 2-8. This might occur when a group of primary school children are asked to spell "cattle" instead of "cat." Making the item easier slides the trace line to the left. The classical psychometric index of difficulty is simply the probability of response alpha. However, modern theories use the amount of an attribute (Θ) necessary to achieve a .5 probability of response alpha, i.e., the "threshold." This reflects the analogy to psychophysics.

The "discrimination" of an item describes the extent to which the probability of response alpha correlates with the attribute. An item with a perfectly flat trace line does not discriminate and should be eliminated from the test. Most models are called "monotone" models in that the probability of response alpha is expected to increase with the attribute in the general form of Fig. 2-7*c*. In that case, making an item more discriminating increases its slope, as depicted by the item designated b in Fig. 2-8, which completes the analogy with the psychometric function of psychophysics. The most common classical index of discrimination is the correlation over people between response alpha and total test score, the "item-total" correlation. The concepts of difficulty and discrimination are logically independent, as an item may be difficult or easy regardless of whether it is discriminating or nondiscriminating. However, modern test theorists stress that the probability of response alpha and the item-total correlation are not independent because, as will be noted in Chapter 4, the proportion of alpha responses places limits on the item-total correlation. In fact, item response theories use the slope of the trace line to describe discrimination, just as in psychophysics.

There must be a large number of persons at each point on the trace line. Save for classes, attributes are continuous, so that it is theoretically possible to make infinitely fine discriminations. The trace line thus shows the expected response probability for people at that level of the attribute or class. This expectation either defines the probability of response alpha for dichotomous items or the mean for Likert or other multicategory items. Such expectations inherently contain error. For example, there is a probability of response alpha at each point for dichotomous items, but there is no certainty as to who will respond alpha and who will respond beta. Multicategory items likewise have a band of error (standard error) surrounding the average. Thus, although the expected score on a 5-point Likert scale for a given point on an attribute might be 3.1, scores at that point probably range from 1 to 5.

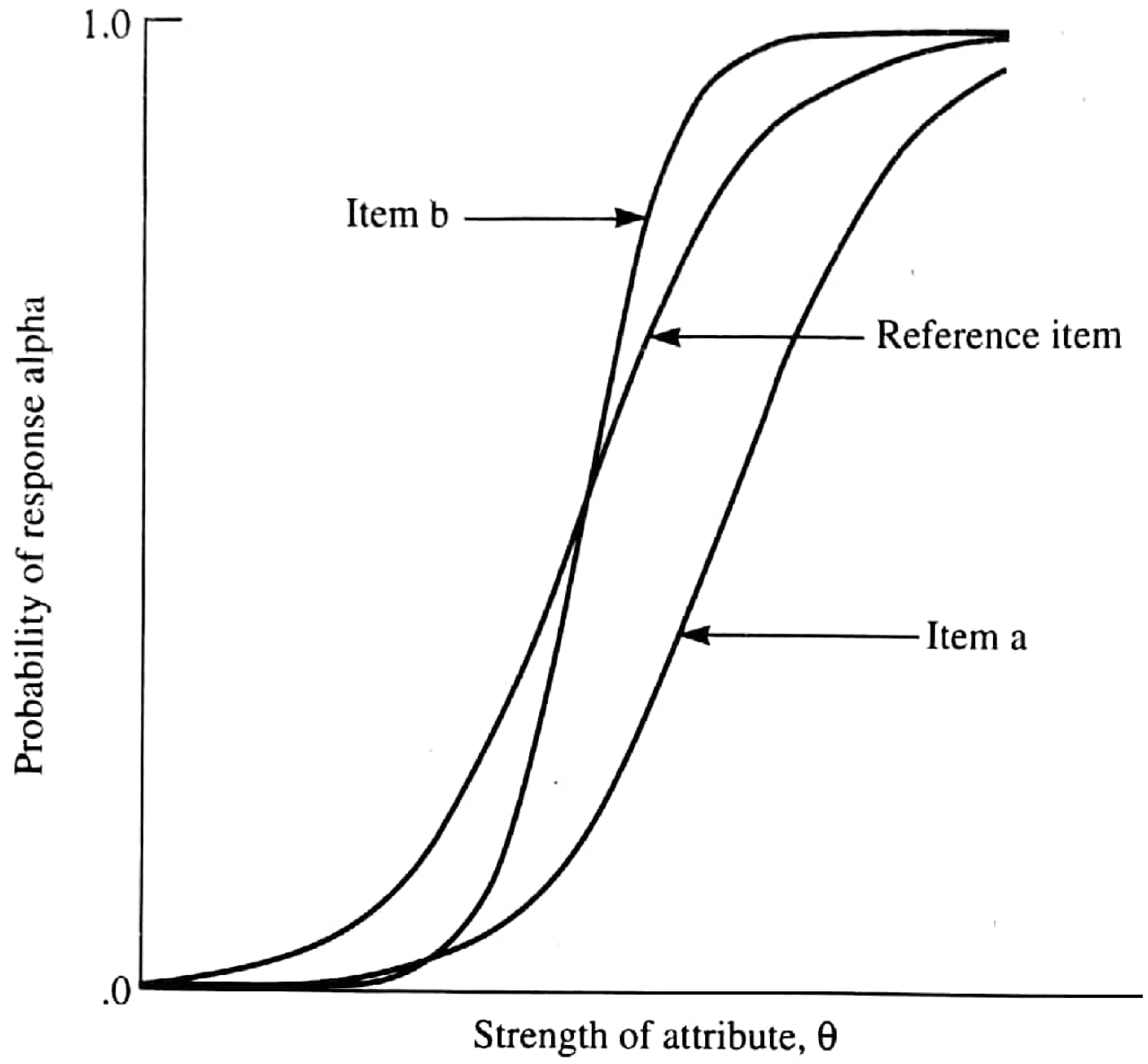


FIGURE 2-8 Effects of making an item more difficult (item a) or more discriminating (item b) relative to a reference item.

DETERMINISTIC MODELS FOR SCALING PEOPLE

Deterministic models are so called because they assume that there is *no error* and so the trace line is a step function as in Fig. 2-7*a* (or Fig. 2-1*a*). The most common form assumes that the probability of response alpha to a dichotomous item at each level of the attribute is 0 up to a point (probability of response beta is 1.0), i.e., the threshold. Beyond this point the probability of response alpha is 1.0. Its discrimination is therefore infinite at the threshold. Figure 2-9 contains a family of such items. Each item has a perfect biserial correlation (an estimated correlation between a dichotomous measure and a continuous measure, assuming that both underlying measures are continuous and normally distributed, see Chapter 4) with the attribute. Consequently each item perfectly discriminates at a particular point of the attribute. This is perhaps a very appealing model because it is exactly what one expects to obtain from measurements of length. Thus, one expects to obtain a trace lines like those in Fig. 2-9 for the following items:

	Yes	No
(a) Are you above 6 feet 6 inches in height?	____	____
(b) Are you above 6 feet 3 inches in height?	____	____
(c) Are you above 6 feet in height?	____	____
(d) Are you above 5 feet 9 inches in height?	____	____
(e) Are you above 5 feet 6 inches in height?	____	____

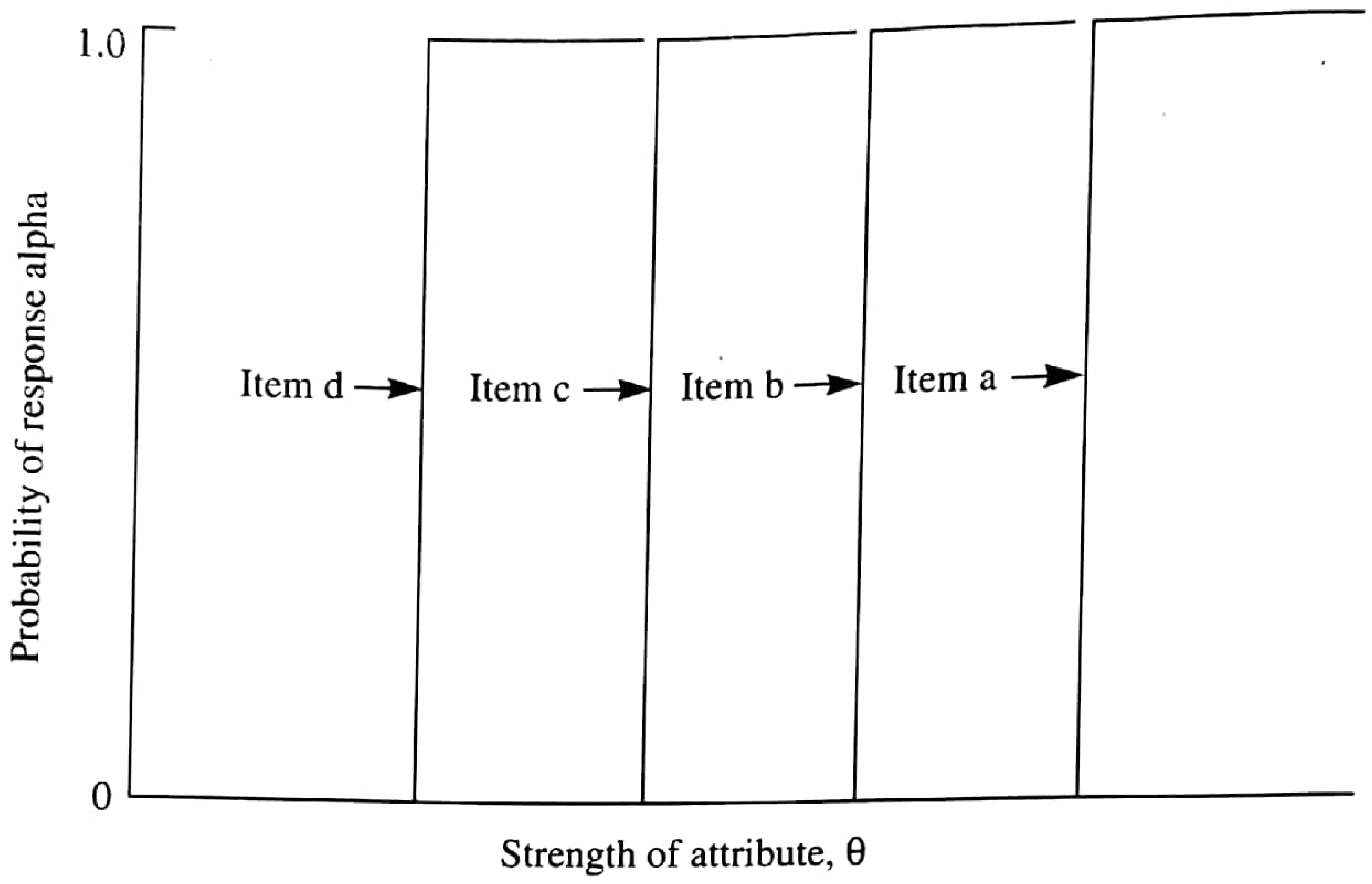


FIGURE 2-9 A family of trace lines that discriminate perfectly at different points and thereby form a monotonic deterministic model (Guttman) scale. Items a to d are progressively easier.

Assume "yes" is response alpha. Any person who answered yes to question (a) would answer yes to the others. Any person who answered *no* to (a) but answered yes to (b) would also answer yes to questions (c) through (e). Five people with different patterns of responses would produce a triangular pattern of data like that in Table 2-5. An X symbolizes a yes answer (response alpha).

The Guttman Scale

Although a trace line usually requires at least some statistical estimation, one can look at data to see if they provide a triangular pattern like that in Table 2-5 (making, however, a subtle logical assumption as discussed below). Some items do produce a pattern of data like that in Table 2-5, perhaps the following:

	Yes	No
(a) The U.S. Congress is the savior of all Americans.	___	___
(b) The U.S. Congress is America's best hope for peace.	___	___
(c) The U.S. Congress is a constructive force in the American political system.	___	___
(d) We should continue our present system of government, including Congress.	___	___

Anyone who answers yes to (a) will probably answer yes to the other items; anyone who answers no to (a) but answers yes to (b) will probably answer yes to the other items; etc. Items that produce a pattern of responses like those in Table ~~2-6~~ form a

2-5

TABLE 2-5 TRIANGULAR PATTERN OF RESPONSES FITTING A GUTTMAN SCALE

	Person				
Item	1	2	3	4	5
a	X				
b	X	X			
c	X	X	X		
d	X	X	X	X	
e	X	X	X	X	X

"Guttman scale." Guttman scales are developed by administering items to a group and then attempting to arrange the responses so that they form the required triangular pattern (see Torgerson, 1958). The data will form a "solid staircase" of alpha responses, and the height of each step will be proportional to the number of people at each level of the attribute. The term "scalogram analysis" describes methods of developing Guttman scales.

Unfortunately, it is very unlikely that the initial set of items will produce a triangular pattern. It is therefore necessary to (1) discard some items and (2) find the best possible ordering among the remaining items. The reproducibility of score patterns is of primary concern regarding the latter issue. If a triangular pattern is obtained, knowing the number of alpha responses allows one to reproduce *all* of an individual's responses. The percentage of people whose patterns are thus reproduced is a basic statistic in scalogram analysis.

Guttman scales could conceivably be developed for any type of dichotomous item such as a spelling test. A triangular pattern of data will be obtained (X denoting a correct spelling) if the items have trace lines like those in Fig. 2-10. If person A has a score of 35 and person B has a score of 34, person A would have to get the same 34 items correct as person B plus the next most difficult item. Knowing how many items an individual passes defines which items are passed.

Figure 2-10 describes a variant upon the Guttman scale which uses nonmonotone items instead of monotone items, i.e., the trace lines go up and then comes down. Our discussion of Guttman scaling in the next paragraph applies to this variant. Subjects falling between two levels of the attribute respond with alpha, and subjects who either fall below the first level or above the second level respond with beta. Each person responds with alpha to only one item. The following four items should fit this model:

	Yes	No
(a) Are you between 6 feet 3 inches tall and 6 feet 6 inches?	____	____
(b) Are you between 6 feet tall and 6 feet 3 inches?	____	____
(c) Are you between 5 feet 9 inches tall and 6 feet?	____	____
(d) Are you between 5 feet 6 inches tall and 5 feet 9 inches?	____	____

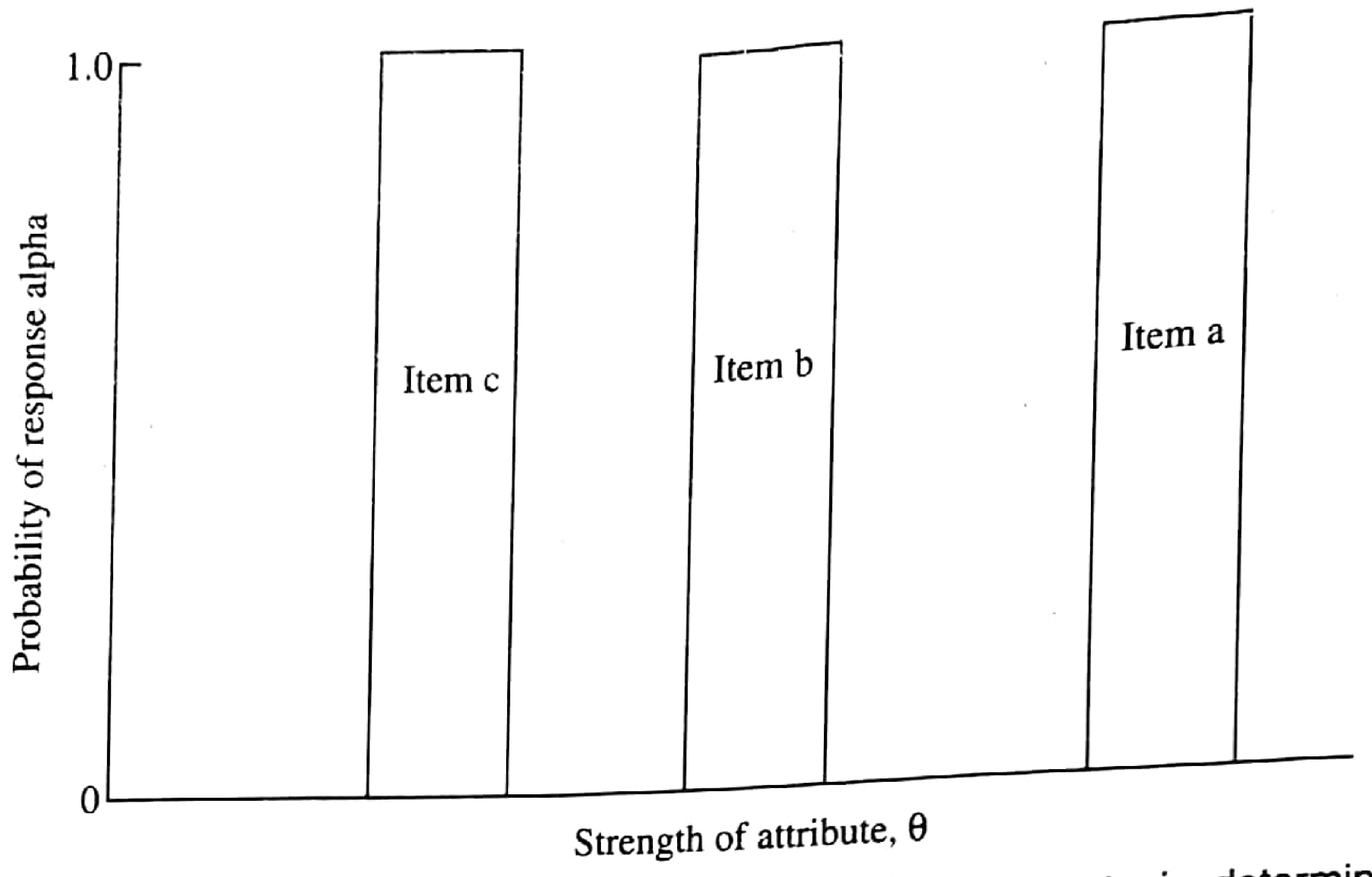


FIGURE 2-10 A family of trace lines that meet the requirements of a nonmonotonic, deterministic scaling model. Items a to c are progressively easier.

Evaluation of the Guttman Scale

The Guttman scale concept has great intuitive appeal, but it is highly unrealistic. First, as rare as step functions are in psychophysics, where there is great control over the stimuli, anything approaching a step function outside that context is even rarer. No item correlates perfectly with any attribute. Although there is no way to obtain the trace line directly, some good approximations are available. Trace lines obtained with virtually all items have a much flatter slope than is consistent with the Guttman model, regardless of whether classical or modern methods are used to estimate them. Individual items rarely correlate higher than .60 with total scores. That is why it is unreasonable to assume a model that assumes perfect biserial correlations between items and an attribute.

Second, having a triangular pattern of data does not guarantee that items have step-function trace lines like those in Fig. 2-11. Items whose thresholds are far enough apart in difficulty will provide a triangular pattern even if their trace lines are fairly flat. This may be illustrated with the following four items:

a Solve for x: $x^2 + 2x + 9 = 16$.
b What does the word "severe" mean?
c How much is 10×38?
d When do you use an umbrella? (given orally).

We have not performed the experiment but suspect that the above four items probably would form an excellent Guttman scale if they were administered to subjects ranging in age from 4 to 16. Anyone who got the first item correct probably could get the others correct. Anyone who failed the first item but got the second correct would probably get the other two correct, etc. This would produce the required triangular pattern of data even though they probably measure different attributes ("factors," in the sense of Chapters 11 through 13). They apparently fit the unidimensional scale model because they are administered to an extremely diverse population. Consequently, it does not follow that having a triangular pattern of data is *sufficient* to establish a unidimensional scale. Because triangular data patterns can be obtained any time items vary greatly in difficulty, Guttman scales seldom have more than eight items. To take an extreme case, three items that are, respectively, passed by 10, 50, and 90 percent of the subjects will probably produce a triangular pattern regardless of their content. Scales which have eight or fewer items can make only gross discriminations among people.

A third criticism of the original Guttman scale was that it provided only an ordinal scale. However, recent methods of statistical estimation considered in Chapter 10 allow Θ to be estimated on an interval scale.

A fourth criticism of the Guttman scale is that it is usually more appropriate to think of items as rubber yardsticks applied by investigators with limited vision rather than as well-defined and well-understood procedures. To complete the analogy, one should think of items as rubber yardsticks that are poor copies of a real yardstick so that some yardsticks may have a zero point at 4 inches. Any single yardstick (item) discriminates poorly. However, the methods discussed in subsequent sections, such as simply adding items evoking response alpha, allow one to combine these various rubber measurements to obtain an approximate linear relationship with "better" yardsticks and thus obtain an interval scale.

In summary, we suggest the deterministic model underlying the Guttman scale is not very applicable to psychological measurement because (1) almost no items fit the model, (2) a triangular pattern is a necessary but not sufficient condition for the fit of the model, (3) the triangular pattern can be (and usually is) an artifact of using a small number of items that vary greatly in difficulty, (4) the model originally provided only an ordinal scale (a problem since overcome), and (5) there are better ways to develop measurement models [Cliff (1983a) presents a well-reasoned defense of Guttman scaling; it is also important to distinguish between Cliff's work in which the Guttman model is applied to a dichotomized composite score and the present discussion, in which individual items are presumed to fit a Guttman scale]. However, impractical models are often very important to the development of more useful models. This is certainly the case with the Guttman scale—the item response theories of Chapter 10 replaced the assumption of a step function with a more realistic ogive of the form presented in Fig. 2-7*b*. The Guttman scale, while unreasonable in itself, is a basic link to modern test theory.

PROBABILISTIC MODELS FOR SCALING PEOPLE

Trace lines that are not step functions like Fig. 2-7*a* describe some probabilistic models. There are numerous types of probabilistic models, depending on what form the trace line is assumed to have.

Nonmonotone Models

Nonmonotone probabilistic models are analogous to nonmonotone deterministic models as discussed above. Trace lines that change slope from positive to negative, or vice versa, at some point are nonmonotone. The only nonmonotone model that has been used assumes trace lines that are in the shape of normal distributions, as depicted in Fig. 2-11 for three items.

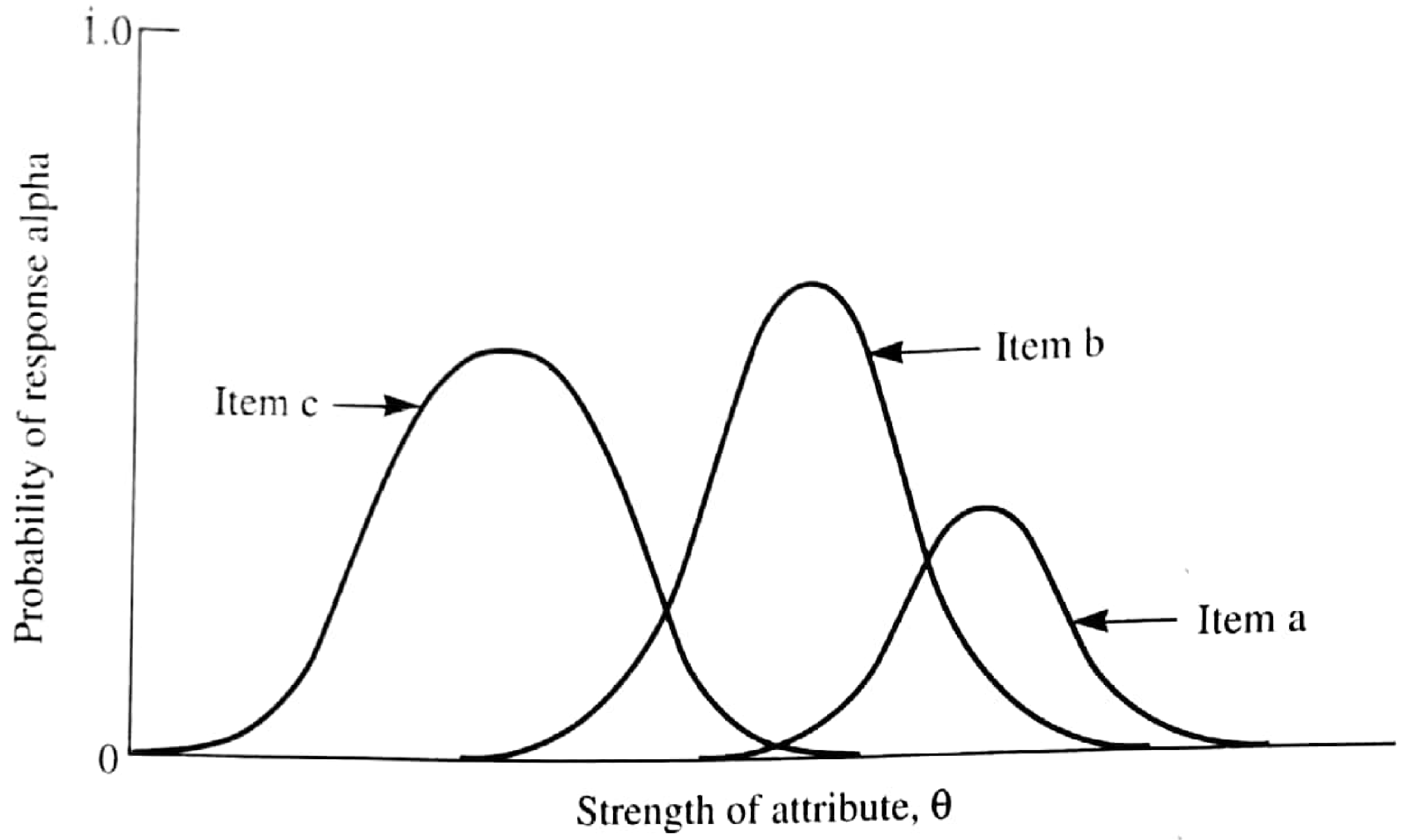


FIGURE 2-11 Nonmonotone, normal trace lines for three items that might be used on a Thurstone scale of attitudes.

The trace lines need not be exactly normal, and their standard deviations need not be equal. This model has been used only to develop a few attitude scales. Since Thurstone developed the scaling procedure, it is referred to as a "Thurstone scale of attitudes." However, this model has little to do with and should not be confused with Thurstone's previously discussed law of comparative judgment. Items at three points on a Thurstone scale of attitudes are as follows:

	Agree	Disagree
(a) I believe that the church is the greatest institution in America today.	___	___
(b) I enjoy a fine ritual service with good music when I go to church.	___	___
(c) The paternal and benevolent attitude of the church is quite distasteful to me.	___	___

A Thurstone scale of attitudes begins with a large pool of attitudinal statements rated by 100 or more raters. Each statement is rated on an scale consisting of about 11 Likert-type steps, perhaps ranging from "strongly favorable" with respect to the attribute to "strongly unfavorable." Note that the raters do not state how *they* feel about the item; they evaluate the *item* itself, so that both a conservative and a liberal rater might rate a given item as "moderately favorable" with respect to a liberal position. Two standards are used to select a set of 10 to 20 items from the initial pool: (1) The ratings of items should have small standard deviations over raters (i.e., the raters should agree among themselves where the items fall on the scale); and (2) means for different items should vary considerably (i.e., items should reflect a wide range of the attribute). A subject's score is the average score of the items he or she endorses.

For example, if a subject agrees with items that have scale scores of 3.0, 3.1, and 3.2 and disagrees with all of the remaining items, that subject is assigned a score of 3.1. Another approach is to assign the scale score of the *highest* item on the scale with which the person agrees. The consistency of judgments can be inferred from the standard deviation of the endorsed items scale scores.

The Thurstone scale of attitudes model states that each item should evoke response alpha (agreement in this case) in only one limited region of the attribute (Θ). Assuming that the trace line has an approximately normal distribution recognizes that items may be endorsed by people with a range of attitudes and, conversely, that people with a given attitude endorse items near and not necessarily at their preferred position. If only people who fall at 3.1 on the scale were to endorse an item judged to be 3.1, the scale would have to have an infinite number of items to capture the one that epitomizes the subject's attitude.

The major fault of the Thurstone scale of attitudes and, for that matter, any other nonmonotone model, is that good nonmonotone items are very difficult to construct. This is especially true for abilities items and, more generally, judgments. The problem is somewhat less severe with sentiments—a person who likes chocolate ice cream may not want it at every possible occasion. However, the model also has logical difficulties with attitudinal statements and sentiments in general. Items fitting this model tend to be "double-barreled" in saying one good thing and one bad thing. This can be seen in the three attitude statements given earlier. Item (*b*) asks subjects to agree simultaneously with two hidden statements:

(b_1) I sometimes go to church.
(b_2) I probably would not go to church if it were not for the fine ritual services and good music.

Likewise, item *(c)* is "triple-barreled" because a subject must agree that the church is paternal, benevolent, and distasteful to agree with it. The three modifiers collectively imply a moderately negative attitude toward the church. One constructs such items only by building two or more statements into what is ostensibly one statement. People who are not skilled at constructing questionnaires often unintentionally construct such ambiguous statements. Some subjects respond to one of the hidden statements, some subjects to another. This is ordinarily not useful in defining a relevant trait. One might as well construct statements like the following: The church is a wonderful, horrible institution.

Another important criticism of nonmonotone probability models is that it is very difficult to think of suitable items to define the ends of the scale. This is illustrated with item (*a*) in the previous example. Who could have a very positive attitude toward the church yet disagree with the statement, "I believe the church is the greatest institution in America today"? Such items will be monotone, continuing to increase in probability of endorsement as the level of the attribute increases.

In summary, nonmonotone probability models at best have limited applicability to the measurement of attitudes. One is probably better off restating the items in a monotone form and using an appropriate model for such items. We now turn to such models.

Monotone Models with Specified Distribution Forms

Some monotone trace line models assume that the trace lines fit a particular statistical function. In particular, those that form the basis of modern psychometrics assume ogives like Fig. 2-7*b*. Another distinguishing characteristic of most of these models is that the pattern of responses defines the scale score rather than simply the number answered in the alpha direction, e.g., correctly. The ideas that are basic to these models have been available for a long time, but computers and recent developments in numerical estimation have spurred their recent growth.

Ogival trace lines are always more discriminating in their steeply ascending middle part than at the extremes. The steeper that section of the trace line, the higher the item-total correlation and other discrimination statistics. If it were a step function, the item would correlate perfectly with the attribute and form part of a Guttman scale. As items correlate less and less with the attribute and therefore become less discriminating, the ogival S shape flattens toward the horizontal.

Ogival models are appealing for two reasons: First, they make good intuitive sense. One can easily think of a critical interval of uncertainty as in psychophysics (see Fig. 2-2*b*) where subjects respond in both directions. This interval of uncertainty is more realistic than the perfect discrimination in a Guttman scale. Moving further away from that zone in either direction markedly reduces the uncertainty. Persons below that zone will choose response beta almost exclusively, and persons above it will choose response alpha almost exclusively. Thus, people of low ability will find a particular item too difficult, and people of high ability will find the same item too easy.

Another reason for the appeal of this model is that it has useful mathematical properties. For example, the sum of a series of ogives is also an ogive of predictable location and slope. The scale score is usually obtained from the probabilities of individual responses. This may require a complex algorithm, but it is a linear function of the attribute under certain assumptions. In contrast, scores derived from item sums typically are not linearly related to the attribute (Lord, 1980, also see Chapter 10), even though this nonlinearity is rarely a major problem. Another useful deduction is that the most discriminating items at any point on the attribute are those whose sum is as steep as possible at that point. This permits some interesting deductions about discriminating at a point, item difficulties, and correlations of items with total scores. It is also possible to deduce the amount of measurement error (unreliability) at different points on the attribute.

Some models make additional assumptions involving correlations among the items or the distribution of the underlying attribute. These assumptions provide a variety of interesting deductions that are useful when the assumptions hold. In particular, one can deduce the score that individuals would make on a test that they have not taken from the score that they made on one that they actually had taken even if the two tests were unequally difficult. These models have been a major, if not the major, focus of psychometric research (see Hulin, Drasgow, & Parsons, 1983; Lord, 1952a; 1974, 1980; Lord & Novick, 1968; Thissen, & Steinberg, 1988; Thissen, Steinberg, & Gerrard, 1986; Wainer & Braun, 1988). At the same time, item response approaches have not supplanted the conventional approach of summing item scores when subjects answer all test items on a test that is administered once.

Monotone Models with Unspecified Distribution Forms

We finally arrive at the model that underlies most scaling—the linear, summative or centroid model. The model makes three major assumptions:

* **1** Each item has a monotonic trace line as in Fig. 2-7*c;* the form of this monotonic trace line can even vary over items.

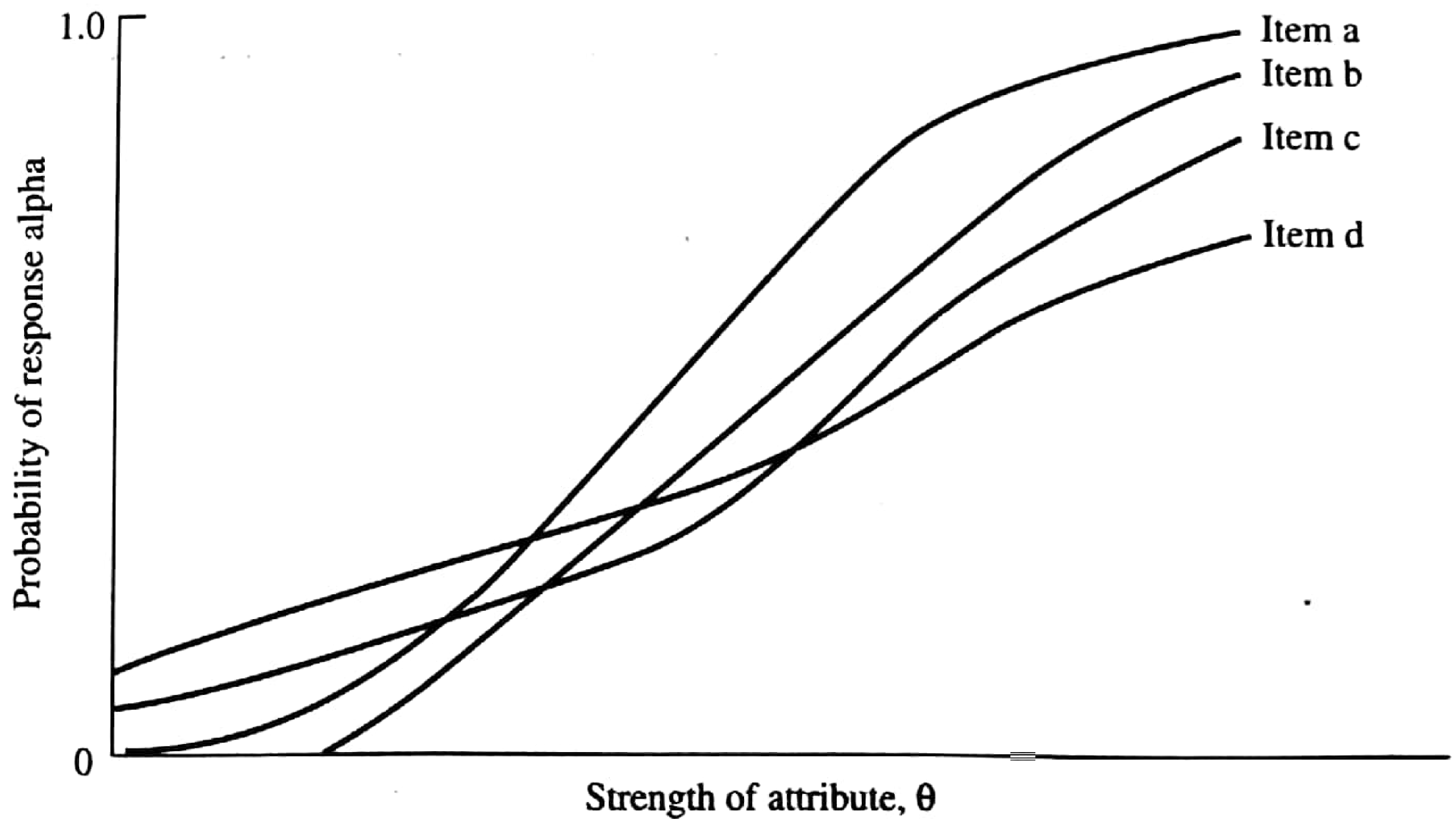
2 The sum of the trace lines for a particular set of items (the trace line for total test scores) is approximately linear. That is, even if items do not all have the same type of monotonic trace line, departures from linearity average out when items are combined.

3 The items as a whole measure only the attribute in question. This is the same as saying that the items have only one factor in common, a point to be discussed in detail in later chapters. It implies that the total score summarizes all the important information about the attribute being measured.

Figure 2-12 contains a family of such trace lines, and Fig. 2-13 presents the sum of these trace lines, the trace line of expected scores on a four-item test.

The model is called "linear" because the score is derived from a linear combination which is a sum of item responses. Even though the underlying mathematics of the linear model is not as elegant as that of modern psychometric models, it is not devoid of such properties given its use of the algebra of linear combinations. This sum does not require each item to have equal weight. The term "centroid" means average—the total score divided by the number of items gives the centroid or average. This is equivalent to weighting each of the *K* items used to generate the score by 1/*K*. Thus, a person's score on a classroom examination would probably be presented as the equally weighted sum, but performance on a series of reaction time trials might presented as a mean—the choice is a matter of convenience.

FIGURE 2-12 A family of four items with monotone trace lines that can be used in a linear model.



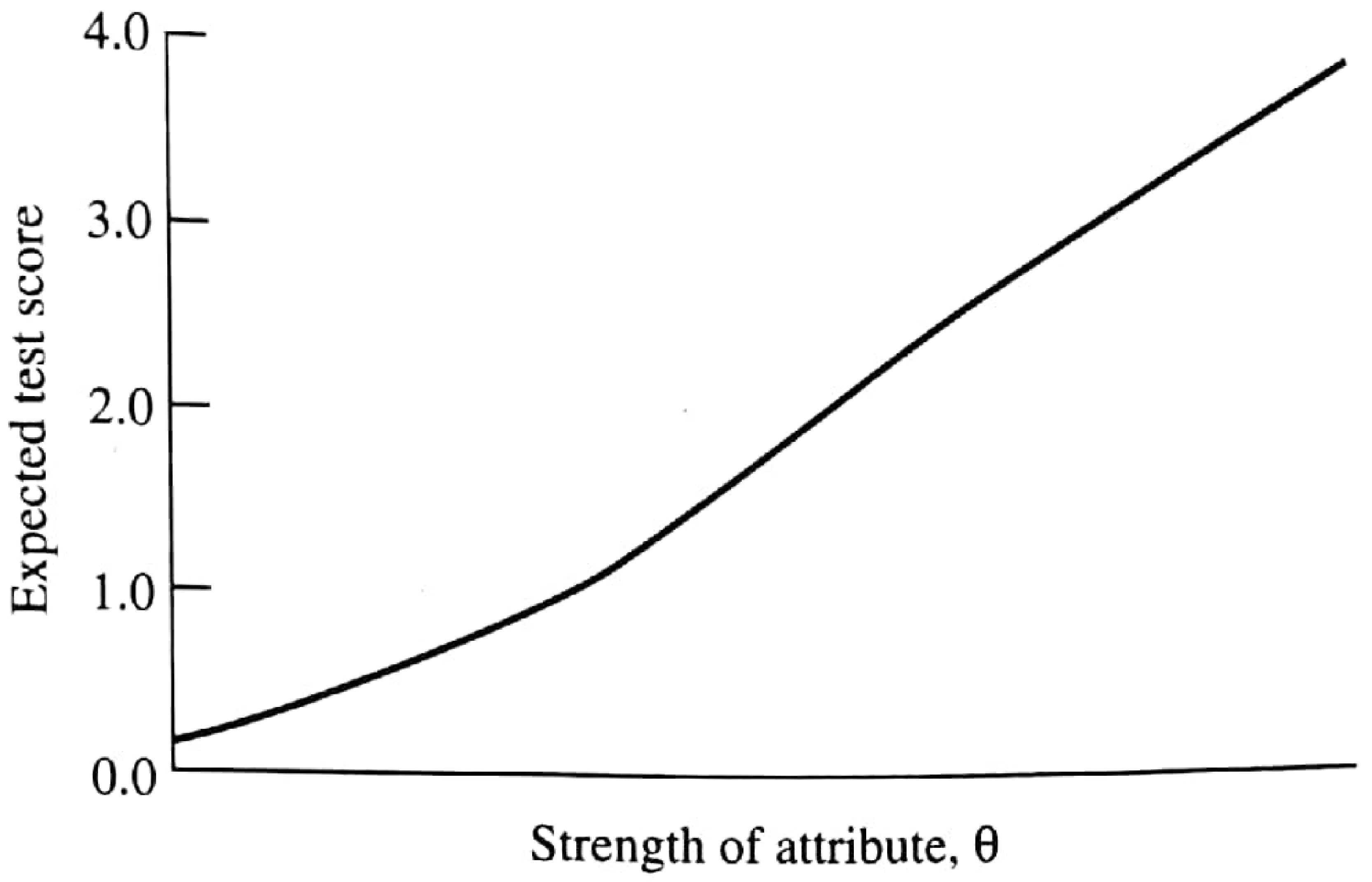


FIGURE 2-13 Expected scores on a four-item test—the sum of trace lines in Fig. 2-12.

We will normally assume that each item is given equal weight, called an equally weighted or unweighted model, but occasionally, weighting items differentially is appropriate. The effects of weighting are often trivial, especially when there are many items, and we will later argue against differentially weighting in most scale applications. The major features of the linear model apply to both weighted and unweighted versions. They also apply to multicategory or continuous items as well as to dichotomously scored items. Two slight drawbacks are that there is no formal rationale, in the representational sense, for a unit of measurement, and that the relation between total score and Θ may be nonlinear. However, as we have stressed, strong relations exist between linear scales and those developed from more complex models, and item sums ordinarily are monotonically related to Θ despite the nonlinearity.

We have come a long way around in this chapter to the conclusion that the most sensible way to measure psychological attributes of people is to do the obvious—sum item scores. The essence and beautiful point of the model is that it does not take individual items very seriously. It recognizes that any individual item has considerable specificity and measurement error. It does not make stringent assumptions about the form of the trace line. The only assumption made is that each item has some form of monotonic trace line. The model is fairly robust with respect to even that point in the sense that it is not highly sensitive to violations of this assumption. Even a few items with slight nonmonotonicities will not seriously affect the adequacy with which the attribute is measured. Items may have a noticeable *false positive* rate in that they may be answered correctly by subjects of the lowest imaginable ability (e.g., through guessing) and a *false negative* rate of being answered incorrectly by able subjects (e.g., through carelessness).

Much of the remainder of the book is based on the linear model, which makes sense and works well in practice. We will be discussing some newer models. However, there presently is no serious challenge to the linear model for most scaling of people and lower animals with respect to psychological attributes in the vast majority of applications.

SUMMARY

Constructing a scale begins with a plan, usually leading to a matrix representation of the data with subjects as rows and stimuli as columns. Two important considerations are whether stimuli or subjects (objects) are to be scaled and the types of judgments to be made by the subjects. Scaling models are typically more critical in scaling stimuli. The different kinds of judgments in turn can be traced back to psychophysics, the study of the relation between physically defined dimensions and their associated responses. There are three main questions in psychophysics: (a) How does one obtain the absolute threshold or point at which a stimulus is perceived 50 percent of the time? (b) How does one obtain the difference threshold (difference limen) or just-noticeable difference (*JND*)? (c) What is the overall relation between variation in a physical dimension and associated responses (psychophysical scaling). In particular, although the absolute threshold was thought of as an all-or-nothing effect (an event was either below threshold and not perceived or above threshold and perceived), nearly all data suggest that the function is continuous, usually in the form of an ogive (S curve).

There are two broad traditions in psychophysics. Fechner's indirect (discriminant) approach stresses ordinal judgments, particularly paired-comparison methods, and requires stimuli be confusable; it leads to a logarithmic relationship between physical magnitudes and associated sensations and can be used with a wide variety of subjects. Stevens' direct approach requires subjects to report intervals or ratios of perceived magnitudes as required. Its major methods are ratio production, ratio estimation, magnitude estimation, bisection, and cross-modal matching. It leads to a power function relating physical magnitude and sensation, although there is no necessary incompatibility between the two laws. Both lead to methods generally important in psychometrics. The Fullerton-Cattell *law* states that equally often noted differences are equal unless always or never noted is a basic link between Fechnerian psychophysics and psychometric theory. It led to Thurstone's law of comparative judgment. In turn, Thurstone scaling is closely related to signal detection theory which stresses the separation of bias in responding from accuracy of discrimination.

The concept of an item trace line (item characteristic curve) which relates the probability of a given response (response alpha) to the magnitude of an underlying attribute is extremely important. The Guttman scale was an early formal model for scaling people. It assumes that the item trace line is a step function. However, the similarity of the trace line to the psychometric function was noted, which suggests that the Guttman scale may be unrealistic. There are newer models, considered in Chapter 10, which make more realistic assumptions.

The simplest model for scaling people simply counts the number of responses in the alpha direction, perhaps weighting certain items over others. The only thing it requires of the item trace line is that it be monotonic. The chapter concluded by noting the utility of this linear (summative, centroid) model.

SUGGESTED ADDITIONAL READINGS

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

Engen, T. (1972a). Psychophysics I: Discrimination and detection. In J. W. Kling & L. A. Riggs (Eds.). *Woodworth & Schlossberg's Experimental Psychology* (3d ed.), vol. 1. New York: Holt, Rinehart, and Winston, chap. 1.

Engen, T. (1972b). Psychophysics II: Scaling Methods. In J. W. Kling & L. A. Riggs (Eds.). *Woodworth & Schlossberg's Experimental Psychology* (3d. ed.), vol. 1. New York: Holt, Rinehart, and Winston, chap. 2.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill, chaps. 2 and 10.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory.* Homewood, Ill.: Dow Jones-Irwin.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104,* 385–395.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,* 118–128.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Woodworth, R. S., & Schlossberg, H. (1954). *Experimental Psychology* (rev. ed.). New York: Holt.