7 Psychoacoustic Methods

Psychophysics is concerned with how we perceive the physical stimuli impinging upon our senses. The branch of psychophysics that deals with the perception of sound is **psychoacoustics**. In defining this term we make a sharp distinction between the physical **stimulus** and the psychological **response** to it. We may think of the sound presented to our ears as the stimulus and of what we hear as the response. For example, what we hear as *loudness* is the perceptual correlate of *intensity*: Other things being equal, a rise in intensity is perceived as an increase in loudness. Similarly, *pitch* is the *perception* related to sound *frequency*: Other things being equal, pitch gets higher as frequency increases.

If there were a single one-to-one correspondence between the physical parameters of sound and how they are perceived, then we could quantify what we hear directly in terms of the attributes of the sound. That would mean that all physically existing sounds could be heard, that all changes in them would be discriminable, and that any change in stimulus magnitude would result in a perceptual change of the same magnitude. This is not the case. It is thus necessary to describe the manner in which sound is perceived, and to attempt to explain the underlying mechanisms of the auditory system. This is the province of psychoacoustics.

SCALES OF MEASUREMENT

The study of auditory perception almost always involves measurements, the assignment of numbers that reflect the phenomena being investigated. Our ability to properly analyze what we find and to arrive at valid interpretations depends on knowing the properties of the measurements made and the qualities of the resulting data. Stevens' (1951, 1958, 1961, 1975) four scales of measurement provide us with this foundation.

Nominal scales are the least restrictive, in the sense that the observations are simply assigned to groups. This is the lowest order of scaling because the nominal label does not tell us anything about the relationship among the groups other than that they are different with respect to some parameter. For example, the nominal scale "gender" enables us to separate people into two categories, "male" and "female." All we know is that the two categories are differentiable, and that we can count how many cases fall into each one. The same would apply to the number of subcompact cars made by different manufacturers. We know that there are so many Fords, Toyotas, etc., but we have no idea of their relative attributes. A nominal scale, then, makes no assumptions about the order among the classes; thus, it is the least restrictive and least informative of the levels of scaling.

Ordinal scales imply that the observations have values, which can be rank-ordered so that one class is greater or lesser than another with respect to the parameter of interest. However, an ordinal scale does not tell us how far apart they are. Consider the relative quality of artistic reproductions. Painter A may produce a better reproduction of the Mona Lisa than painter B, who in turn makes a better copy than painter C, and so on. However, there may be one magnitude of distance between A and B, a second distance between B and C, and still a third distance between C and D. An ordinal scale thus gives the rank order of the categories ($A > B > C \ldots$), but does not specify the distances between them. Whereas the nominal scale allows us to express the **mode** of the data (which category contains more cases than any other), ordinal scales permit the use of the **median** (the value with the same number of observations above and below it). However, the lack of equal distances between values precludes the use of most mathematical operations. Sometimes the nature of the categories enables some of them to be rank-ordered, but not others. This constitutes a *partially ordered scale* (Coomb, 1953), which lies between the nominal and ordinal scales.

An **interval scale** specifies both the order among categories and the fixed distances among them. In other words, the distance between any two successive categories is equal to the distance between any other successive pair. Interval scales, however, do not imply a true zero reference point. Examples are temperature (in degrees Celsius or Fahrenheit) and the dates on a calendar. In contrast to nominal and ordinal data, equal distances between category values make it possible to use most mathematical operations with interval data. For example, the central tendency of interval data may be expressed as a *mean* (average). However, interval data cannot be expressed as proportions (ratios) of one another, because a true zero point is not assumed. It is also possible to rank the categories in such a way that there is an ordering of the distances between them. For example, the distances between successive categories may become progressively longer, as follows:

A – B – – C – – – D – – – – E – – – – – F – – – – – – G . . .

This is an *ordered metric scale* (Coomb, 1953). An ordered metric scale actually falls between the definitions of ordinal and interval scales, but may be treated as an interval scale (Abelson and Tukey, 1959).

Ratio scales include all the properties of interval scales as well as an inherent zero point. The existence of a *true zero* point permits values to be expressed as ratios or in decibels, and the use of all mathematical operations. As the most restrictive level, ratio scales give the most information about the data and their interrelationships. Examples are length, time intervals, and temperature (in kelvins), as well as loudness (sone) and pitch (mel) scales.

MEASUREMENT METHODS

Establishing relationships between the sound presented and how the subject perceives it is a primary goal. To accomplish this goal, the investigator contrives a special situation designed

to home in on the relation of interest. An experimental situation is used to avoid the ambiguities of presenting a stimulus and, in effect, asking the open-ended question "What did you hear?" Instead, the stimulus and response are clearly specified, and then some aspect of the stimulus (intensity, frequency, etc.) is manipulated. The subject's task is to respond in a predetermined manner so that the investigator can get an unambiguous idea of what was heard. For example, one may vary the intensity of a tone and ask the subject whether it was heard during each presentation. The lowest level at which the sound is heard (the transition between audibility and inaudibility) might be considered an estimate of **absolute sensitivity**. Alternatively, two tones might be presented, one of which is varied in frequency. The subject is asked whether the varied tone is higher (or lower) in pitch, and the smallest perceivable frequency difference—the just noticeable difference (jnd)—might be considered an estimate of **differential sensitivity**.

We must also distinguish between what the subject actually hears and the manner in which he responds. The former is **sensory capability** or **sensitivity**, and the latter is **response proclivity**. For the most part, we are interested in sensory capability. Response proclivity reflects not only the subject's sensitivity, but also the biases and criteria that affect how he responds. We therefore try to select measurement methods and techniques that minimize the effects of response bias. An excellent discussion of the many details to be considered in psychoacoustic experiments is given in Robinson and Watson (1973). In this chapter, we shall be concerned with classical psychophysical methods, adaptive techniques, and some aspects of scaling. Chapter 8 covers the theory of signal detection.

CLASSICAL METHODS OF MEASUREMENT

There are three classical psychophysical methods: limits, adjustment, and constant stimuli.

Method of Limits

In the **method of limits**, the stimulus is under the investigator's control and the subject simply responds after each presentation. Suppose we are interested in the **absolute sensitivity** or **threshold** for a particular sound. The sound is presented at a level expected to be well above threshold. Since it is clearly audible, the subject responds by saying that he heard the sound (+) in Fig. 7.1. The level of the sound is then decreased by a discrete amount (2 dB in Fig. 7.1) and presented again. This process is repeated until the subject no longer perceives the sound (−), at which point the series (or run) is terminated. This example involves a descending run. In an ascending series, the sound is first presented at a level known to be below the threshold and is increased in magnitude until a positive (+) response is obtained. The odd-numbered runs in Fig. 7.1 are descending series and the even-numbered runs are ascending. Since the crossover between "hearing" and "not hearing" lies somewhere

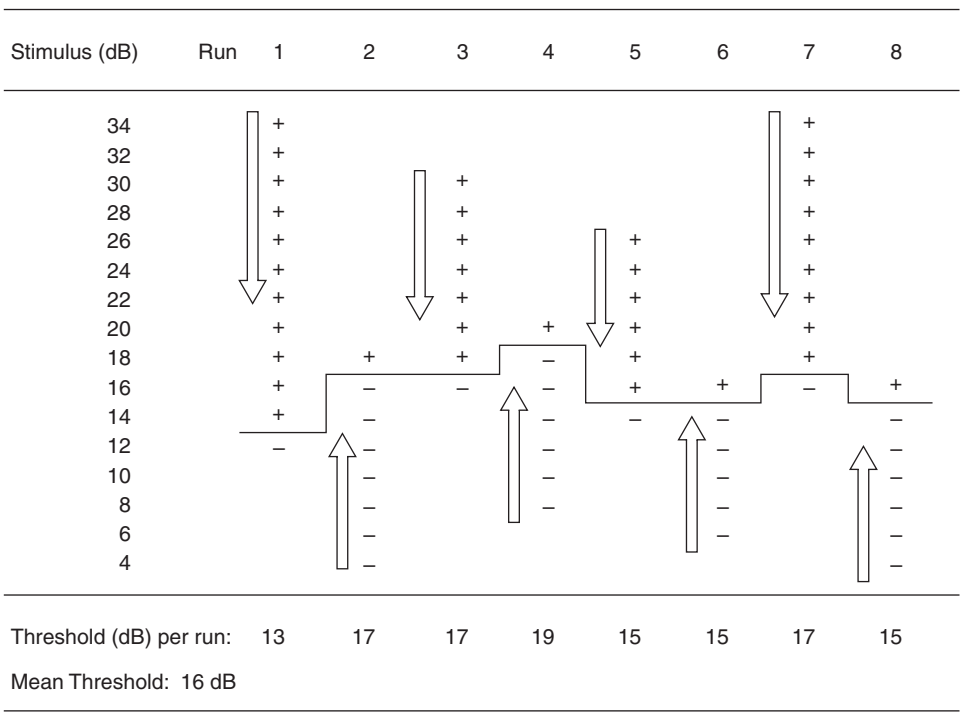

Figure 7.1 An example of the method of limits in hypothetical threshold experiments.

between the lowest audible level and the highest inaudible one, the "threshold" for each series may be taken as the halfway point between them. The subject's threshold is obtained by averaging the thresh old levels across runs. This average is 16 dB for the data in Fig. 7.1.

Several forms of **response bias** are associated with the method of limits. Since a series either ascends or descends, and is terminated by a change in response, the subject may anticipate the level at which his response should change from "no" to "yes" in an ascending run and from "yes" to "no" in a descending series. **Anticipation** thus results in a lower (better) ascending threshold because the subject anticipates hearing the stimulus, and a higher (poorer) descending threshold since he anticipates not hearing it. An opposite affect is caused by **habituation**. Here, the subject does not change his response from "no" to "yes" during an ascending run until the actual threshold is exceeded by a few trials (raising the measured threshold level), and he continues to respond "yes" for one or more descending trials after the sound has actually become inaudible (lowering the measured threshold level). These biases may be minimized by using an equal number of ascending and descending test runs in each threshold determination. These runs may be presented alternatively (as in the figure) or randomly. A second way to minimize these biases is to vary the starting levels for the runs. Both tactics are illustrated in Fig. 7.1.

The method of limits is also limited in terms of step size and inefficiently placed trials. Too large a step size reduces accuracy because the actual threshold may lie anywhere between two discrete stimulus levels. For example, a 10-dB step is far less precise than a 2-dB increment; and the larger the increment between the steps, the more approximate the result. Too large a step size may place the highest inaudible presentation at a level with a 0% probability of response, and the lowest audible presentation at a level with a 100% probability of response. The 50% point (threshold) may be anywhere between them! To make this point clear, consider the psychometric functions in Fig. 7.2. A **psychometric function** shows the probability (percentage) of responses for different stimulus levels. Figure 7.2a shows the psychometric function for a particular sound. It is inaudible (0% responses) at 13 dB and is always heard (100% responses) at 21 dB. It is customary to define the **threshold** as the level at which the sound is heard 50% of the time (0.5 probability). The threshold in Fig. 7.2a is thus 17 dB. Suppose we try to find this threshold by using a 10-dB step size, with increments corresponding to 14 dB, 24 dB, etc. Notice that this step size essentially includes the whole psychometric function, so that we do not know where the responses change from 0% to 100%, nor do we know whether they do so in a rapid jump (a step function) or along a function where gradual changes in the proportion of "yes" responses correspond to gradual changes in stimulus level. The result is low precision in estimating the location of the 50% point. However, a large step size is convenient in that it involves fewer presentations (and thus shorter test time), since responses go from "yes" to "no" in very few trials, each of which is either well above or well below threshold.

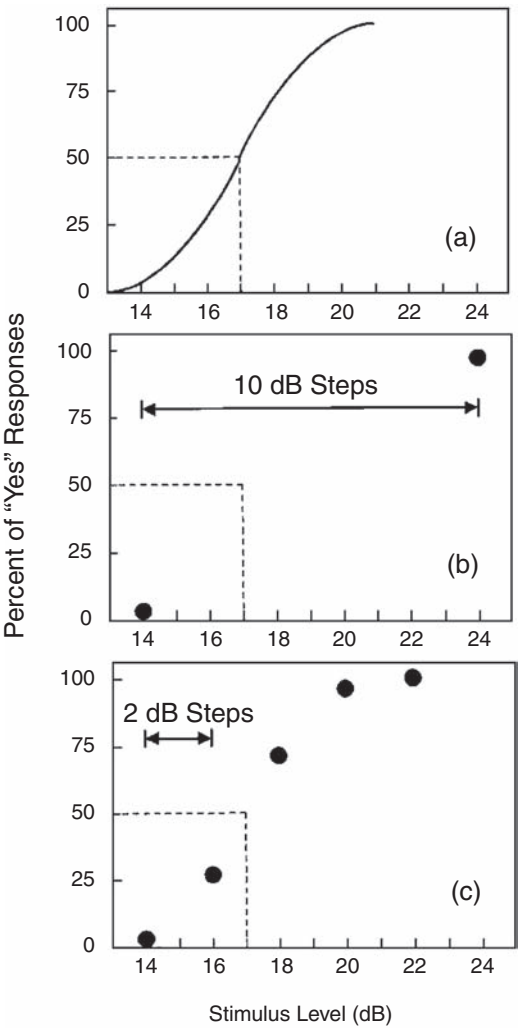

Figure 7.2 (a) Psychometric function showing the 50% threshold at 17 dB. (b) Responses obtained at near 0% at 14 dB and 100% at 24 dB with a 10-dB step size. (c) Percent responses at various levels with a 2-dB step size. The 50% threshold is shown on each graph. The 2-dB and 10-dB step sizes are illustrated at the top of the figure.

A smaller step size permits a more precise estimate of threshold because the reversals from "yes" to "no" (and vice versa) are better placed (closer) in relation to the 50% point. The relationship of a 2-dB step to the psychometric function is shown in Fig. 7.2c, which gives the probability of a response in 2-B intervals. Notice that these points are better placed than those for the 10 dB step size in Fig. 7.2b. For this reason, even though there may be "wasted" presentations due to test levels well above or below the threshold, the method of limits with an appropriate step size is still popular. This is particularly true in pilot experiments and in clinical evaluations, both of which take advantage

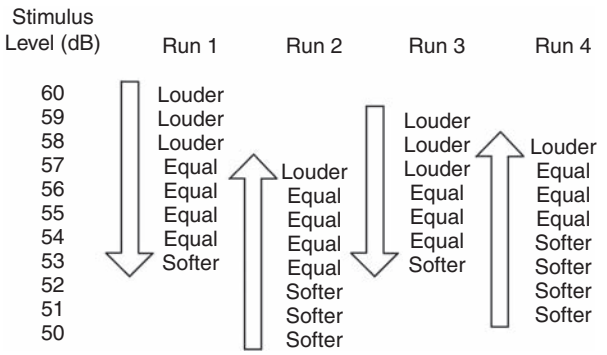

Figure 7.3 An example of the method of limits in a hypothetical discrimination experiment.

of the speed with which thresholds are estimated by the method of limits. The clinical method of limits (e.g., ASHA, 2005), however, is actually a hybrid technique with characteristics of the staircase method, discussed below.

The method of limits may also be used to determine **differential thresholds**. In this case, two stimuli are presented in each trial, and the subject is asked whether the second is greater than, less than, or equal to the first with respect to some parameter. The first stimulus is held constant, and the second is varied by the investigator in discrete steps. The procedure is otherwise the same as for determining thresholds, although the findings are different. Suppose the subject is to make an equal loudness judgment. The method of limits would result in a range of intensities in which the second stimulus is louder than the first, a range in which the second is softer, and a range in which the two sounds appear equal. In Fig. 7.3, the average **upper limen** (halfway between "higher" and "equal") is 57 dB, and the average **lower limen** (halfway between "equal" and "lower") is 53.5 dB. The range between these values is the **interval of uncertainty**, which is $57 - 53.5 = 3.5$ dB wide in this example. Although there is a range of "equal" judgments, we may estimate the "equal level" to lie halfway between the upper and lower limens, at 55.25 dB in this example. This point is commonly referred to as the **point of subjective equality (PSE)**. The **just noticeable difference (jnd)** or **difference limen (DL)** is generally estimated as one-half of the uncertainty interval, or 1.75 dB for the data in Fig. 7.3.

Method of Adjustment

The **method of adjustment** differs from the method of limits in two ways. First, the stimulus is controlled by the subject instead of by the investigator. In addition, the level of the stimulus is varied continuously rather than in discrete steps. As in the method of limits, the level is adjusted downward from above threshold until it is just inaudible, or increased from below threshold until it is just audible. Threshold is taken as the average of the just audible and just inaudible levels. To obtain an estimate of differential sensitivity, the subject adjusts the level of one sound until it is as loud as a standard sound, or adjusts the frequency of one sound until it has the same pitch as the other.

The stimulus control (generally a continuous dial) must be unlabeled and should have no detents that might provide tactile cues that could bias the results. Furthermore, a second control is sometimes inserted between the subject's dial and the instrumentation, allowing the investigator to vary the starting point of a test series by an amount unknown to the subject. This strategy avoids biases based on the positioning of the response dial and to the use of dial settings as "anchors" from one series to the next. Even with these precautions, however, it is difficult for the investigator to exercise the same degree of control over the procedure as in the method of limits. Furthermore, the subject may change his criterion of audibility during test runs, introducing another hard-to-control bias into the method of adjustment.

Just as anticipation and habituation affect the results obtained with the method of limits, stimulus persistence (perseveration) biases the results from the method of adjustment. Persistence of the stimulus means that a lower threshold is obtained on a descending run because the subject continues to turn the level down below threshold as though the sound were still audible. Thus, we may think of this phenomenon as **persistence of the stimulus**, or as **perseveration of the response**. In an ascending trial, the absence of audibility persists so that the subject keeps turning the level up until the true threshold is passed by some amount, which has the opposite effect of raising the measured threshold level. These biases may be minimized by using both ascending and descending series in each measurement. Another variation is to have the subject **bracket** his threshold by varying the level up and down until a just audible sound is perceived. After the ending point is recorded, the investigator may use the second stimulus control discussed above to change the starting level by an amount unknown to the subject, in preparation for the next trial.

Method of Constant Stimuli

The **method of constant stimuli** (or **constants**) involves the presentation of various stimulus levels to the subject in random order. Unlike the methods of limits and adjustments, the method of constants is a nonsequential procedure. In other words, the stimuli are not presented in an ascending or descending manner. A range of intensities is selected which, based upon previous experience or a pilot experiment, encompasses the threshold level. A step size is selected, and the stimuli are then presented to the subject in random order. In an absolute sensitivity (threshold) experiment, an equal number of stimuli are presented at each level. The subject indicates whether the stimulus presentation has been perceived during each test trial. In a differential sensitivity (DL) experiment, the subject's task would be to say whether two items are the same or different.

In an experiment to determine the threshold for a tone by using the method of constant stimuli, one might randomly present tones in 1-dB increments between 4 and 11 dB, for a total of 50 trials at each level. Sample results are tabulated

Table 7.1 Threshold of a Tone Using the Method of Constant Stimuli

Stimulus level (dB)	Number of responses	Percent of responses
11	50	100
10	50	100
9	47	94
8	35	70
7	17	34
6	3	6
5	0	0
4	0	0

Table 7.2 Data from an Experiment on Differential Sensitivity for Intensity Using the Method of Constant Stimuli

Level of second tone (dB)	Percentage of louder judgments
70	100
68	95
66	85
64	70
62	55
60	35
58	20
56	10
54	8
52	5
50	0

in Table 7.1. When these data are graphed in the form of a psychometric function (Fig. 7.4), the 50% point corresponds to 7.5 dB, which is taken as the threshold.

Table 7.2 shows the results of an experiment using the method of constants to find differential sensitivity for intensity. Two tones are presented and the subject is asked whether the second tone is louder or softer than the first. The intensity of the second tone is changed so that the various stimulus levels are presented randomly. Table 7.2 shows the percentage of presentations in which the subject judged the second tone to be louder than the first tone at each of the levels used. (The percentage of "softer" judgments is simply obtained by subtracting the percentage of "louder" judgments from 100%. Thus, the 60-dB presentations of the second tone were "softer" 100% − 35% = 65% of the time.) Figure 7.5 shows the psychometric function for these data. Because the intensity at which the second tone is judged louder 50% of the time is also the tone for which it was judged softer half of the time, the 50% point is where the two tones were perceived as equal in loudness. This is the PSE. In experiments of this kind, the 75% point is generally accepted as the threshold for "louder" judgments. (If we had also plotted "softer" judgments, then the 75% point on that psychometric function would constitute the "softer" threshold.) The DL is taken as the difference in stimulus values between the PSE and the "louder" threshold. For the data in Fig. 7.5, this difference is 64.8 dB − 61.5 dB = 3.3 dB.

The method of constant stimuli enables the investigator to include "catch" trials over the course of the experiment. These are intervals during which the subject is asked whether a tone was heard, when no tone was really presented. Performance on catch trials provides an estimate of guessing, and performance on real trials is often corrected to account for this effect (see Chap. 8). This correction reduces, but does not completely remove, response biases from the results.

The method of constants has the advantage over the methods of limits and adjustments of greater precision of measurement, and, as just mentioned, has the advantage of allowing direct estimation of guessing behavior. However, it has the disadvantage of inefficiency, because a very large number of trials are needed

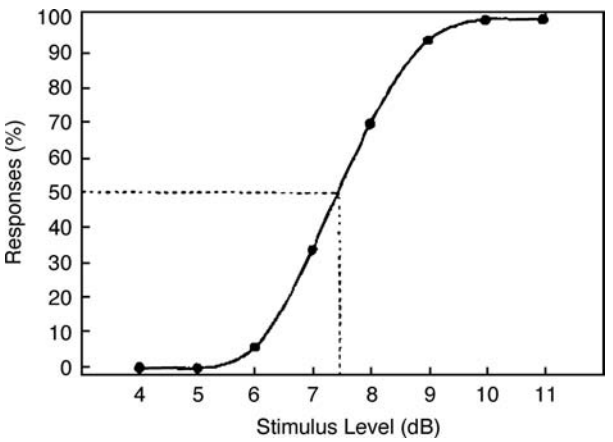

Figure 7.4 Psychometric function based on data from Table 7.1 obtained by using the method of constant stimuli. The threshold corresponds to 7.5 dB.

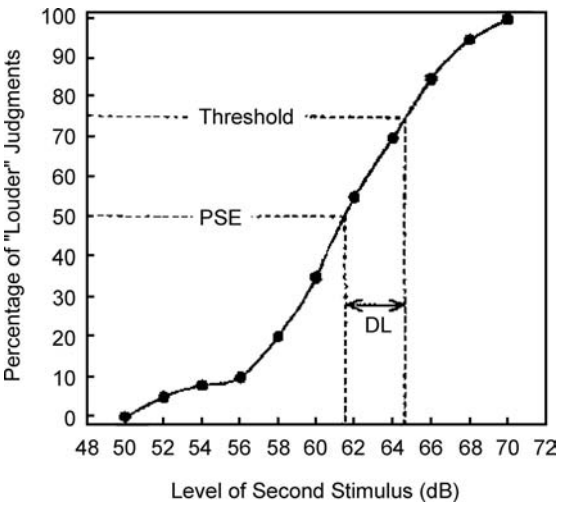

Figure 7.5 Psychometric function for a differential sensitivity experiment showing the point of subjective equality (PSE), "higher" threshold, and difference limen (DL). The data are from Table 7.2.

to obtain the data. Most of these trial points are poorly placed relative to the points of interest (generally the 50% and 75% points), so that the method of constants costs heavily in time and effort for its accuracy. The prolonged test time increases the effects of subject fatigue and the difficulty of maintaining motivation to respond.

FORCED CHOICE METHODS

Until now, we have focused for the most part on a "yes/no" testing approach. However, other formats are used as well and are actually more commonly employed in actual experiments. These approaches involve forced choice paradigms in which the subject is presented with two or more alternatives from which he must choose a response. Suppose, for example, that we want to find out whether a subject can hear a tone in the presence of a noise. In a "yes/no" experiment the subject hears one stimulus presentation, which might be a just a tone, or perhaps a noise alone versus a tone-plus-noise combination. In either case, the subject's task is to indicate whether the tone was there or not ("yes" or "no"). In a **two-alternative forced choice (2AFC) method**, the subject is presented with two stimuli in succession, only one of which contains the tone. After listening to both stimuli, he must decide whether the tone was present in the first one or the second. Similarly, in a 4AFC experiment, the subject must decide which of four successive stimuli includes the tone. Because the two or more presentations occur as successive *intervals*, we could also say that the subject must decide which interval contained the stimulus. Therefore, these experiments are often called **2- (or more) interval forced choice methods** (hence, 2-IFC, 3-IFC, etc.). These topics are covered further in the context of the theory of signal detection in the next chapter.

ADAPTIVE PROCEDURES

In an **adaptive procedure**, the level at which a particular stimulus is presented to the subject depends upon how the subject responded to the previous stimuli (Wetherill and Levitt, 1965; Levitt, 1971; Bode and Carhart, 1973). Broadly defined, even the classical method of limits can be considered an adaptive method because of its sequential character and the rule that stimuli are presented until there is a reversal in the subject's responses from "yes" to "no" or vice versa. However, use of the term "adaptive procedures" has come to be associated with methods that tend to converge upon the threshold level (or some other target point), and then place most of the observations around it. This approach, of course, maximizes the efficiency of the method because most of the test trials are close to the threshold rather than being "wasted" at some distance from it. It also has the advantage of not requiring prior knowledge of where the threshold level is located, since adaptive methods tend to home in on the threshold regardless of the starting point, and often include step sizes which are large at first and then become smaller as the threshold level is approached. As a result, both efficiency and precision are maximized.

Bekesy's Tracking Method

Bekesy (1960/1989) devised a **tracking method** which shares features with both the classical methods of adjustment and limits and with adaptive procedures. The level of the stimulus changes at a fixed rate (e.g., 2.5 dB/s) under the control of a motor-driven attenuator, and the direction of level change is controlled by the subject via a pushbutton switch. The motor is also connected to a recorder, which shows the sound level as a function of time (Fig. 7.6) or frequency. The pushbutton causes the motor to decrease the sound level when it is depressed and to increase the level when it is up. The subject is asked to press the button whenever he hears the tone and to release it whenever the tone is inaudible. Thus, the sound level is increased toward threshold from below when the tone is inaudible and decreased toward threshold from above when the sound is heard. The threshold is thus tracked by the subject, and its value is the average of the midpoints of the excursions on the recording (once they are stabilized).

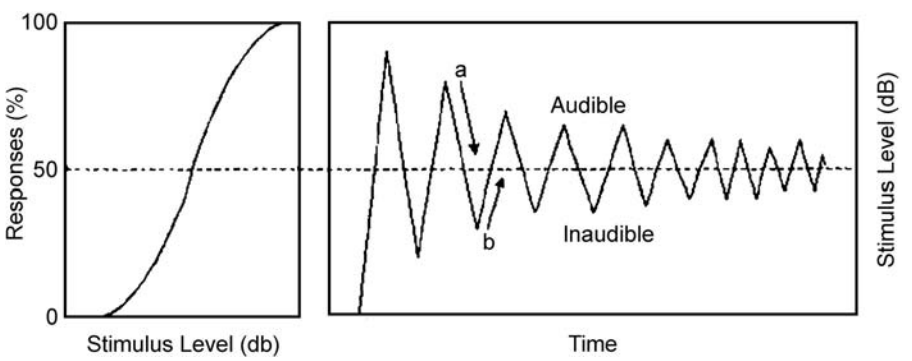

Figure 7.6 Bekesy's tracking method. (a) Intensity decreases as the subject depresses the response button when he hears the sound. (b) Intensity increases as the subject releases the button when he cannot hear the sound. The midpoints of the excursions correspond to the 50% point on the psychometric function shown to the left of the tracing.

Tracking has the advantages of speed and reasonable precision. It is, of course, subject to several sources of response bias. At fast attenuation rates (intensity change speeds), the subject's reaction time can substantially affect the width of the tracking excursions and the precision of measurement. For example, if the tone increases and decreases in level at 5 dB/s and a subject has a 2-s reaction time, then the motor will have advanced the stimulus level (and pen position on the recorder) 10 dB above threshold before the button is finally depressed. Precision is improved and reaction time becomes less critical at reasonably slower attenuation rates, although the tracking may meander somewhat on the recording as the subject's criterion for threshold varies.

Simple Up-Down or Staircase Method

The **simple up-down (or staircase) method** involves increasing the stimulus when the subject did not respond to the previous stimulus presentation and decreasing the intensity when there was a response to the prior stimulus (Dixon and Mood, 1948; Levitt, 1971). It differs from the method of limits in that testing does not stop when the responses change from "yes" to "no" or from "no" to "yes." Similar to the method of limits, the stimuli are changed in discrete steps.

Figure 7.7 shows the first six runs of a staircase procedure to find the threshold of a tone using a 2-dB step size. Here, a run is a group of stimulus presentations between two response reversals. In other words, a descending run starts with a positive response and continues downward until there is a negative response, while an ascending run begins with a negative response and ends with a positive one. Because stimulus intensity is always increased after a negative (−) response and decreased after a positive (+) response, the staircase method converges upon the 50% point on the psychometric function. The procedure is continued through at least six to eight reversals (excluding the first one), and the threshold value is then calculated as the average of the midpoints of the runs, or as the average of their peaks and troughs (Wetherill, 1963; Wetherill and Levitt, 1965). The latter method appears to give a somewhat better estimate. The precision of the method can be increased by first estimating the threshold with a larger step size, and then using a smaller step size (generally half that of the previous one) to locate the threshold in the vicinity of the first estimate (Wetherill, 1963). For example, if the average of six runs using a 4-dB step is 10 dB, a second group of runs using a 2-dB step might begin at 10 dB in order to obtain a more precise estimate of the threshold.

The simple up-down method has several advantages and limitations (Levitt, 1971). It quickly converges upon the 50% point so that most trials are efficiently placed close to the point of interest. It also has the advantage of being able to follow changes (drifts) in the subject's responses. On the other hand, the subject may bias his responses if he realized that the stimuli are being presented according to a sequential rule, which depends on the way he responds. As with the method of limits, if the step size is too small, a large number of trials are wasted, and if the step is too large, they are badly placed for estimating the 50% point. Another limitation is that only the 50% point can be converged upon with the simple up-down rule.

Parameter Estimation by Sequential Testing

Parameter estimation by sequential testing (PEST) is an adaptive procedure, which uses changes in both the direction and step size of the stimulus to home in on a targeted level of performance (Taylor and Creelman, 1967; Taylor, Forbes, and Creelman, 1983). The investigator may set the target value to any location on the psychometric function he chooses (for example, 50% or 80%). However, we will concentrate here only on the 50% point in order to make clear the salient features which distinguish the PEST procedure. As in the simple up-down method, positive responses are followed by decreases in stimulus level because the threshold is probably lower, and negative responses are followed by increases in intensity because the threshold is probably higher. The difference is that PEST includes a series of *rules for doubling and halving the stimulus level* depending upon the previous sequence of responses.

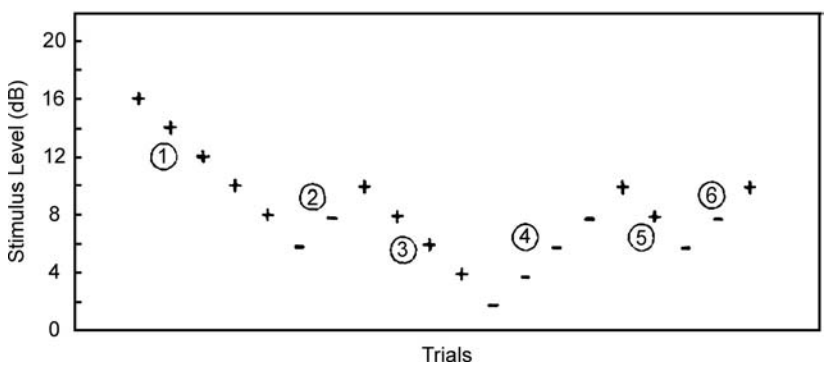

Figure 7.7 The first six runs of a threshold search using the simple up-down or stair case method. Each (+) indicates a positive response and each (−) indicates a negative response. Odd numbers are descending runs and even numbers are ascending runs. The first reversal is generally omitted from the threshold calculation.

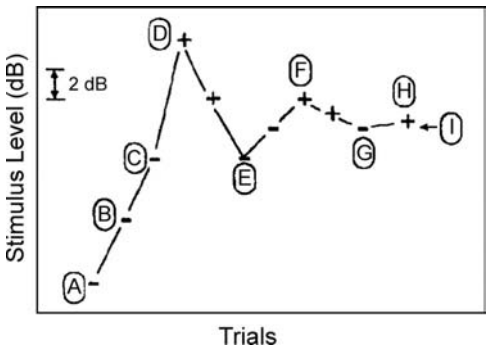

Figure 7.8 An example of how the threshold is obtained with the PEST procedure. Points identified by letters are discussed in the text. Point I is the estimate of threshold.

At each stimulus level, PEST in effect asks whether the threshold has been exceeded. The level is then changed so that the maximum amount of information is obtained from the next trial. To do this, the step size is varied in the manner specified in Fig. 7.8. Although it is most efficient to know the approximate location of the threshold range in advance, it is not essential. Suppose we begin testing at some value below threshold corresponding to point A in Fig. 7.8. Since the subject gives a negative response, the stimulus is presented at a higher level (B). This level also produces no response and the stimulus is raised by the same amount as previously and is presented again (C). Since there is still no response, the stimulus level is again increased. However, PEST has a rule, which states that if there is a negative response on two successive presentations in the same direction, then the step size is doubled for the next presentation. Thus, the next stimulus is presented at level D. The doubling rule ensures that a minimal number of trials are wasted in finding the range of interest.

A positive response at level D indicates that the threshold has been exceeded. As in the staircase method, the direction of the trials is changed after a response reversal. However, the PEST procedure also halves the step size at this point. The halving rule causes the stimuli to be presented closer to the threshold value. Thus, precision is improved as the threshold is converged upon. Since D is followed by another positive response, the stimulus is then presented at a lower level (E). A negative response at E causes the direction of stimulus change to be changed again, and the step size is halved compared to the previous one. The stimulus is heard again at the next higher level (F), so the direction is changed again and the step size is again halved. Stimuli are now presented in a descending run until there is a negative response (G). Halving the step size and changing direction results in a positive response at H, indicating that the threshold lies somewhere between points G and H. Since this interval represents an acceptable degree of precision, the procedure is terminated. The level at which the next stimulus would have been presented is taken as the threshold. This level is point I, which lies halfway between levels C and H. Note on the scale for Fig. 7.8 that the step size between E and F is 2 dB, between F and G is 1 dB, and between G and H is 0.5 dB. This observation highlights the rapidity with which PEST results in a precise threshold estimate.

Block Up-Down Methods

Suppose we are interested in the 75% point on the psychometric function. One way to converge upon the point is to modify the simple up-down procedure by replacing the single trial per stimulus level with a block of several trials per level. Then, by adopting three out of four positive responses (75%) as the criterion per level, the strategy will home in on the 75% point. If blocks of five were used with a four out of five criterion, then the 80% point would be converged upon. The procedure may be further modified by changing the response from *yes-no* to a *two-alternative* (*interval*) *forced choice*. In other words, the subject is presented with two stimulus intervals during each trial and must indicate which of the intervals contains the stimulus. This is the **block up-down temporal interval forced-choice (BUDTIF) procedure** (Campbell 1963). Using the two-interval forced choice method allows the investigator to determine the proportion of responses to the no-stimulus interval—the "false alarm" rate. We shall see when the theory of signal detection is discussed in the next chapter that this distinction is important in separating sensitivity from bias effects.

The BUDTIF procedure is illustrated in Fig. 7.9. Note that each block is treated as though it were one trial in a staircase procedure. Since the target point has been preselected as 75%,

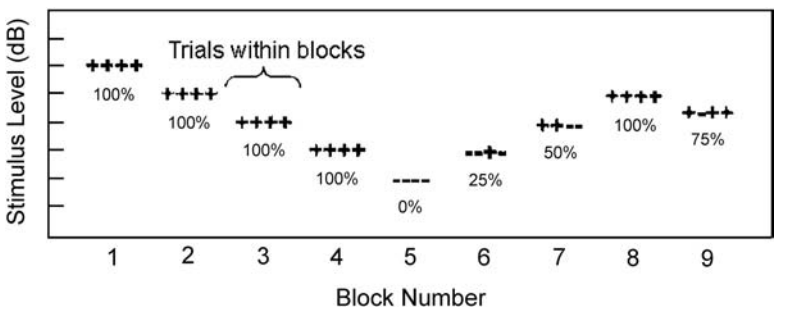

Figure 7.9 An example of convergence upon the 75% point of the psychometric function using BUDTIF.

stimulus intensity is raised whenever there are less than three out of four correct responses in a block and is decreased when all four are correct. Testing is terminated when three out of four correct responses are obtained. The last level is the target level (75% in this case). Notice that since blocks of trials are presented at each level, poor placement of the initial test level will cause many wasted trials in converging upon the target range.

A modification of BUDTIF replaces the two-alternative forced-choice paradigm with the familiar yes-no response. This adaptation is called the **block up-down yes-no (BUDYEN)** method (Campbell and Counter, 1969). However, the BUDYEN paradigm is less advantageous than its forced-choice predecessor because an estimate of false alarms is not obtained (Creelman and Taylor, 1969).

Transformed Up-Down or Staircase Procedures

The simple up-down method converges on the 50% point of the psychometric function because each positive response leads to a decrease in stimulus level and each negative response leads to an intensity increase. If the up-down rule is modified so that stimulus level is changed only after certain sequences have occurred, then the procedure will home in on other points on the psychometric function (Wetherill and Levitt, 1965; Levitt and Rabiner, 1967; Levitt, 1971, 1978). These other target points depend upon the particular set of sequences chosen by the investigator.

We will go over the fundamental principles of transformed up-down methods because they are ubiquitous in hearing science research. When the target is 50% on the psychometric function, as in the simple up-down method, the chances of a positive response to stimuli well below the 50% point are very small. Similarly, it is likely that stimuli presented at levels well above the 50% point will frequently be heard. However, as the intensity corresponding to 50% is approached, the chances of positive and negative responses become closer and closer. At the 50% point, the probability of a positive response is the same as of a negative one. This is, of course, exactly what we mean by 50%. Now, suppose that the total probability of all responses is 1.00. If we call the probability of a positive response (p), then the probability of a negative response would be $(1 - p)$. At the 50% point

$$p = (1 - p) = 0.5 \tag{7.1}$$

In other words, the probability of a positive response at the 50% point is 0.5, which is also the probability of a negative response. In effect, the simple up-down rule forces the intensity to the point on the psychometric function where the probabilities of positive and negative responses are equal (0.5 each).

Other target levels can be converged upon by changing the up-down rule so that the probabilities of increasing and decreasing stimulus intensity are unequal. This is done by setting the criteria for increasing stimulus level (the "up rule") to be a certain response sequence, and those for decreasing stimulus level (the "down rule") to be other response sequences. An example will demonstrate how the transformed up-down method works.

Suppose we are interested in estimating a point above 50% on the psychometric function, say 70%. To accomplish this, we would increase the stimulus level after a negative response $(-)$ or a positive response followed by a negative one $(+, -)$, and lower the stimulus level after two successive positives $(+, +)$. In other words, we have established the following rules for changing stimulus level:

$$\begin{aligned} &\text{Up rule :} \quad (-) \text{ or } (+, -) \\ &\text{Down rule :} \quad (+, +) \end{aligned} \tag{7.2}$$

As with the simple staircase rule, levels well above the target will often yield $(+, +)$ responses, and those well below will tend to have $(-)$ or $(+, -)$ responses. However, at the target level, the probability of increasing the stimulus level will be

$$(1 - p) + p(1 - p) \tag{7.3}$$

and the probability of two successive positive responses $(+, +)$ will be

$$p \times p \text{ or } p^2 \tag{7.4}$$

The up-down strategy will converge on the point where the up and down rules have the same probabilities (0.5). In other words, the probability of the transformed positive response $(+, +)$ at the target is

$$p^2 = 0.5 \tag{7.5}$$

Since we are interested in the probability (p) of a single positive response, which is the square root of p^2, we simply find the square root of $p^2 = 0.5$, and we obtain

$$p = 0.707 \tag{7.6}$$

Converting to percent, the transformed procedure just outlined homes in on the 70.7% point of the psychometric function, which is a quite acceptable estimate of the 70% point.

To converge on the 29.3% of the psychometric function (which is a reasonable estimate of the 30% point), we might choose to increase the stimulus after a sequence of two successive negative responses $(-, -)$, and to decrease stimulus level after a positive response $(+)$ or a negative response followed by a positive one $(-, +)$.

The 70.7% and 29.3% **transformed up-down strategies** are illustrated in Fig. 7.10. As with the simple up-down method, each transformed strategy would be continued through six to eight reversals, and the average of the peaks and valleys would be taken as the target level. Because these two points are on the rising portion of the psychometric function and are equidistant from 50%, a reasonably good estimate of the 50% point can be obtained by averaging the levels from 70.7% to 29.3%. To increase efficiency, one might start with a large step size, and then halve it in the target range for increased precision.

Other target points can be converged upon by various sequences of positive and negative responses, and different sequences may be used to converge upon the same target points

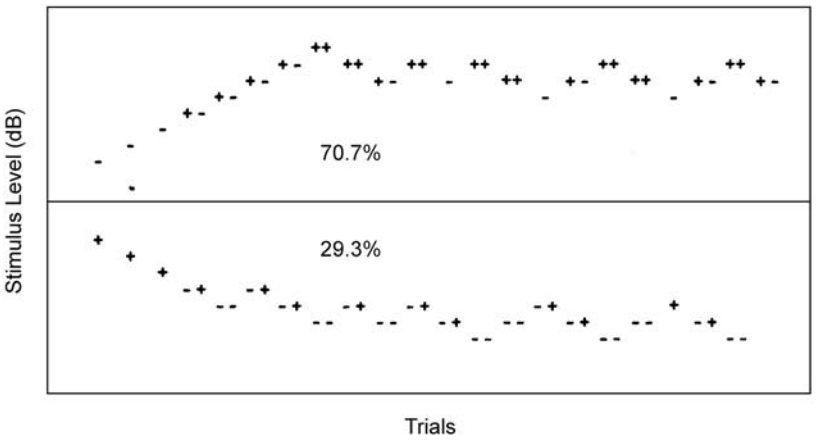

Figure 7.10 Examples of how the transformed up-down procedure converges upon the 70.7% point (*upper frame*) and the 29.3% point (*lower frame*) of the psychometric function.

(Levitt and Rabiner, 1967; Levitt, 1971, 1973). Transformed up-down procedures can also be used in the measurement of subjective judgments, such as for loudness balances (Jesteadt, 1980). In addition, both simple and transformed up-down procedures are particularly adaptable to testing various aspects of speech recognition functions under a variety of conditions (e.g., Levitt, 1971; Bode and Carhart, 1974; Plomp and Mimpen, 1979; Duquesnoy, 1983; Dubno, Morgan, and Dirks, 1984; Gelfand, Ross, and Miller, 1988).

A useful approach to minimizing biases is to interleave different testing strategies (Levitt, 1968). In other words, two points on the psychometric function are converged upon during the same test session. This is done by switching in an approximately random manner between the two test strategies. For example, two reversals on the 29.3% strategy might be followed by a few reversals on the 70.7% strategy, then the investigator would return to where he left off on the 29.3% sequence, and so forth. Such interleaving can also be applied to other psychoacoustic methods. Of course, greater flexibility and ease of measurement is made possible when the procedure is automated, and, almost needless to say, computerized administrations of these procedures are the norm.

Modifications, Other Procedures, and Comparisons

Numerous approaches have been introduced that modify the methods already discussed and/or combine adaptive procedures with maximum likelihood, Bayesian, or other techniques (Hall, 1968, 1981, 1983; Pentland, 1980; Watson and Pelli, 1983; Emmerson, 1984; Findlay, 1978; Simpson, 1989; Kaernbach, 1991; King-Smith, Grigsby, Vingrys, et al., 1994; Kontsevich and Tyler, 1999; Remus and Collins, 2008). The *maximum likelihood methods* use the history of the subject's responses combined with certain assumptions about the nature of the psychometric function to estimate where the threshold (actually the midpoint of the function) lies after each response. This estimated value then becomes the level of the next stimulus presentation. For example, Hall's (1981, 1983) hybrid procedure combines aspects of maximum likelihood methods with features of PEST. In the *Bayesian adaptive procedures*, the step size is adaptive rather than fixed and each stimulus value is calculated based on a running update of the probability distribution.[1]

Many studies have compared the various adaptive methods and between adaptive and more traditional approaches (Pentland, 1980; Shelton et al., 1983; Shelton and Scarrow (1984); Taylor et al., 1983; Hesse, 1986; Marshall and Jesteadt, 1986; Madigan and Williams, 1987; Kollmeier, Gilkey and Sieben, 1988; Simpson, 1988, 1989; Leek, 2001; Marvit, Florentine, and Buus, 2003; Amitay, Irwin, Hawkey, et al., 2006; Rowan, Hinton, and Mackenzie, 2006; Remus and Collins, 2008). Generally speaking, thresholds obtained with the various adaptive approaches tend to be relatively close to each other. For example, Shelton et al. (1983) found that thresholds were nearly the same when obtained using the transformed up-down, PEST, and maximum likelihood procedures. However, it does appear that somewhat better performance is obtained with forced choice compared to nonforced choice paradigms (Taylor et al., 1983; Hesse, 1986; Marshall and Jesteadt, 1986; Kollmeier et al., 1988), and with adaptive step sizes and Bayesian approaches than with fixed step size methods (Pentland, 1980; Leek, 2001; Marvit et al., 2003; Remus and Collins, 2008). One should consult these papers when deciding upon the most appropriate approach for a given experiment.

DIRECT SCALING

The methods discussed so far in this chapter, in which the subject's task is to detect the presence of or small differences between

[1] For readily available explanations of Bayesian probabilities and methods see, e.g., http://www.bayesian.org/bayesexp/bayesexp.html, or http://drambuie.lanl.gov/~bayes/tutorial.htm.

stimuli, are often referred to as **discriminability** or **confusion scales**. In contrast, **direct scaling** involves having the subject establish perceptual relationships among stimuli (**magnitude and ratio scales**) or to divide a range of stimuli into equally spaced or sized perceptual categories (**category or partition scales**). In other words, the subject must specify a perceptual continuum that corresponds to a physical continuum. Two types of continua may be defined (Stevens, 1961). **Prothetic continua**, such as loudness, have the characteristic of *amount*. They are *additive* in that the excitation due to an increase in stimulus level is added to the excitation caused by the intensity which was already present. On the other hand, pitch has the characteristic of *kind* and azimuth has the characteristic of *location*. These are **metathetic continua** and are *substantive* rather than additive. In other words, a change in the pitch corresponds to a substitution of one excitation pattern, as it were, for another.

Ratio Estimation and Production

In **ratio estimation**, the subject is presented with two stimuli differing in terms of some parameter and is asked to express the subjective magnitude of one stimulus as a ratio of the other. Subjective values are thus scaled as a function of the physical magnitudes. Suppose two 1000-Hz tones with different intensities are presented to a subject, who must judge the loudness of the second tone as a ratio of the first. He might report that the intensity of the second tone sounds one-half, one-quarter, twice, or five times as loud as the first tone.

Ratio production, or **fractionalization**, is the opposite of ratio estimation in that the subject's task is to adjust the magnitude of a variable stimulus so that it sounds like a particular ratio (or fraction) of the magnitude of a standard stimulus. For example, the subject might adjust the intensity of a comparison tone so that it sounds half as loud as the standard, twice as loud, etc. Fractionalization has been used in the development of scales relating loudness to intensity (Stevens, 1936) and pitch to frequency (Stevens, Volkmann, and Newman, 1937; Stevens and Volkmann, 1940).

Magnitude Estimation and Production

In **magnitude estimation**, the subject assigns to physical intensities numbers that correspond to their subjective magnitudes. This may be done in two general ways (Stevens, 1956, 1975). In the first method, the subject is given a standard or reference stimulus and is told that its intensity has a particular value. This reference point is called a **modulus**. He is then presented with other intensities and must assign numbers to these, which are ratios of the modulus. Consider a loudness scaling experiment in which the subject compares the loudness of variable tones to a standard tone of 80 dB. If the 80-dB standard is called 10 (modulus), then a magnitude estimate of 1 would be assigned to the intensity 1/10 as loud, 60 would be assigned to the one that is 6 times as loud, etc. The relationship between these magnitude estimates and intensity is shown by the closed circles in Fig. 7.11.

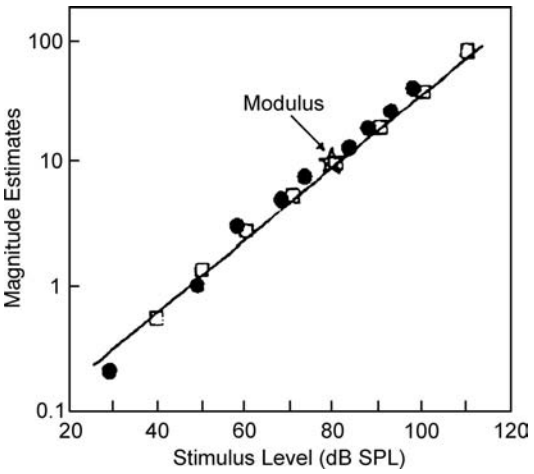

Figure 7.11 Magnitude estimations of loudness obtained with a modulus (*closed circles*) and without a modulus (*open squares*) as a function of stimulus intensity based on data from Stevens (1956).

An alternative approach is to omit the modulus. Here, the subject is presented with a series of stimuli and is asked to assign numbers to them reflecting their subjective levels. The results of such an experiment are shown by the open squares in Fig. 7.11. As the figure shows, magnitude estimates obtained with and without a modulus result in similar findings.

The reverse of magnitude estimation is **magnitude production**. In this approach, the subject is presented with numbers representing the perceptual values and must adjust the physical magnitude of the stimulus to correspond to the numbers.

Absolute magnitude estimation (AME) and absolute **magnitude production (AMP)** involve the performance of magnitude estimates (or productions) without any specified or implied reference value, and with each estimate (or production) made without regard to the judgments made for previous stimuli (Hellman and Zwislocki, 1961, 1963, 1968; Hellman, 1976, 1981; Zwislocki and Goodman, 1980; Zwislocki, 1983a; Hellman and Meiselman, 1988). There has been some discussion regarding this approach (e.g., Mellers, 1983a, 1983b; Zwislocki, 1983b). However, the convincing preponderance of evidence reveals that it is valid, reliable, and efficient, and that AMEs and AMPs are readily performed by naive clinical patients as well as laboratory subjects (Hellman and Zwislocki, 1961, 1963, 1968; Hellman, 1976, 1981; Zwislocki and Goodman, 1980; Zwislocki, 1983a; Hellman and Meiselman, 1988).

Subject bias causes magnitude estimation and production to yield somewhat different results, especially at high and low stimulus levels. Specifically, subjects tend not to assign extreme values in magnitude estimation, or to make extreme level adjustments in magnitude production. These bias effects are in opposite directions so that the "real" function lies somewhere between the ones obtained from magnitude estimations and productions. This is illustrated in Fig. 7.12 by the

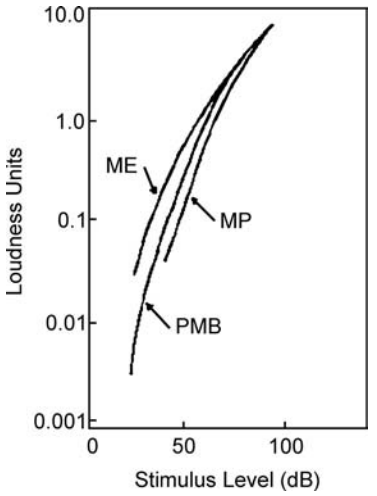

Figure 7.12 Bias effects in magnitude estimation (ME) and magnitude production (MP) are minimized by geometric averaging in the method of psychological magnitude balance (PMB). *Source*: Adapted from Hellman and Zwislocki (1968) with permission of *J. Acoust. Soc. Am.*

divergence of the magnitude estimation and magnitude production functions. An unbiased function may be obtained by using the method of **psychological magnitude balance** (Hellman and Zwislocki, 1963, 1968). This is done by calculating the geometric mean of the corresponding magnitude estimations and magnitude productions along the intensity axis or the loudness axis. An example is illustrated by the curve labeled PMB in Fig. 7.12.

Cross-Modality Matches

A scaling approach related to magnitude estimation and production is called **cross-modality matching** (Stevens and Guirao, 1963; Stevens and Marks, 1965, 1980; Stevens, 1975; Hellman and Meiselman, 1988, 1993). In this technique, the subject is asked to express the perceived magnitude for one sense in terms of another sensory modality. For example, *loudness* (an auditory perception) might be expressed in terms of *apparent line length* (a visual perception). A very useful variation of this approach has been developed and applied by Hellman and Meiselman (1988, 1993). In this method, the slope of the power function for loudness is derived from that for line length combined with the cross-modality match between loudness and line length.

Category Rating of Loudness

Category rating methods are often used in loudness measurements, particularly in clinical assessments related to hearing aids. These methods involve presenting sounds to the listener at various levels, who gives a loudness rating to each of them based on a list of descriptive loudness categories (e.g., Allen, Hall, Jeng, 1990; Hawkins, Walden, Montgomery, and Prosek, 1987; Cox, Alexander, Taylor, and Gray, 1997). For example, in the Contour Test developed by Cox et al. (1997), the listener assigns numerical loudness ratings to pulsed warble tone stimuli using a seven-point scale from *1* for "very soft," to *7* for "uncomfortably loud." Sherlock and Formby (2005) found no significant differences between sound levels rated "uncomfortably loud" using this approach and directly measured loudness discomfort levels (Sherlock and Formby, 2005).

REFERENCES

Abelson, RP, Tukey, JW. 1959. Efficient conversion of nonmetric information into metric information. In: ER Tufte (ed.), *The Quantitative Analysis of Social Problems*. Reading, MA: Addison Wesley, 407–417.

Allen, JB, Hall, JL, Jeng, PS. 1990. Loudness growth in $^1/_2$-octave bands (LGOB)—A procedure for the assessment of loudness. *J Acoust Soc Am* 88, 745–753.

American Speech-Language-Hearing Association (ASHA). 2005. *Guidelines for manual pure-tone threshold audiometry*. Rockville, MD: ASHA.

Amitay, S, Irwin, A, Hawkey, DJC, Cowan, JA, Moore, DR. 2006. A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *J Acoust Soc Am* 119, 1616–1625.

Bekesy, G. 1960/1989. *Experiments in Hearing*. New York, NY: McGraw-Hill. [Republished by the Acoustical Society of America].

Bode, DL, Carhart, R. 1973. Measurements of articulation functions using adaptive test procedures. *IEEE Trans Audiol Electroacoust* AU-21, 196–201.

Bode, DL, Carhart, R. 1974. Stability and accuracy of adaptive tests of speech discrimination. *J Acoust Soc Am* 56, 963–970.

Campbell, RA. 1963. Detection of a noise signal of varying duration. *J Acoust Soc Am* 35, 1732–1737.

Campbell, R.A, Counter, SA. 1969. Temporal energy integration and periodicity pitch. *J Acoust Soc Am* 45, 691–693.

Coomb, CH. 1953. Theory and methods of measurement. In: L Festinger, D Katz (eds.), *Research Methods in the Behavioral Sciences*. New York, NY: Holt, Rinehart, and Winston, 471–535.

Cox, RM, Alexander, GC, Taylor, IM, Gray, GA. 1997. The Contour Test of loudness perception. *Ear Hear* 18, 388–400.

Creelman, CD, Taylor, MM. 1969. Some pitfalls in adaptive testing: Comments on "Temporal integration and periodicity pitch". *J Acoust Soc Am* 46, 1581–1582.

Dixon, WJ, Mood, AM. 1948. A method for obtaining and analyzing sensitivity data. *J Am Stat Assn* 43, 109–126.

Dubno, JR, Dirks, DD, Morgan, DE. 1984. Effects of age and mild hearing loss on speech recognition in noise. *J Acoust Soc Am* 76, 87–96.

Duquesnoy, AJ. 1983. Effect of a single interfering noise or speech sound upon the binaural sentence intelligibility of aged persons, *J Acoust Soc Am* 74, 739–743.

Emmerson, PL. 1984. Observations on a maximum likelihood method of sequential threshold estimation and a simplified approximation. *Percept Psychophys* 36, 199–203.

Findlay, JM. 1978. Estimates on probability functions: A more virulent PEST. *Percept Psychophys* 23, 181–185.

Gelfand, SA, Ross, L, Miller, S. 1988. Sentence reception in noise from one versus two sources: Effects of aging and hearing loss. *J Acoust Soc Am* 83, 248–256.

Hall, JL. 1968. Maximum-likelihood sequential procedure for estimation of psychometric functions. *J Acoust Soc Am* 44, 370.

Hall, JL. 1981. Hybrid adaptive procedure for estimation of psychometric functions. *J Acoust Soc Am* 69, 1763–1769.

Hall, JL. 1983. A procedure for detecting variability of psychophysical thresholds. *J Acoust Soc Am* 73, 663–669.

Hawkins, DB, Walden, BE, Montgomery, A, Prosek, RA. 1987. Description and validation of an LDL procedure designed to select SSPL-90. *Ear Hear* 8, 162–169.

Hellman, RP. 1976. Growth of loudness at 1000 and 3000 Hz. *J Acoust Soc Am* 60, 672–679.

Hellman, RP. 1981. Stability of individual loudness functions obtained by magnitude estimation and production. *Percept Psychophys* 29, 63–78.

Hellman, RP, Meiselman, CH. 1988. Prediction of individual loudness exponents from cross-modality matching. *J Speech Hear Res* 31, 605–615.

Hellman, RP, Meiselman, CH. 1993. Rate of loudness growth for pure tones in normal and impaired hearing. *J Acoust Soc Am* 93, 966–975.

Hellman, RP, Zwislocki, JJ. 1961. Some factors affecting the estimation of loudness. *J Acoust Soc Am* 33, 687–694.

Hellman, RP, Zwislocki, J. 1963. Monaural loudness function of a 1000-cps tone and interaural summation. *J Acoust Soc Am* 35, 856–865.

Hellman, R.P, Zwislocki, J. 1968. Loudness summation at low sound frequencies. *J Acoust Soc Am* 43, 60–63.

Hesse, A. 1986. Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility, and efficiency. *Acustica* 59, 263–266.

Jesteadt, W. 1980. An adaptive procedure for subjective judgments. *Percept Psychophys* 2S, 85–88.

Kaernbach, C. 1991. Simple adaptive testing with the weighted up-down method. *Percept Psychophys* 49, 227–229.

King-Smith, PE, Grigsby, SS, Vingrys, AJ, Benes, SC, Supowit, A. 1994. Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Res* 34, 885–912.

Kollmeier, B, Gilkey, RH, Sieben, UK. 1988. Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *J Acoust Soc Am* 83, 1852–1862.

Kontsevich, LL, Tyler, CW. 1999. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Res* 39, 2729–2737.

Leek, MR. 2001. Adaptive procedures in psychophysical research. *Percept Psychophys* 63, 1279–1292.

Levitt, H. 1968. Testing for sequential dependencies. *J Acoust Soc Am* 43, 65–69.

Levitt, H. 1971. Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49, 467–477.

Levitt, H. 1978. Adaptive testing in audiology. *Scand Audiol* Suppl 6, 241–291.

Levitt, H, Rabiner, LR. 1967. Use of a sequential strategy in intelligibility testing. *J Acoust Soc Am* 42, 609–612.

Madigan, R, Williams, D. 1987. Maximum-likelihood procedures in two alternative forced-choice: evaluation and recommendations. *Percept Psychophys* 42, 240–249.

Marshall, L, Jesteadt, W. 1986. Comparison of pure-tone audibility thresholds obtained with audiological and two-interval forced-choice procedures. *J Speech Hear Res* 29, 82–91.

Marvit, P, Florentine, M, Buus, S. 2003. A comparison of psychophysical procedures for level-discrimination thresholds. *J Acoust Soc Am* 113, 3348–3361.

Mellers, BA. 1983a. Evidence against "absolute" scaling. *Percept Psychophys* 33, 523–526.

Mellers, BA. 1983b. Reply to Zwislocki's views on "absolute" scaling. *Percept Psychophys* 34, 405–408.

Pentland, A. 1980. Maximum likelihood estimation: The best PEST. *Percept Psychophys* 28, 377–379.

Plomp, R, Mimpen, AM. 1979. Speech-reception threshold for sentences as a function of age and noise. *J Acoust Soc Am* 66, 1333–1342.

Remus, JJ, Collins, LM. 2008. Comparison of adaptive psychometric procedures motivated by the Theory of Optimal Experiments: Simulated and experimental results. *J Acoust Soc Am* 123, 315–326.

Robinson, DE, Watson, CS. 1973. Psychophysical methods in modern psychoacoustics. In: JV Tobias (ed.), *Foundations of Modern Auditory Theory, Vol. 2.* New York, NY: Academic Press, 99–131.

Rowan, D, Hinton, K, Mackenzie, E. 2006. Comparison of Levitt- and Zwislocki-type adaptive procedures for stimulus placement in human listeners. *J Acoust Soc Am* 119, 3538–3541.

Shelton, BR, Scarrow, I. 1984. Two-alternative versus three-alternative procedures for threshold estimation. *Percept Psychophys* 35, 385–392.

Shelton, BR, Picardi, MC, Green, DM. 1983. Comparison of three adaptive psychophysical procedures. *J Acoust Soc Am* 71, 1527–1532.

Sherlock, P, Formby, C. 2005. Estimates of loudness, Loudness discomfort, and the auditory dynamic range: Normative estimates, comparison of procedures, and test-retest reliability. *J Am Acad Audiol* 16, 85–100.

Simpson, WA. 1988. The method of constant stimuli is efficient. *Percept Psychophys* 44, 433–436.

Simpson, WA. 1989. The step method: A new adaptive psychophysical procedure. *Percept Psychophys* 45, 572–576.

Stevens, SS. 1936. A scale for the measurement of a psychological magnitude: Loudness. *Psychol Rev* 43, 405–416.

Stevens, SS. 1951. Mathematics, measurement, and psychophysics. In: SS Stevens (ed.), *Handbook of Experimental Psychology*. New York, NY: Wiley.

Stevens, SS. 1958. Problems and methods in psychophysics. *Psychol Bull* 55, 177–196.

Stevens, SS. 1956. The direct estimation of sensory magnitudes—loudness. *Am J Psychol* 69, 1–25.

Stevens, SS. 1961. The psychophysics of sensory function. In: WA Rosenblith (ed.), *Sensory Communication.* Cambridge, MA: MIT Press.

Stevens, SS. 1975. *Psychophysics*. New York, NY: Wiley.

Stevens, SS, Guirao, M. 1963. Subjective scaling of length and area and the matching of length to loudness and brightness. *J Exp Psychol* 66, 177–186.

Stevens, JC, Marks, LM. 1965. Cross-modality matching of brightness and loudness. *Proc Natl Acad Sci U S A* 54, 407–411.

Stevens, JC, Marks, LM. 1980. Cross-modality matching functions generated by magnitude estimation. *Precept Psychophys* 27, 379–389.

Stevens, SS, Volkmann, J. 1940. The relation of pitch to frequency: A revised scale. *Am J Psychol* 53, 329–353.

Stevens, SS, Volkmann, J, Newman, EB. 1937. A scale for the measurement of the psychological magnitude pitch. *J Acoust Soc Am* 8, 185–190.

Taylor, MM, Creelman, CD. 1967. PEST: Efficient estimates on probability functions. *J Acoust Soc Am* 41, 782–787.

Taylor, MM, Forbes, SM, Creelman, CD. 1983. PEST reduces bias in forced choice psychophysics. *J Acoust Soc Am* 74, 1367–1374.

Watson, AB, Pelli, DG. 1983. QUEST: A Bayesian adaptive psychometric method. *Percept Psychophys* 33, 113–120.

Wetherill, GB. 1963. Sequential estimation of quantal responses. *J R Stat Soc* 25, 1–48.

Wetherill, GB, Levitt, H. 1965. Sequential estimation of points on a psychometric function. *Br J Math Stat Psychol* 18, 1–10.

Zwislocki, JJ. 1983a. Group and individual relations between sensation magnitudes and their numerical estimates. *Percept Psychophys* 33, 460–468.

Zwislocki, JJ. 1983b. Absolute and other scales: The question of validity views on "absolute" scaling. *Percept Psychophys* 33, 593–594.

Zwislocki, JJ, Goodman, DA. 1980. Absolute scaling of sensory magnitudes: A validation. *Percept Psychophys* 28, 28–38.

8 Theory of Signal Detection

The previous chapter addressed itself to the classical and modern psychoacoustical methods and the direct scaling of sensory magnitudes with respect to hearing. It left essentially unresolved, however, the problem of how to effectively separate sensitivity from response proclivity. In this chapter, we shall approach this problem from the standpoint of the theory of signal detection.

FACTORS AFFECTING RESPONSES

The theory of signal detection (Swets, 1965; Greene and Swets, 1974; Egan, 1975) provides the best approach to separate the effects of sensitivity from those of response bias. We might think of the **theory of signal detection (TSD)** as asking the question, "what led to a "yes" (or "no") decision?" as opposed to "what did the subject hear (or not hear)?"

Suppose a subject were asked to say "yes" when he hears a tone during a test trial and "no" when a tone is not heard. A large number of trials are used for each of several stimulus levels, and half of those at each level are "catch trials" during which signals are not actually presented. There are thus four possible outcomes for each test trial. Two of them are correct:

1. A **hit** occurs when the signal is present and the subject says "yes."
2. A **correct rejection** occurs when the signal is absent and the subject says "no." The other two alternatives are wrong:
3. The signal is present but the subject says "no." This is called a **miss**.
4. The signal is absent but the subject says "yes". Here a **false alarm** has occurred.

A convenient way to show these possible stimulus and response combinations is to tabulate them in a stimulus–response matrix, which is illustrated in Fig. 8.1.

The stimulus–response table is generally used to summarize the results of all trials at a particular test level; there would thus be such a table for each stimulus level used in an experiment. For example, Fig. 8.2 shows the results of 100 trials containing a signal and 100 catch trials. The subject responded to 78 of the signal trials (so that the probability of a hit was 0.78), did not respond to 22 signals (the probability of a miss is 0.22), said "yes" for 17 out of 100 catch trials (the probability of a false alarm is 0.17), and said "no" for the remaining absent-stimulus trials (the probability of a correct rejection is 0.83). One is tempted to say that the percent correct at this stimulus level is 78% (the hit rate), but the fact that the subject also responded 17 times when there was no stimulus present tells us that even the 78% correct includes some degree of chance success or guessing. One way to account for this error is to use the proportion of false alarms as an estimate of the overall guessing rate and to correct the hit rate accordingly. The traditional formula to correct the hit rate for chance success is

$$\mathrm{p\,(hit)_{corrected}} = \frac{\mathrm{p\,(hit)} - \mathrm{p\,(false\ alarm)}}{1 - \mathrm{p\,(false\ alarm)}}$$

In other words, the probability p of a hit corrected for chance success is obtained by dividing the difference between the hit rate and the false alarm rate by 1 minus the false alarm rate. [If this seems odd, recall that the total probability of all catch trials is 1.0, so that 1 − p(false alarm) is the same as the probability of a correct rejection.] Thus, for this example:

$$\mathrm{p\,(hit)_{corrected}} = \frac{0.78 - 0.17}{1.0 - 0.17} = \frac{0.61}{0.83} = 0.735$$

The original 78% correct thus falls to 73.5% when we account for the proportion of the "yes" responses due to chance.

Correcting for chance success is surely an improvement over approaches that do not account for guessing, but it still does not really separate the effects of auditory factors (sensitivity) and nonauditory factors. In essence, this process highlights the importance of nonauditory factors in determining the response, because the very fact that the subject said "yes" to catch trials and "no" to stimulus trials indicates that his decision to respond was affected by more than just sensitivity to the stimulus. The theory of signal detection is concerned with the factors that enter into this decision.

Let us, at least for the moment, drop the assumption that there is some clear-cut threshold that separates audibility from inaudibility and replace it with the following assumptions of TSD. First, we assume that there is always some degree of noise present. This may be noise in the environment, instrumentation noise, or noise due to the subject's moving around and fidgeting. Even if all of these noises were miraculously removed, there would still remain the subject's unavoidable physiological noises (heartbeat, pulse, breathing, blood rushing through vessels, stomach gurgles, etc.). Indeed, the noise is itself often presented as part of the experiments. For example, the task may be to detect a tone in the presence of a noise. Since there is always noise, which is by nature random, we also assume that the stimulation of the auditory system varies continuously. Finally, we shall assume that all of the stimulation occurs (or is at least measurable) along a single continuum. In other words, the subject must decide whether the stimulation in the auditory system (e.g., energy) is due to **noise alone (N)** or to **signal-plus-noise (SN)**. This process may be represented by distributions along a **decision axis** like the one in Fig. 8.3. Here, the abscissa may be conceived of as representing the energy contained in the noise and in the noise plus signal. The x-axis may also be conceived of as representing the *magnitude of sensory activation* resulting from such stimulation. The ordinate denotes the probability of an event occurring. Hence, the N distribution shows the

STIMULUS	RESPONSE: Yes	RESPONSE: No
Present	Hit	Miss
Absent	False Alarm	Correct Rejection

Figure 8.1 Stimulus–response matrix or table showing the four possible outcomes for any given test trial. Correct responses may be "hits" or "correct rejections," whereas errors may also be of two possible types, "misses" or "false alarms."

probability of occurrence of a noise alone as a function of x, and the SN curve shows the chances of a signal-plus-noise as a function of x. The convention is to use the term "probability density" (as opposed to "probability") for the y-axis in order to reflect the fact that values of x change continuously rather than in discrete steps. The subject's response is a decision between "yes" ("I hear the signal as well as the noise") and "no" ("I hear the noise alone").

The N and SN distributions in Fig. 8.3 show the probability functions of noise alone (N) and signal-plus-noise (SN). We might think of these curves as showing the chances (or likelihood) of there being, respectively, a noise alone or a signal-plus-noise during a particular test trial. Obviously, there must always be more energy in SN than in N, due to the presence of the signal. The separation between the N and SN curves thus becomes a measure of sensitivity. This is an unbiased measure because the separation between the two curves is not affected by the subject's criteria for responding (biases). The separation is determined solely by the energy in the signals and the sensitivity of the auditory system. This separation is measured in terms of a parameter called **d prime (d′)**. The value of d′ is equal to the difference between the means ($\bar{x}$) and the N and SN distributions divided by their standard deviation (σ):

$$d' = \frac{\bar{x}_{SN} - \bar{x}_{N}}{\sigma}$$

Comparing Figs. 8.3 8.3a and 8.3b, we see that the greater the separation between N and SN distributions, the larger the value of d′. This value does not change even when different experimental methods are used (Swets, 1959).

STIMULUS	RESPONSE: Yes	RESPONSE: No	
Present	0.78	0.22	1.00
Absent	0.17	0.83	1.00

Figure 8.2 Hypothetical results in the form of proportions for 100 test trials actually containing stimuli and 100 test trials actually without stimuli ("catch trials").

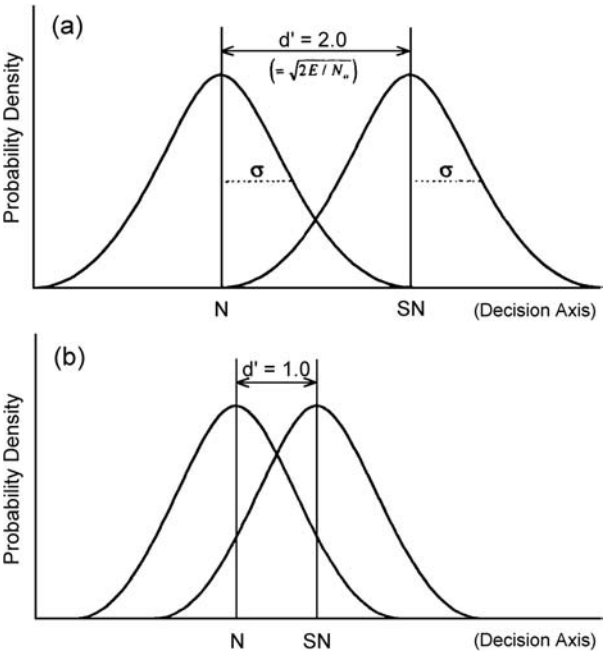

Figure 8.3 The separation between the distribution for the noise alone (N) and the distribution for the signal-plus-noise (SN) determines the value of d′.

Several points will be of interest to the quantitatively oriented reader. It is assumed that SN and N are normally distributed with equal variances. Since σ is the square root of the variance, and the variances of SN and N are assumed to be equal, then only one value of σ need be shown. The value of d′ is equal to the square root of twice the energy in the signal (2E) divided by the noise power (N_O) in a band that is one cycle wide (Swets, Tanner, and Birdsall, 1961), or

$$d' = \sqrt{\frac{2E}{N_O}}$$

Tables of d′ are available in the literature (Elliot, 1964); however, a corrected value of d′ may be a more valid measure because the standard deviation of SN is actually larger than that of N in some cases (Theodore, 1972).

How, then, does a subject decide whether to say "yes" or "no" for a given separation between the N and SN curves? Consider the N and SN distributions in Fig. 8.4. A vertical line has been drawn through the overlapping N and SN distribution in each frame of this figure. This line represents the subject's **criterion** for responding. Whenever the energy is greater than that corresponding to the criterion the subject will say "yes." This occurs to the right of the criterion along the x-axis. On the other hand, the subject will say "no" if the energy is less than (to the left of) the criterion value. The value (or placement) of this criterion depends on several factors, which we will examine next.

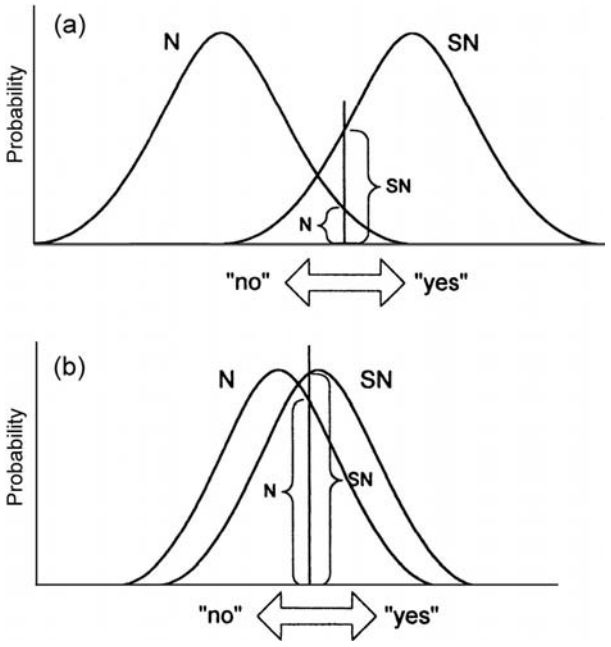

Figure 8.4 Criterion points (shown by vertical lines) for two degrees of overlapping of the noise alone (N) and signal-plus-noise (SN) distributions. The probabilities corresponding to the SN and N distributions at the criterion point are highlighted by brackets. Values of x below (to the left of) the criterion result in "no" decisions and those greater than (to the right of) the criterion yield "yes" decisions.

The first factor affecting the criterion may be expressed by the question "What is the probability that there is a noise alone compared to the probability that there is a signal-plus-noise for a given value of x?" For any point along the decision axis, this question is the same as comparing the height of the N curve with the height of the SN curve (Fig. 8.4). Otherwise stated, this value is the ratio of the likelihoods that the observation is from the N versus SN distributions for the two overlapping curves at any value of x. The ratio of these two probabilities is called **beta** (β). The value of the criterion is affected by the amount of overlap between the N and SN distributions, and by what the subject knows about the relative chances of a signal actually being presented.

Comparison of Figs. 8.4a and 8.4b shows how overlapping of the N and SN functions affects this ratio. At any point, the heights of the two curves becomes close as the separation between them decreases from that in Fig. 8.4a to that in Fig. 8.4b. An **ideal observer**, which is actually a mathematical concept rather than a real individual, would place the criterion point at the ratio which minimizes the chances of error, that is, at the point at which misses and false alarms are minimized. However, the placement of the criterion point will also be adjusted somewhat by what the subject knows about the chances of occurrence of a noise alone versus a signal-plus-noise. Let us now address ourselves to this factor.

Up to this point, it has been assumed that N and SN will be presented on a fifty-fifty basis. However, if the subject knows that a signal will actually be presented one-third of the time, then he will of course adjust his criterion β accordingly. In other words, he will adopt a stricter criterion. Alternatively, if the subject knows that a signal will occur more often than the noise alone, then he will relax his criterion for responding, adjusting for the greater chances of the signal actually being presented. The theoretical ideal observer always knows these probabilities; a real subject is often, but not always, told what they are.

The last factor that we will discuss which affects the final value of the criterion β has to do with how much a correct response is worth and how much a wrong response will cost. We are therefore concerned with the chance of an error associated with a particular criterion for responding. These chances are shown in Fig. 8.5. The subject will say "no" whenever the actual presentation falls to the left of the criterion, and will say "yes" when the presentation is to the right of the criterion. As a result of the fact that the N and SN distributions are overlapping, it turns out that there will be both "yes" and "no" decisions for a certain proportion of *both* signal and no-signal presentations. With the criterion placed as shown in the figure, most of the "yes" decisions will be in response to actual SN presentations; that is, the subject will say "yes" when there actually was a signal present. Recall that such a correct identification of the presence of the signal is called a *hit*. On the other hand, a certain percentage of the N trials will fall to the right of the criterion, so the subject will say "yes" even though there was actually no signal presented. This incorrect decision that a signal was present even though it really was not there is a *false alarm*. A stimulus–response table similar to the one in Fig. 8.1a is shown next to the N and SN distributions in Fig. 8.5 to illustrate how

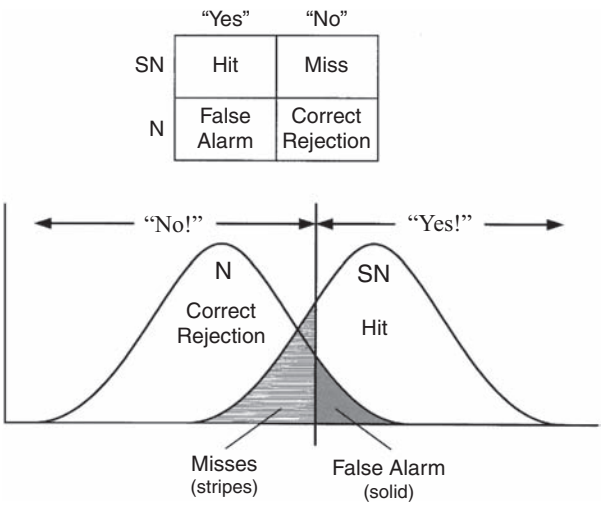

Figure 8.5 The four possible outcomes of a "yes" or "no" response based upon a given criterion value (*vertical line*). The corresponding stimulus–response table is shown to the right.

the two curves and the criterion relate to the possible outcomes of an experiment.

Now, suppose that a subject is told that it is imperative that he never miss a signal. He would thus move the criterion point toward the left to increase the hit rate; however, this shift would also have the effect of increasing the number of false alarms. This result would occur because moving the criterion toward the left increases the proportions of both the N and SN curves that fall inside of the "yes" region. On the other hand, suppose that the subject were advised that a false alarm is the worst possible error. Under such circumstances, the criterion point would be shifted toward the right, minimizing false alarms. Of course, this shift would also increase the number of misses, because a larger portion of the SN curve would now be in the "no" region.

Instead of telling the subject that one or another type of response is more (or less) important, the subject might be given a nickel for each correct response, lose three cents for a false alarm, etc. This, too, would cause the subject to adjust the criterion point so as to maximize the payoff associated with his responses. In effect, then, a set of values is attached to the responses so that each correct response has a **value** and each erroneous response has a **cost**.

An optimum criterion point (**optimum** β) is based upon the probabilities of the noise alone (p_N) and of the signal-plus-noise (p_{SN}) combined with the payoff resulting from the costs and values of each response. The *payoff* is the net result of the values of hits (V_H) and correct rejections (V_{CR}) and of the costs of misses (C_M) and false alarms (C_{FA}). In other words,

$$\text{optimum } \beta = \left(\frac{p_N}{p_{SN}}\right)\left(\frac{V_{CR} - C_{FA}}{V_H - C_M}\right)$$

The decision criterion is an attempt to maximize the payoff associated with the task. However, the subject in the real world is either not aware of all factors, or not able to use them as efficiently as the mathematically ideal observer. Therefore, the actual performance observed in an experiment generally falls short of what would have resulted had the subject been an ideal observer.

In summary, two types of information are obtained from the subject's responses in a TSD paradigm. One of these, d′, is a measure of *sensitivity*, which is determined strictly by the separation between the noise and signal-plus-noise distributions and by the ability of the auditory system to make use of this separation. The other measure is the subject's *criterion* for responding, which does not affect the actual measure of sensitivity.

How can we show all of this information at the same time in a meaningful manner? Consider the effects of several different response criteria for the same value of d′. These criteria may be obtained by changing the directions given to the subject, or by changing the payoff scheme. Another way would be to have the subject rank the degree of certainty with which he makes each yes/no decision (see the discussion of TSD methods, below, for the rationale of this approach).

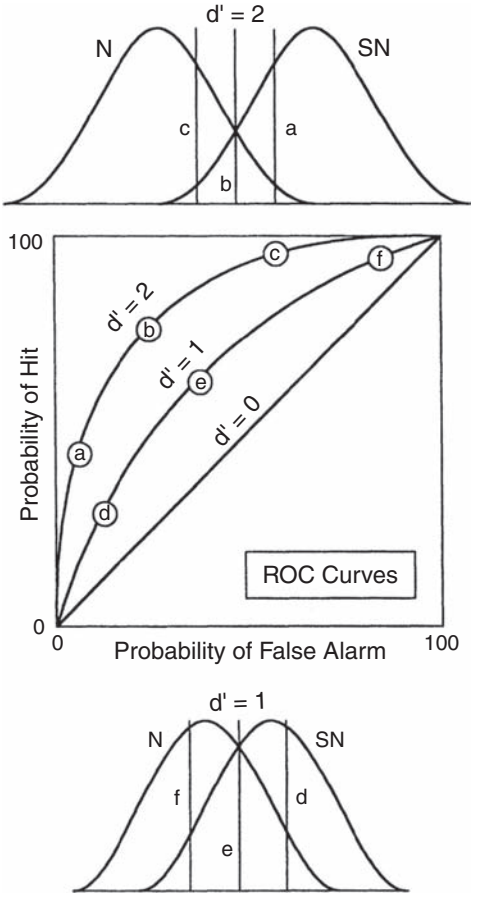

Figure 8.6 Relationships between ROC curves (center of figure) and the noise alone (N) and signal-plus-noise (SN) distributions for two values of d′. Sensitivity is depicted by the distance of the ROC curve from the diagonal. The upper set of distributions shows $d' = 2$, and the lower distributions show $d' = 1$. Various response criteria are indicated by the letters a–f.

For a given amount of sensitivity (i.e., a given value of d′), different criteria will result in different proportions of hits and false alarms. This result is shown for two arbitrarily selected values of d′ in Fig. 8.6. We may plot the proportions of hits versus false alarms for each criterion point, as in the center of Fig. 8.6. Such a graph is called a **receiver-operating characteristic** or **ROC curve**. Notice that the ROC curve allows both the effects of sensitivity and response criterion to be illustrated at the same time. Sensitivity is shown by the distance of the ROC curve from the diagonal (at which $d' = 0$), or by the area under the ROC curve. On the other hand, the response criterion is indicated by the particular point along the ROC curve. Specifically, points a, b, and c in the figure (where $d' = 2$) differ in terms of sensitivity from points d, e, and f (for which $d' = 1$). However, even though points a, b, and c are the same in terms of sensitivity ($d' = 2$), they differ from each other in terms of response criteria. A similar relationship exists among points d, e, and f, where $d' = 1$.

PSYCHOPHYSICAL METHODS IN TSD

Yes/No Methods

This discussion of the theory of signal detection has dealt primarily with the yes/no method. To recapitulate: The subject is presented with a large number of trials for each stimulus level, and a large proportion of these are actually catch trials. For each trial, the subject says "yes" when a signal is detected and "no" when a signal is not detected. Fundamentally, then, the yes/no method in TSD is somewhat akin to the classical method of constant stimuli, although there are obvious differences in the number of trials, the large proportion of catch trials, and the manner of data analysis.

As in the classical methods, the TSD experiment is easily modified for use in a study of differential sensitivity. In this case, two signals are presented in a trial and the subject's task is to say "yes" ("they are different") or "no" ("they are not different").

The yes/no method is actually a subcategory of a larger class of experiments in which each trial contains one or two alternative signals. In this general case, the subject's task is to indicate which of two possible signals was present during the trial. For example, in the differential sensitivity experiment mentioned in the last paragraph, the decision is "same" versus "different", or else the subject might be asked to decide between two alternative speech sounds (e.g., /p/ and /b/) while some parameter is varied. In this light, the yes/no method might be thought of as a **single-interval forced-choice** experiment. In other words, the subject is presented with a stimulus interval during each trial and is required to choose between signal-plus-noise (one of the alternatives) and noise alone (the other alternative).

Two-Interval and N-Interval Forced-Choice Methods

Just as the subject may be asked to choose between two alternatives in a single-interval trial, he might also be asked to decide which of two successive intervals contained a signal. This approach is called the **two-interval forced-choice (2IFC)** or **two-alternative forced-choice (2AFC)** method. In this method, a test trial consists of two intervals, A and B, presented one after the other. One of the intervals (SN) contains the signal and the other one (N) does not. The subject must indicate whether the signal was presented in interval A or in interval B.

Experiments involving a choice between more than two choices are termed **multiple** or **N-interval** (or **alternative**) **forced-choice**, where N refers to the number of choices. For example, a 4IFC task would include four intervals (alternatives) in each test trial, among which the subject must choose the one that contained the signal.

Confidence Rating Methods

Recall that various points along the same ROC curve represent different response criteria with the same sensitivity d′. We might think of the response criterion as a reflection of how much confidence a subject has in his decision. In other words, a **strict criterion** means that the subject must have a great deal of confidence in his decision that the signal is present before he is willing to say "yes." In this case, the criterion value 3 is pushed toward the right along the decision axis. Alternatively, a **lax criterion** means that the subject does not require as much confidence in his "yes" decision, which moves the criterion point toward the left.

We might apply this relationship between the confidence in the decision and the criterion point by asking the subject to rate how much confidence he has in each of his responses. For example, the subject might be instructed to rate a "yes" response as "five" when he is absolutely positive that there was a signal, and "four" when he thinks there was a signal. A rating of "three" would mean "I'm not sure whether there was a signal or no signal." "Two" would indicate that there probably was no signal present, and a rating of "one" would suggest that the subject is positive that a signal was not presented. This procedure is the same as adopting a series of criterion points located successively from right to left along the decision axis. Thus, the use of **confidence ratings** enables the experimenter to obtain several points along the ROC curve simultaneously. This approach results in data comparable to those obtained by the previously discussed methods (Egan, Schulman, and Greenberg, 1959).

SOME IMPLICATIONS OF TSD

The theory of signal detection has importance in psychoacoustics because its application allows the experimenter to ferret out the effects of sensitivity and response criterion. Furthermore, TSD lends itself to experimental confirmation and can be used to test theories and their underlying assumptions. A key application of TSD has been the testing of the classical concept of threshold as an absolute boundary separating sensation from no sensation. It is implicit in this discussion that such a concept of a clear-cut threshold is not supported by TSD. However, the more general concept of threshold remains unresolved. Threshold theory is beyond the scope of this text. The interested student is therefore referred to the very informative discussions that may be found in the papers by Swets (1961) and by Krantz (1969).

REFERENCES

Egan, JP. 1975. Signal Detection *Theory and ROC Analysis*. New York, NY: Academic Press.

Egan, JP, Schulman, AI, Greenberg, GZ. 1959. Operating characteristics determined by binary decisions and by ratings. *J Acoust Soc Am* 31, 768–773.

Elliot, PB. 1964. Tables of d'. In: JA Swets (ed.), *Signal Detection and Recognition by Human Observers.* New York, NY: Wiley, 651—684.

Greene, DM, Swets, JA. 1974. *Signal Detection Theory and Psychophysics*. New York: Krieger.

Krantz, DH. 1969. Threshold theories of signal detection. *Psychol Rev* 76, 308–324.

Swets, JA. 1959. Indices of signal delectability obtained with various psychophysical procedures. *J Acoust Soc Am* 31, 511–513.

Swets, JA. 1961. Is there a sensory threshold? *Science* 134, 168–177.

Swets, JA (ed.). 1965. *Signal Detection and Recognition by Human Observers.* New York, NY: Wiley.

Swets, JA, Tanner, WP Jr, Birdsall, TG. 1961. Decision processes in perception. *Psychol Rev* 68, 301–340.

Theodore, LH. 1972. A neglected parameter: Some comments on "A table for calculation of d′ and β". *Psychol Bull* 78 260–261.