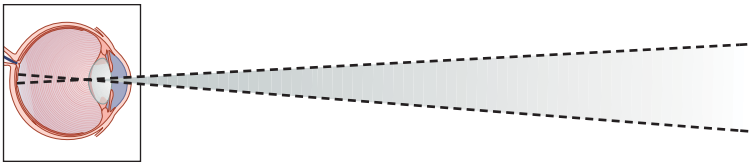

11

Psychophysical Methodology

Many of the routine procedures used in eye care, including the determination of a patient's visual acuity, refractive status, visual fields, and color vision status, are psychophysical in nature. Knowledge of psychophysical theory and methodology better enables the clinician to properly perform these procedures and interpret their results in a meaningful manner.

The past several decades have seen substantial progress in our understanding of basic visual processes, and these insights are increasingly being applied in routine and advanced eye care. Clinical applications largely take the form of noninvasive psychophysical tests that enable the clinician to diagnose disease at an early stage and to monitor the effectiveness of treatment.

THRESHOLD

Psychophysical experiments and psychophysically based clinical procedures frequently involve the determination of a threshold, the minimum quantity of a stimulus that can be detected. For example, in a visual acuity test, the threshold is the minimum angle of resolution (MAR). For a visual field test, the threshold is the minimum light intensity that can be detected (Fig. 11–1A).[1]

The determination of a threshold is complicated because humans (and other animals) are not perfect observers. A perfect observer would give the same threshold each time it is measured. In practice, threshold varies on repetition of its measurement.

1. A distinction can be made between psychophysics and perception. In psychophysical experiments, a physical aspect of a stimulus is adjusted until a threshold is reached. For instance, the intensity of a stimulus is increased until the observer detects it. Alternatively, the observer could be asked to report when a stimulus appears colored, flickers, moves, etc. Perceptual experiments, in comparison, may require the observer to describe what he or she perceives. An observer could be presented with a complex scene, and asked to describe certain aspects of that scene. Both psychophysical and perceptual experiments utilize quantitative measures.

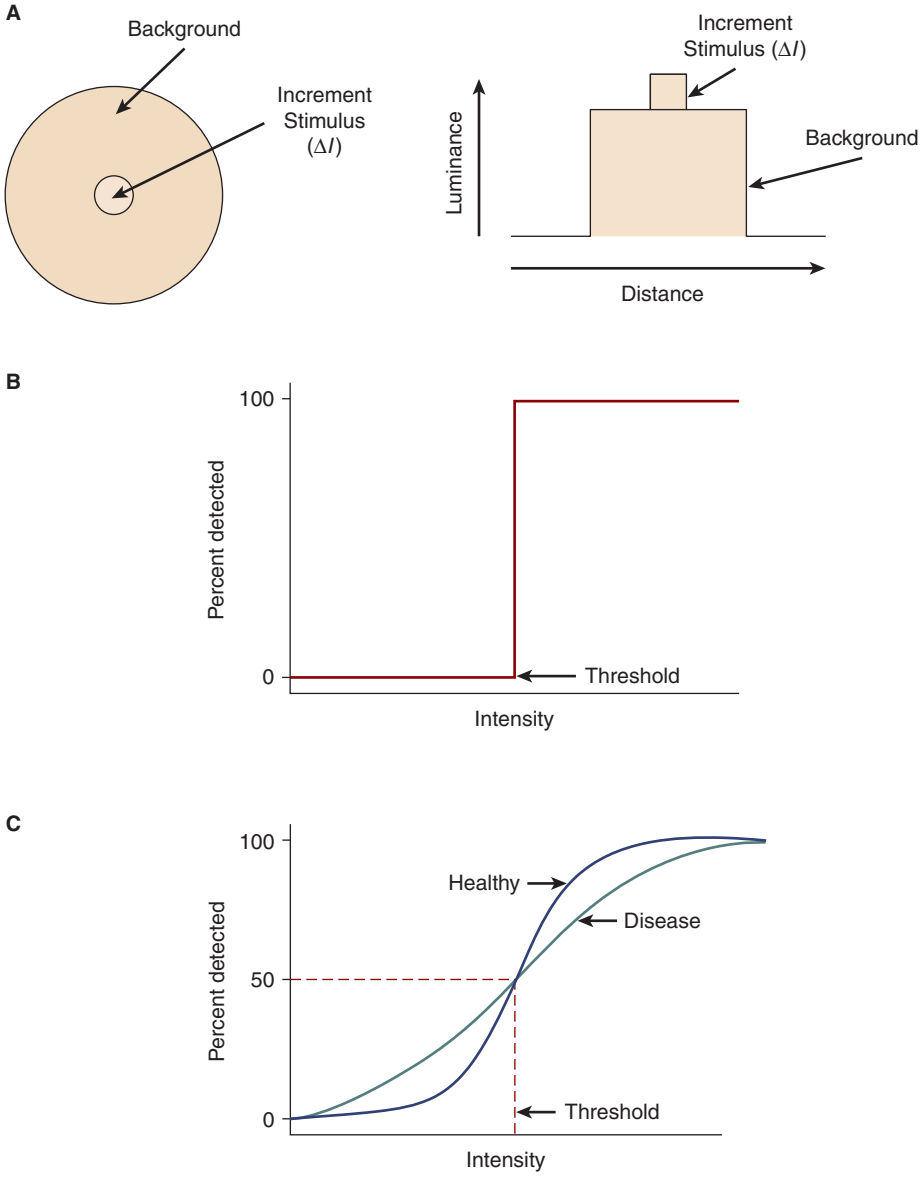

Figure 11–1. **A.** The diagram on the left shows an increment stimulus on a background. This is how the stimulus appears to the observer. The diagram on the right is a luminance profile of the stimulus and background. **B.** A FOS curve for a hypothetical ideal observer. Such observers do not actually exist. **C.** FOS curves for a healthy observer and one with ocular disease. Note that the latter is flatter, making the determination of a threshold more difficult.

Figure 11–1 shows the results of a psychophysical experiment conducted with a hypothetical ideal (perfect) observer and real observer. It involves the determination of an increment threshold, such as when performing visual field testing. The observer is required to detect a light (stimulus) that is flashed on a surrounding background. The task is repeated for a range of stimulus intensities, from dim to intense, and the percentage of stimuli detected is plotted as a function of stimulus intensity to produce a **frequency of seeing (FOS) curve,** also referred to as a **psychometric function.**

An ideal observer manifests an unambiguous threshold (see Fig. 11–1B). Below the threshold intensity, she never sees the stimulus, and above this intensity, she always sees the stimulus. In comparison, a real observer produces results similar to those in Fig. 11–1C. As the intensity of the stimulus is increased, the probability of seeing the stimulus increases. There is no clearly defined intensity, however, below which the stimulus is never seen and above which it is always seen.

Since there are no perfect observers, threshold is based on theoretical considerations. It is usually defined as the intensity that results in detection of the stimulus on one-half of the presentations. This value is read off the FOS curve, as indicated in Fig. 11–1C.

Humans are not perfect observers because they are complex biological systems, not simple mechanical devices. A stimulus results in neural activity. If this neural activity is sufficiently strong, the stimulus is seen. Random neural noise is inherent within the visual system, however, and the signal produced by the stimulus must be perceived as different than this neural noise. As discussed later in this chapter, neural noise can be thought of as varying over time. At any given moment, the amount of neural noise is unpredictable. Therefore, the threshold is variable. Attention, motivation, and fatigue can also affect threshold.

Clinical Highlight

A diseased visual system is noisier than a healthy one, making the FOS curve less steep (more flat). Examining Fig. 11–1C, we see that it is more difficult to accurately ascertain a threshold for a flat curve because there is a broader range of values that could correspond to 50% detection. This complicates the measurement of visual fields in diseased eyes. Moreover, the visual system is noisier peripherally than centrally, making the assessment of peripheral visual fields in the diseased eye yet more of a challenge.

DETERMINATION OF THRESHOLD

The scientist or clinician may choose among several methods to measure a threshold, with the most suitable method determined by the nature of the experiment or clinical procedure. In the following discussion, we introduce the primary methods of threshold determination and discuss some of their advantages and disadvantages.

Method of Ascending Limits

Consider an increment threshold procedure. In the method of ascending limits, the stimulus is initially below threshold. It is not visible. During a trial, which consists of a number of stimulus presentations, the stimulus intensity is increased systematically until the observer reports that it is visible. Several trials may be performed. The results are averaged to obtain a threshold.

The method of ascending limits is particularly advantageous in dark adaptometry, where it is important that the state of retinal adaptation be minimally affected by the stimulus (see Chapter 3).[2] A potential disadvantage of ascending limits, which may lead to an inaccurate result, is observer anticipation. If the stimulus starts at the same intensity on each trial, the observer may, in an effort to be consistent, anticipate when he or she "should" see the stimulus based on when he or she saw it on the previous trial. Beginning each trial at a different intensity can mitigate this disadvantage.

Method of Descending Limits

The method of descending limits is essentially the reverse of the method of ascending limits. A trial commences with a clearly visible stimulus (i.e., the stimulus is above threshold)[3] and the visibility is decreased systematically until it can no longer be seen. In the method of ascending limits, the stimulus is initially not visible and becomes visible; in the method of descending limits, the reverse occurs.

Descending limits is commonly used to determine visual acuity. The patient is asked to read down the Snellen eye chart, which consists of optotypes that become progressively smaller from top to bottom (and more difficult to resolve). The threshold MAR is determined when the optotypes are too small to be resolved (see Chapter 7).

When using descending limits, the initial stimulus presentations are visible, serving to familiarize the patient with the task. Consider the clinical determination of visual acuity. The patient obtains practice by reading the large optotypes at the top of the chart, thereby increasing the clinician's confidence that the patient understands the task.

Similar to the method of ascending limits, the method of descending limits may be contaminated by observer anticipation. In an increment threshold experiment where each trial commences at the same intensity, the observer may anticipate when he or she "should" no longer see the stimulus. This can be addressed by starting each trial at a different intensity.

2. Consider the determination of one point on a dark adaptation curve (see Fig. 3–10). If this threshold were obtained by the method of descending limits (our next topic), the initial stimulus presentations might be intense, bleaching substantial amounts of photopigment. This could lead to an inaccurate threshold measurement—the measured threshold would be higher than the actual threshold.
3. A stimulus that is above threshold is sometimes referred to as suprathreshold stimulus.

Clinical Highlight

The manner in which a clinical psychophysical test is performed can influence the result. Suppose you determine the visual acuity of a patient who is worried he may be losing his vision, and who wishes to be sure that you take his symptoms seriously. If asked to read down an eye chart—method of descending limits—it is possible that the patient may stop reading letters when he notices blur and has uncertainty regarding what he is seeing. In this case, where the patient has been allowed to determine the threshold criterion, the measured visual acuity could be, say, 20/30.

How would the result be different if you set the criterion by instructing the patient to continue to read down the chart even if he says he cannot clearly see the letters? You encourage the patient to guess at what he sees, and to not worry about being wrong. Using this forced choice methodology—forced choice because the patient must guess at the letters—it would not be surprising to find the acuity to be 20/20 rather than 20/30.

Visual acuity should be measured in a standardized manner that ensures that the practitioner, not the patient, sets the threshold criterion. Not only will different patients have different criteria (some will be adventurous guessers and others will be cautious nonguessers), the criterion that a given patient uses may vary from visit to visit depending on his mood or other factors. Since treatment decisions are often made based on visual acuity and/or changes in visual acuity over time, it is important to know that our measurements are reliable and not unnecessarily confounded by issues of threshold criteria.

Given these considerations, what termination rules should be employed when measuring visual acuity? Asking the patient to continue to guess at optotypes once it is obvious that they are not visible can be frustrating for the patient and time consuming. Carkeet (2001) suggests that when using charts with a Bailey-Lovie design (Chapter 7), the test be terminated after the patient makes four or more mistakes on a line.

These same considerations hold for other psychophysical tests that could be improperly administered by allowing the patient to set the threshold criteria. In stereopsis testing, where the patient selects which of the images is elevated, an answer of "I can't tell" should prompt in the practitioner to ask the patient to guess which image is elevated.

Staircase Method

The staircase method of threshold determination is a combination of ascending and descending limits. Suppose a stimulus is presented in discrete steps of increasing visibility called an ascending staircase (Fig. 11–2). Eventually, the observer reports seeing the stimulus. At this point, the staircase is reversed, and the visibility of the stimulus is reduced until the observer reports that it cannot be detected (descending staircase). The staircase is again reversed, and the stimulus intensity is increased until it is visible. Threshold is taken to be the stimulus intensity at one of the reversals, for example, the fourth reversal.

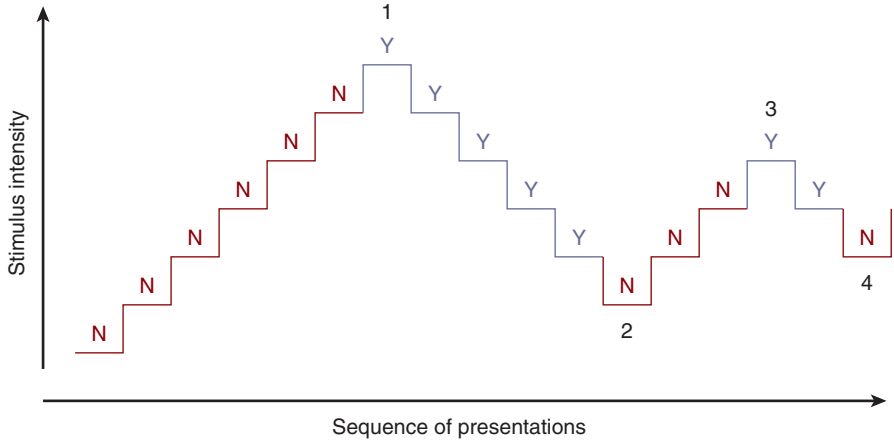

Figure 11–2. Staircase method for threshold determination. The stimulus intensity is increased from nonseeing (N) to seeing (Y). A reversal occurs at point 1 and, subsequently, the intensity is decreased until another reversal occurs at point 2. Threshold could be taken as the intensity at, for example, reversal 4.

Clinical Highlight

This strategy provides a quick and reliable method of determining a threshold. It is commonly used in psychophysical experimentation and automated visual field testing.

Method of Constant Stimuli

In the method of constant stimuli, the stimulus visibility is varied randomly from presentation to presentation. Because the observer is typically asked whether or not he or she sees the stimulus, this method is sometimes referred to a "yes–no" procedure. Blank trials, in which no stimulus is presented, are typically included. The number of times that the observer reports seeing the stimulus during a blank trail (false positive or false alarm responses) is, as we shall learn, important for the analysis of the data.

Consider the determination of an increment threshold. Suppose there are 220 trials, with the stimulus presented numerous times at each of 10 intensities. Twenty of the trials are blanks. Stimuli are presented in random order with the blanks randomly interspersed. On each presentation, the observer is asked whether or not he or she sees the stimulus. A FOS curve is plotted and 50% visibility is taken as the threshold.

Let us take this a step further. Suppose the subject reports seeing a stimulus on 145 trials. The stimulus is present, but not seen (misses) on 60 trials. How many times was the stimulus present and seen (hits), falsely seen in a blank trial (false positives or alarms), and not seen during a blank presentation (correct rejects)? Figure 11–3 shows how these values can be calculated by completing a chart that contains cells for hits, misses, false positives, and correct rejects.

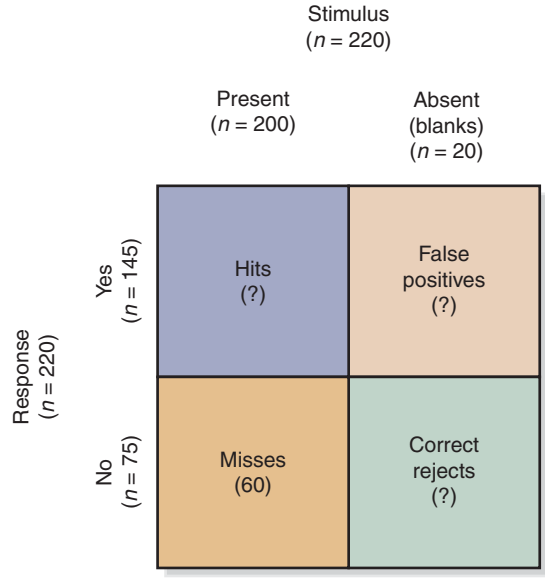

Figure 11–3. Outcome matrix for the yes–no experiment discussed in the text. Using this matrix, the number of hits, false positives, and correct rejects can be calculated to be 140, 5, and 15, respectively.

The method of constant stimuli is so named not because the stimulus is kept constant from presentation to presentation, but because the procedure is designed to maintain the observer's expectations at the same level from presentation to presentation. The observer has no valid basis on which to anticipate the visibility of an upcoming presentation. Because of this, the method of constant stimuli can provide valuable information for laboratory experiments. It is, however, time consuming and not typically indicated for clinical applications.

Errors can occur when using the method of constant stimuli (and other psychophysical methods) with a subject who has a bias. For instance, the subject may have a lax threshold criterion that results in a high false positive rate. Suppose the false positive rate is 20%. How does this affect the FOS curve? It will not start out at 0% detected, as in Fig. 11–1C, but at the false positive rate of 20%. To obtain an accurate threshold, we must compensate for the false positive rate so that the FOS curve runs from 0% to 100% correct.

Now consider a subject with the opposite bias—a strict threshold criterion where the subject says "no" when the stimulus is clearly visible. In this case, the psychometric function does not reach 100% detection, as in Fig. 11–1C, because even at high intensities the stimulus is not always seen.

How do we correct for a strict criterion? We include trials where the stimulus is clearly above threshold (i.e., suprathreshold stimuli). When the subject says she does not see such a stimulus, we call this response a "false negative." The FOS curve is corrected for the false negative rate, allowing threshold to be read off it.

A challenge with correcting the FOS curve is that the subject may not show a consistent bias. At one point, she may use a lax criterion, but later on in the experiment may switch over to a stricter criterion. The use of forced choice methods, which we discuss later in this chapter, may help to get around this.

Method of Adjustment

In the method of adjustment, the subject adjusts the stimulus intensity until it is barely visible (or invisible), allowing for a relatively quick threshold determination. This method may suffer, perhaps more than those previously mentioned, from anticipation and variations in the observer's threshold criterion.

Forced Choice Method

The previously discussed methods of determining a threshold share a common flaw: not all observers use the same criteria when deciding whether or not they see a stimulus. For example, observers with strict threshold criteria do not report seeing a stimulus until they are absolutely certain they see it. This results in a relatively high threshold (low sensitivity). Other observers who have lax criteria report seeing a stimulus even though they may have a great deal of uncertainty regarding their decision. The result is a relatively low threshold (high sensitivity).

Not only may the threshold criteria vary from observer to observer, it may vary from trial to trial for the same observer. At certain times during an experiment, an observer may be more willing to guess that he or she sees a stimulus than at other times. Moreover, an observer may use one set of criteria for one type of stimulus and another set for a different stimulus. For example, the criteria used by an observer to detect stimuli under photopic conditions could be different than those applied under scotopic conditions. These variations in threshold criteria potentially complicate the interpretation of experimental results.

In forced choice methodology, the effects of the observer's criteria are minimized by forcing him or her to choose between several alternative choices, one of which contains the stimulus. In the example shown in Fig. 11–4A, the stimulus is randomly presented in one of the two windows. The other window does not contain a stimulus. It is blank. The observer is forced to choose which window contains the stimulus. Because a response of "I cannot see the stimulus" is not acceptable, the role of the observer's threshold criterion is reduced.

To construct a psychometric function, a large number of trials are presented using stimuli of various visibilities.[4] If the experiment forces the observer to choose between two alternatives, as in the previous example, it is referred to as a two-alternative forced choice (2AFC) experiment. A psychometric function for a 2AFC experiment is given in Fig. 11–4A. Note that the lowest percentage correct is 50%

4. In forced choice experiments, a trial consists of the stimulus and the blank (or blanks).

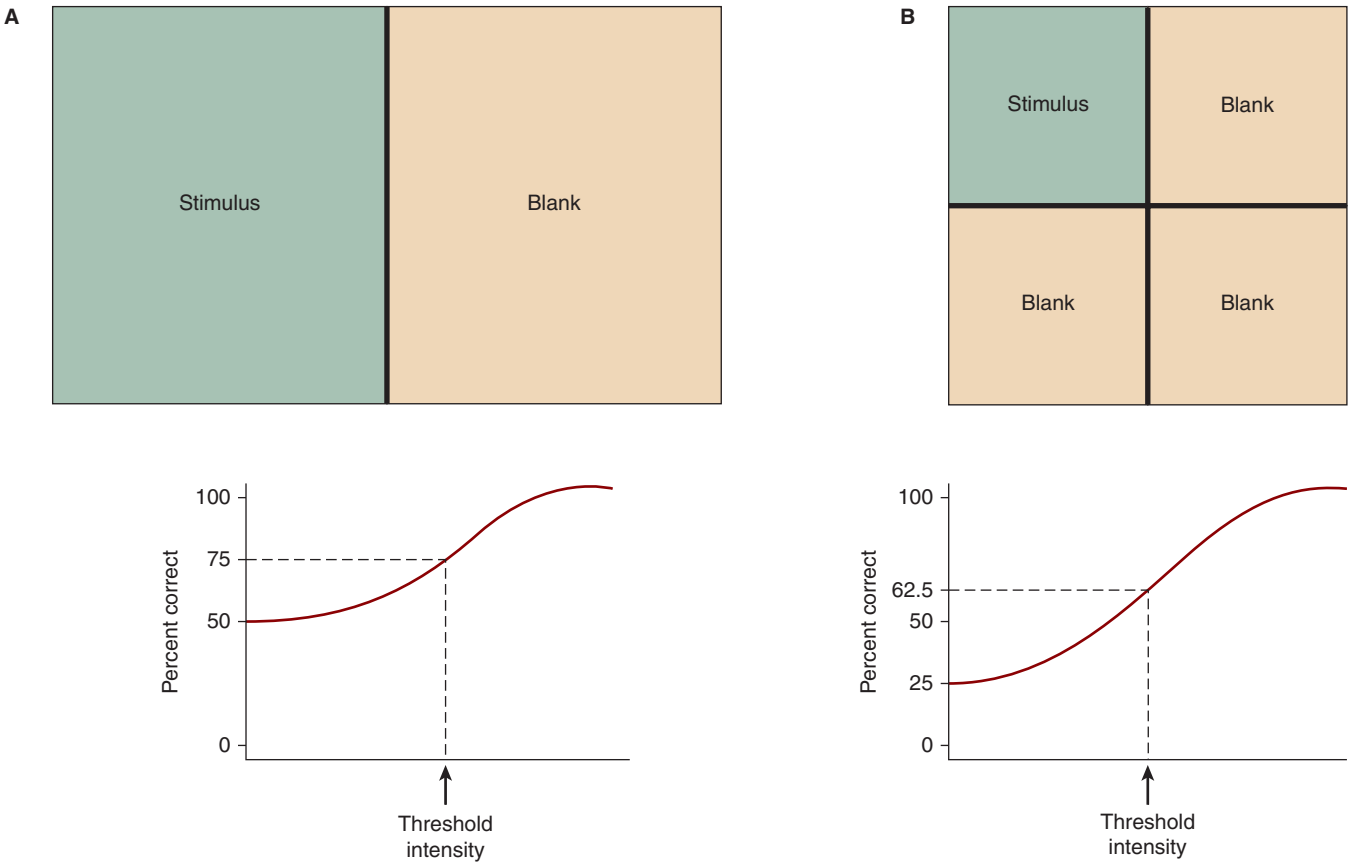

Figure 11–4. Two-alternative forced choice (2AFC) methodology. The top diagram shows the stimulus array, and the bottom figure shows a typical psychometric function for a 2AFC experiment. **B.** Stimulus array and psychometric function for a four-alternative forced choice (4AFC) experiment. Note that this psychometric function is steeper than for the 2AFC experiment.

(chance performance). It is not 0% because even when the observer cannot see the stimulus, he or she is expected to guess correctly 50% of the time.[5] The threshold value is typically taken as the point midway between chance performance and perfect performance. As indicated in Fig. 11–4A, the threshold for a 2AFC experiment is 75%.

Forced choice experiments can present the participant with more than two choices. Consider Fig. 11–4B, which shows a four-alternative forced choice (4AFC) stimulus array in which the stimulus is randomly presented in one of the four windows. The observer is forced to choose the correct window. Note that the psychometric function for this experiment shows chance performance of 25%. Threshold is taken as the point midway between chance performance and 100%, which is 62.5%.

Increasing the number of choices typically increases the complexity of the experiment and causes it to take longer to perform. Is there an advantage to using more than two choices? The answer can be seen by studying Fig. 11–4. Note that the psychometric function is flatter for 2AFC compared to 4AFC. This increases the chance of error when reading the threshold off the graph because there are many points on the flat function that fall close to 75% correct—any noise in the data can make it difficult to choose accurately the stimulus that corresponds to the 75% point. Because the psychometric function for 4AFC is steeper, the threshold can be ascertained with more confidence.

Compared to other psychophysical methods, forced choice typically results in lower thresholds. When observers are forced to guess, they often do remarkably well despite claims that they cannot see the stimulus.

Forced choice preferential viewing techniques have been successfully used to determine the visual acuity and other visual capabilities of infants (see Chapter 17) (Teller et al., 1974). In these procedures, the experimenter observes the infant's gaze as the infant views a 2AFC display (see Fig. 17–13). The experimenter is forced to choose the location of the stimulus based on observation of the infant's eyes. By using a forced choice methodology, the criterion (i.e., strict or lax) used by the experimenter is minimized as a confounding factor.

SIGNAL DETECTION THEORY

The threshold that is determined in an experiment or clinical procedure may be influenced by a number of factors, including decision criteria, attention, motivation, and internal neural noise. Signal detection theory provides a useful model to predict the effects of certain of these factors (Swets et al., 1961).

The theory assumes that within the visual system there is a randomly fluctuating level of background neural activity,—the so-called noise. A stimulus produces a neural signal that is superimposed on this neural noise. The observer's task is to differentiate the signal and noise combination from the background noise alone. An analogy would be listening to white noise (static) and attempting to distinguish a discrete increase in the level of noise (i.e., the signal) from the white noise itself.

5. Because there are two windows and the stimulus must be in one of the two windows, a blindfolded observer would guess correctly on one half of the trials.

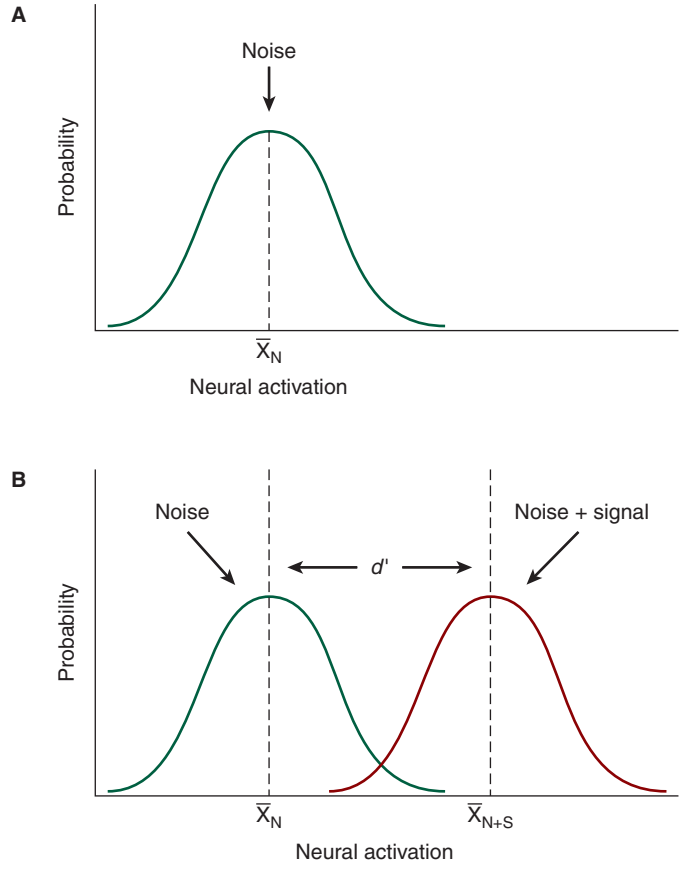

Figure 11–5. **A.** Neural noise is randomly distributed over time. **B.** A stimulus elicits a signal that can be added to the noise distribution to produce a noise + signal distribution. The detectability, *d'*, is a measure of the strength of the stimulus.

A key element of this model is that the neural noise—neural activation that is present in the absence of a stimulus—is randomly distributed over time. Figure 11–5A shows the probability of a given level of neural activation at any instant in time. At some times, there is much noise, and at other times, there is little noise. Returning to the white noise analogy, one can think of the level of static as randomly fluctuating over time.

A stimulus causes a constant level of neural activation (a signal) that is added to a noise distribution (N) to produce a noise plus signal distribution (N + S) (see Fig. 11–5B). It is important to keep in mind that neural noise is present in the absence of the signal, and the signal is superimposed on this noise. The observer's task is to determine if what he or she is seeing (or hearing) is noise or signal plus noise.

The larger the signal, the easier it is for the observer to distinguish the signal plus noise from noise alone (Fig. 11–6A). As the signal becomes larger, the distributions of N and N + S become further apart, and the detectability (d') of the stimulus increases.[6] With a very large d', there is no overlap of the distributions; therefore, there is no uncertainty regarding whether a stimulus is present.

The situation is not so clear-cut when the stimulus is weak, resulting in substantial overlap of the N and N + S distributions (see Fig. 11–6B). If the stimulus is delivered when the noise is low, the resulting level of neural activation (for example, the level indicated by *point 1*) is ambiguous. There is no way for the observer to be certain whether the stimulus is absent or present because this level of neural activation can be produced by either the signal plus noise or noise alone. If, however, the stimulus is delivered at a point in time when the noise is very high, the resulting level of neural activation (for example, the level indicated by *point 2*) is unambiguous. This level of neural noise occurs only when the stimulus is present.

Effect of Observer Criterion

Signal detection theory allows us to predict how the observer's criterion affects stimulus detection. To understand this, consider Fig. 11–7A, which shows the results of a yes–no experiment. The dashed line represents a lax criterion, such as may be adopted by, for example, an intern or resident, who when learning to examine a patient's retina, wishes to be certain that he or she detects any deviation from the norm, no matter how small. Any level of neural activation above the lax criterion line alerts the clinician to a possible abnormality, whereas levels of activation below this line do not elicit a response.

The observer's responses fall into the four categories labeled in Fig. 11–8. According to signal detection theory, if the stimulus results in neural activity that exceeds the threshold criterion, the result is a **hit.** If the activity resulting from the stimulus does not exceed the criterion, there is a **miss.** On those occasions that no stimulus is present (i.e., a blank trial), but the neural activity exceeds the criterion, there is a **false positive** (or **false alarm**). Finally, if the neural noise is below the criterion during a blank trial, the result is a **correct reject.**

Returning to the example of the intern or resident, note that a lax criterion results in very few misses, but many hits and a substantial number of false positives (see Fig. 11–7A). This is the outcome we would expect if a lax criterion were employed when examining the retina. There may be few misses of retinal disease, yet there may be many false alarms.

Now consider the case of a strict criterion, such as that adopted by a deer hunter who wishes to be absolutely certain that the target in his or her sights is indeed a deer (see Fig. 11–7B). Note that d' is the same as for the previously discussed lax criterion. The only difference between the two examples is the location of the criterion line. The strict criterion results in fewer hits than does a lax criterion.

6. Detectability refers to the difference between the means of the N and N + S distributions.

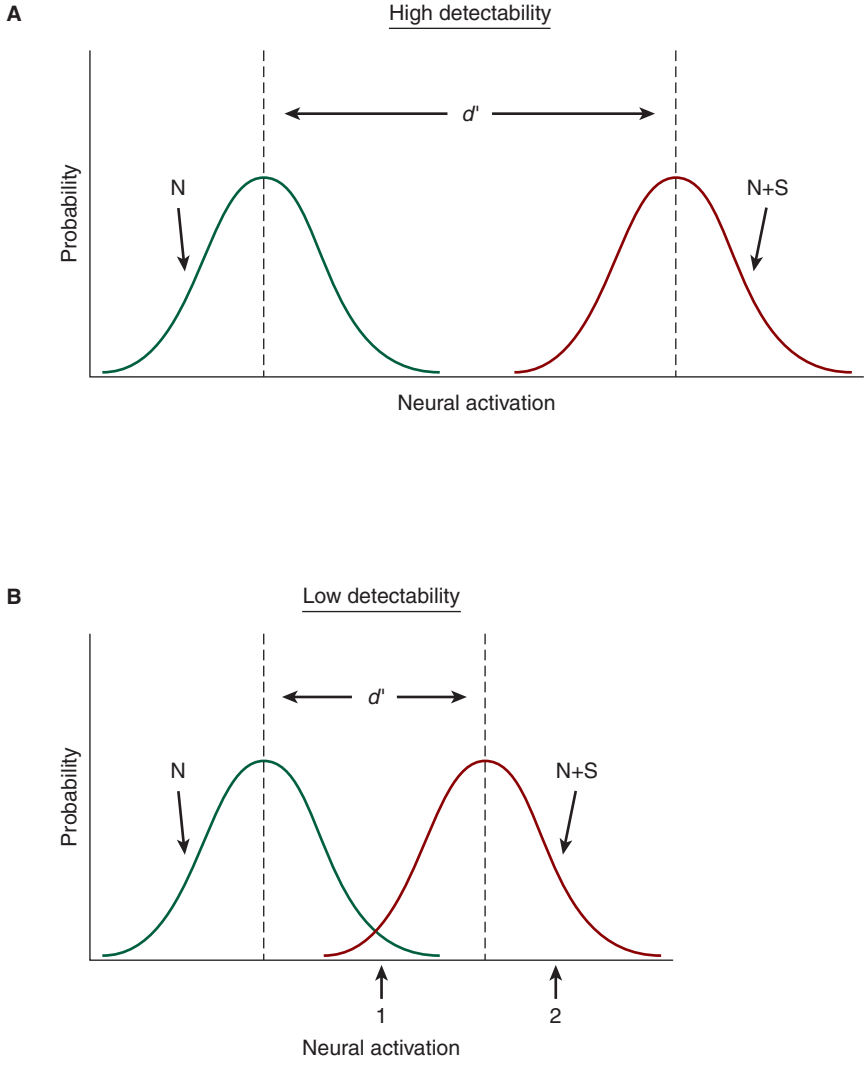

Figure 11–6. **A.** The larger the detectability of a stimulus, the greater the separation of the noise and noise plus signal distributions. For the high detectability illustrated, it is easy for the observer to determine if the level of neural activation is due to noise alone or signal plus noise. **B.** When the detectability is low, a stimulus may result in a level of neural activity that could be produced by noise or stimulus plus noise (indicated by point 1). If the stimulus is delivered at a time when the noise level is high, the resultant neural activation is unambiguously due to signal plus noise (point 2).

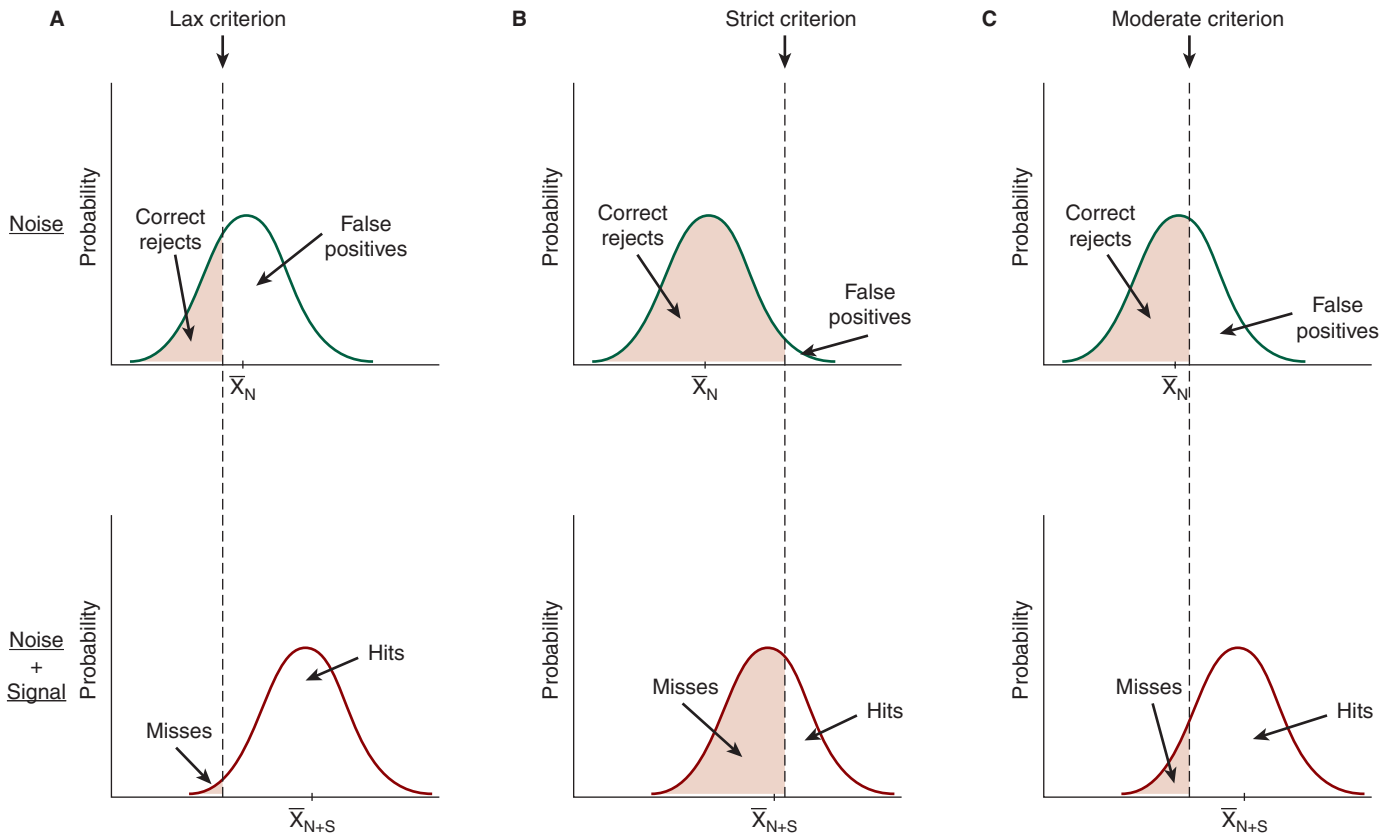

Figure 11–7. For each of these three examples, the detectability is the same; however, the threshold criterion varies from lax to strict to moderate.

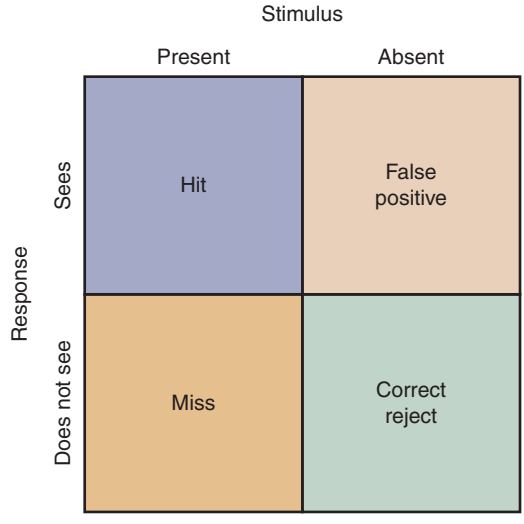

Figure 11–8. Possible outcomes for a signal detection experiment.

A practical implication of this, with regard to the example, is that some deer will not be shot. The payoff, however, is the low number of false positives. By employing a strict criterion, the hunter rarely shoots at targets other than deer.

Receiver Operating Characteristic Curves

A receiver operating characteristic (ROC) curve, which shows the probability of a hit as a function of the probability of a false positive, allows us to predict the effect of the observer criteria for a given detectability (d'). Consider Fig. 11–9A, which shows an ROC curve where $d' = 1$, and L represents a lax criterion, S a strict criterion, and M a moderate criterion. This graph summarizes nicely what we have already learned. For a lax criterion, the probability of a hit is high, but so is the probability of a false positive. For a strict criterion, the probability of a hit is low, but so is the probability of a false positive.

As illustrated in Fig. 11–9B, the ROC is a straight line when $d' = 0$, and becomes more curved as the detectability increases. In the special case of $d' = 0$, the stimulus is so weak that it produces virtually no signal. The straight-line function tells us that no matter what the criteria—lax or strict—the proportion of hits matches the proportion of false positives.

Now consider the other extreme, where the stimulus produces an infinitely large signal (i.e., $d' = \infty$). As can be seen in Fig. 11–9B, the observers' criterion has no effect on the proportion of hits and false positives. The observer always sees the stimulus, and there are never false positives.

How is the family of ROC curves in Fig. 11–9B generated? For a given detectability, the observer's criterion can be controlled by providing rewards for hits and

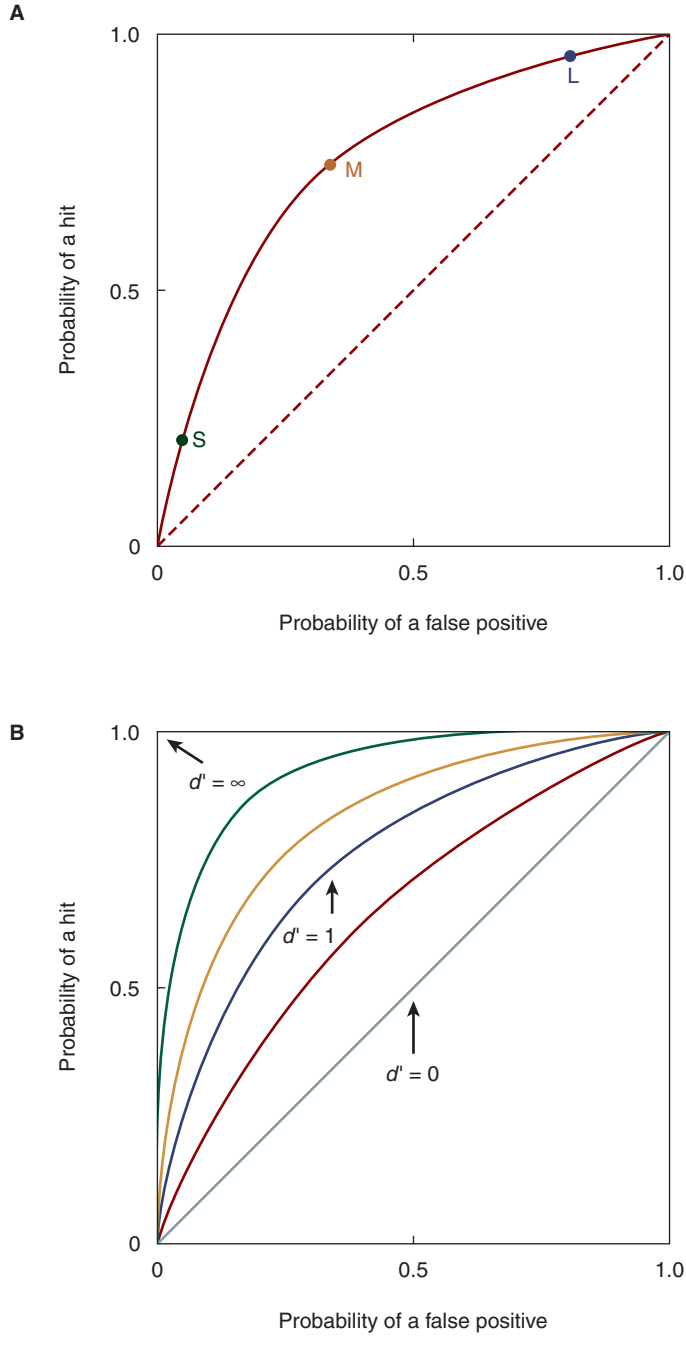

Figure 11–9. **A.** Receiver operating characteristic (ROC) curve, where $d' = 1$ (solid curve). *S, M,* and *L* represent strict, moderate, and lax criteria, respectively. **B.** ROCs for various levels of detectability.

penalties for false positives (a payoff matrix). For example, a subject may be told that each time she correctly responds to the stimulus (a hit), she will receive $5; however, incorrectly reporting the stimulus is there when it isn't (a false positive) will result in a penalty of $20. The result is a strict criterion. To induce the subject to adopt a lax criterion, the penalty for a false positive should be minimized relative to the reward for a hit (e.g., $25 reward for a hit and $5 penalty for a false positive).

Clinical Highlight

ROC curves are commonly used in the medical literature to assess the utility of clinical diagnostic procedures. Two important considerations when evaluating a new procedure are its sensitivity and specificity. Sensitivity refers to the percentage of patients with a given condition that are correctly identified as having the condition (hits). In comparison, specificity refers to the percentage of patients who are healthy and test negative (correct rejects). An ideal test has both high sensitivity and specificity.

Consider a hypothetical new test for glaucoma that compares differences in visual latency between inferior and superior visual fields. The premise is that since glaucoma often develops vertically asymmetrically, with the upper or lower nasal field affected first (a nasal step), patients with early glaucoma may manifest differences between superior and inferior field visual latencies. Suppose we conduct an experiment using both healthy patients and those with glaucoma, and arrive at the results in Table 11–1.

TABLE 11–1. HYPOTHETICAL GLAUCOMA TEST

Difference in Latency	Patients with Glaucoma ($N = 310$)	Healthy Patients ($N = 400$)
<25 ms	10	250
25–49 ms	25	90
50–75 ms	100	50
>75 ms	175	10

If we were to define abnormal as 25 millisecond, the sensitivity of the test would be 96.7% (300/310), while the specificity would be 62.5% (250/400). That is, 96.7% of the patients with glaucoma test positive, while only 62.5% of the patients who do not have glaucoma test negative. Just like an ROC experiment, we can change the threshold criteria. Let us say we adopt a more strict criterion, selecting 50 millisecond as the cutoff for abnormal. With this stricter criterion, the sensitivity falls to 91.7% (275/300), but the specificity rises to 82.5% (340/400). The test has become less sensitive, but more specific. As with ROC psychophysical experiments, there is a trade-off between hits (sensitivity) and correct rejects (specificity) as we change the threshold criteria.

Figure 11–10 shows ROC curves for two clinical tests, which are identified as "better" and "weak." Note that for the ordinate, "probability of a hit" is relabeled as "sensitivity" and for the abscissa "probability of a false positive" (which is equal to 1 − correct rejects) is relabeled as "1 − specificity." How can the ROC curves

Figure 11–10. ROC curves for two hypothetical clinical tests. For the function labeled "better" the area under the curve is large, allowing criteria to be located that optimize both the sensitivity and sensitivity of the test. Point *A* represents such a criterion. For the "weak test" curve, which has a slope of 1.0, any increase in sensitivity causes specificity to decrease by the same amount. For example, if the sensitivity increases from 0.50 to 0.90, the specificity decreases from 0.50 to 0.10.

be used to evaluate these clinical tests? For a test to be useful, there should be a cutoff criterion for abnormal that results in both relatively high sensitivity and specificity. For the "better" test, such a criterion can be found. Criterion A, for example, results in a sensitivity of 90% and specificity of 80%.

Now consider the test that produces the "weak" ROC curve. When the sensitivity is 90%, as indicated by point B, the specificity is a lowly 10%. The only way to increase specificity is to decrease sensitivity by the same amount, an undesirable situation. For instance, when the specificity increases from 10% to 50% (point C) the sensitivity decreases from 90% to 50%. A clinical test is made more powerful by optimizing the area under the ROC curve, which is the case for the "better" test, because there is a region of the curve (the ascending portion) where large improvements in sensitivity are accompanied by relatively small decreases in specificity.

WEBER'S LAW

Up to now, we have discussed threshold without consideration of the background against which the stimulus is detected. In many psychophysical procedures and experiments, however, the task is to discriminate between the combination of stimulus and background, and background alone.

This can be understood by considering an increment threshold experiment, as illustrated in Fig. 11–11A. The observer's task is to detect the increment stimulus, ΔI, which is flashed on the background, I_b. The threshold increment is sometimes referred to as a **just noticeable difference (JND)** or **difference limen (DL).** Another way of thinking of this task is that the observer must discriminate between the combined stimulus and background, and the background itself.

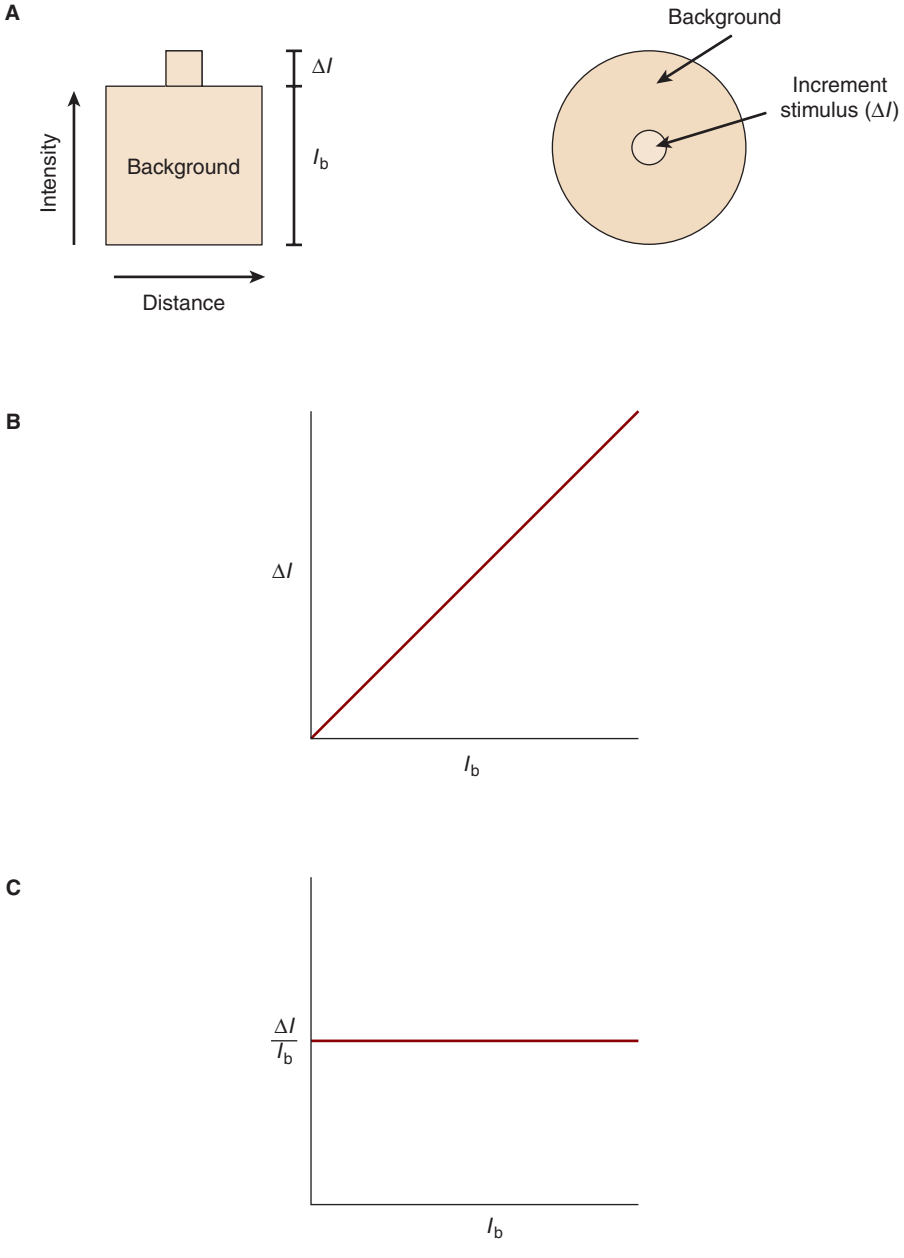

Figure 11–11. **A.** Increment threshold arrangement that can be used to demonstrate Weber's law. A luminance profile is on the left, and the observer's view of the stimulus is on the right. **B.** Graphical representation of Weber's law showing that the ratio of ΔI to I_b is a constant. **C.** Figure B replotted to show $\Delta I/I_b$ as a function of I_b.

The JND is not a constant, but changes as the background changes. As the background intensity increases, the JND also increases such that the ratio of the JND to the background intensity remains constant (see Figs. 11–11B, C). This is referred to as Weber's law, and it is expressed mathematically as

$$\Delta I = KI_b$$

or

$$K = \frac{\Delta I}{I_b}$$

where

ΔI = increment threshold (JND)
I_b = background illumination
K = Weber's constant or fraction

Weber's law applies not only to vision, but also to other senses as well. Consider the task of discriminating between two weights. Suppose the observer can barely tell the difference a dumbbell that weighs 10 lb (let us call it the background weight) and one that weighs 11 lb. In this case, the JND (or increment threshold) is 1 lb. Weber's constant is calculated as $(11 - 10)/10 = 0.10$.

If the background weight were 50 lb, what would now be the JND?

$$\Delta I = KI_b$$
$$(X - 50 \text{ lb})/X = 0.10$$
$$X = 55 \text{ lb}$$

where X is the combination of the increment and background weights. The increment threshold is calculated as follows:

$$\Delta I = X - 50 \text{ lb}$$
$$\Delta I = 5 \text{ lb}$$

To maintain a constant Weber's fraction of 0.10, the JND is 5 lb rather than the original 1 lb. The observer will barely be able to tell the difference between weights of 50 and 55 lb (and certainly will not be able to distinguish between 50 and 51 lb).

As discussed in Chapter 3, the visual system follows Weber's law over much (but not all) of its operational range, with different fractions for scotopic and photopic vision.[7] As the background becomes more intense (within the Weber regions), the increment threshold increases. Consequently, absolute sensitivity decreases, while relative sensitivity remains constant. This process—sensitivity regulation—results in a constant contrast threshold regardless of the background

7. Recall that the Weber's fraction for scotopic vision is 0.14 and for photopic vision is 0.015.

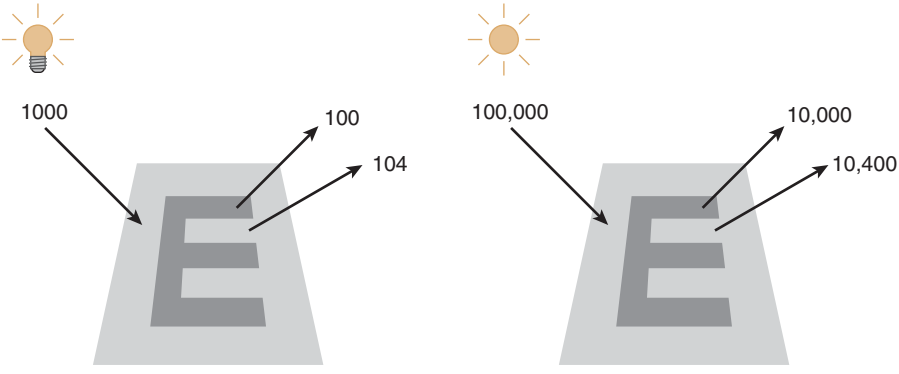

Figure 11–12. The optotype **E** reflects 10% of the light incident upon it, and the gray background reflects 10.40%. Whether under dim photopic illumination (1000 units of light) or bright illumination (100,000 units of light), the ratio of the optotype to background luminance (contrast) is the same. Because the visual system has evolved to detect contrast, rather than absolute luminance, the appearance of the **E** is the same under both dim and bright lighting conditions (as long as the light levels fall within Weber's region). This phenomenon is referred to as lightness constancy.

brightness. For scotopic vision, this contrast threshold is 0.14 (or 14%); for photopic vision it is 0.015 (or 1.5%).

An important result of sensitivity regulation is illustrated in Fig. 11–12, which shows a dark optotype on a gray background under dim and bright photopic illumination. The optotype reflects 10% of the light that falls on it, and the background reflects 10.40%; consequently, the optotype contrast is approximately 2%. When the illumination increases, contrast remains constant because the amount of light reflected from both the dark and gray surfaces increases at the same rate, resulting in a constant ratio of optotype to background luminance. Because Weber's fraction for photopic vision is approximately 0.015, the optotype is at threshold (barely detectable) under both lighting conditions. The optotype does not become more visible when more light is reflected from it.[8] Rather, because the contrast remains constant, the appearance remains the same, a phenomenon referred to as **lightness constancy.**[9]

What is the appearance of a stimulus of constant luminance when viewed against backgrounds of various luminances? Figure 11–13 demonstrates that brightness depends on the background, a phenomenon referred to as **simultaneous contrast.** Consistent with Weber's law, the contrast of the stimulus, not its luminance, is the key factor in predicting its appearance. Using a more complex configuration, Fig. 11–14 demonstrates how compelling this effect can be.

8. This is true only at those light levels where Weber's law is followed.
9. Another example of lightness constancy is the appearance of a black and white striped shirt viewed indoors under dim photopic conditions and outdoors under bright photopic conditions. The shirt appears about the same under both conditions.

Figure 11–13. Although the central squares are physically identical (i.e., they have the same luminance), each has a different brightness. As the background becomes darker, the central square appears brighter. This phenomenon is referred to as simultaneous contrast.

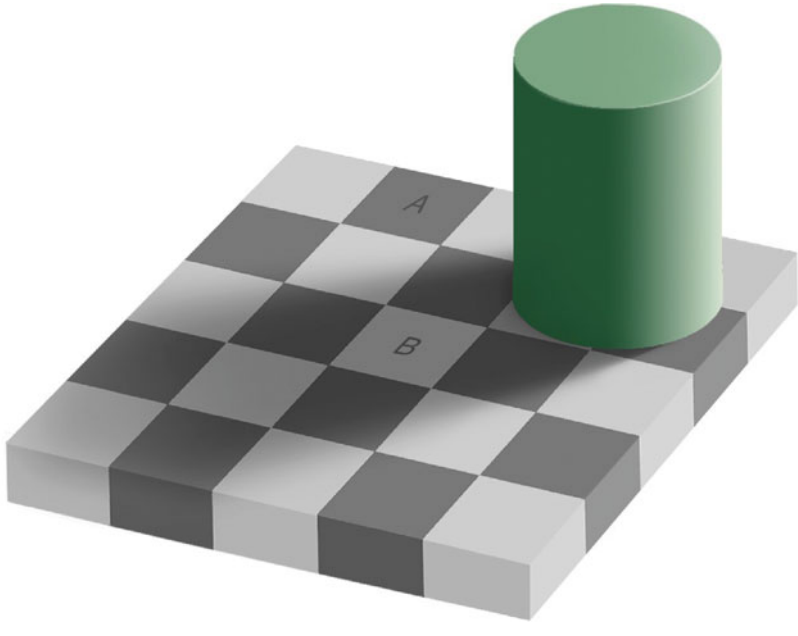

Figure 11–14. Which is brighter, square **A** or **B**? Although **B** is seen by most people as brighter, both **A** and **B** have the same luminance. The illusion is due to the visual system's tendency to analyze brightness in the context of surrounding elements. In this case, the surrounding squares and shadow cast by the cylinder contribute to the effect. (*Created by Professor Edward H. Adelson. http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html.*)

MAGNITUDE OF SENSATION

The discussion thus far has been limited to threshold stimuli. It is of interest, however, to quantify the growth in magnitude of sensation as the intensity of a stimulus is increased to suprathreshold levels. If the intensity of a light bulb is doubled, will it appear twice as bright? What is the relationship between magnitude of sensation and stimulus intensity?

Fechner (1860) attempted to answer this question by assuming that Weber's law applies to suprathreshold stimuli. According to Fechner's model, if the intensity of a light were to increase by 5 JNDs, it would appear five units brighter. This would result in a log relationship between intensity and sensation, as illustrated in Fig. 11–15A. **Fechner's log law** is mathematically expressed as

$$S = c \log I$$

where

S = magnitude of sensation (e.g., brightness)
I = stimulus intensity
c = constant related to Weber's constant

Fechner provided empirical evidence for this law by **indirect scaling.** Essentially, Fechner determined JNDs and assumed that all JNDs produce equal differences in the magnitude of sensation. This assumption is incorrect.

Stevens (1957) applied a remarkably simple paradigm to address this problem—he simply asked observers to directly assess the intensity of suprathreshold stimuli. This is referred to as **direct scaling** or **magnitude estimation.** When investigating brightness, lights of various intensities are presented, and the observer is asked to assign a number to the perceived brightness of each of the stimuli. For example, a very dim light may be labeled 1, and a very bright light may be labeled 10. Sensation, indicated by the numerical values assigned by the observer, is plotted as a function of light intensity.

The results of such an experiment, plotted on linear coordinates, are given in Fig. 11–15A. When plotted on log–log coordinates, the relationship is seen to be a power function (see Fig. 11–15B). Mathematically, these results are expressed as **Stevens' power law:**

$$S = I^c$$

where

S = magnitude of sensation (e.g., brightness)
I = stimulus intensity
c = constant

By magnitude estimation, Stevens was able to show that the growth in magnitude of sensation follows a power relationship, rather than a log relationship as postulated

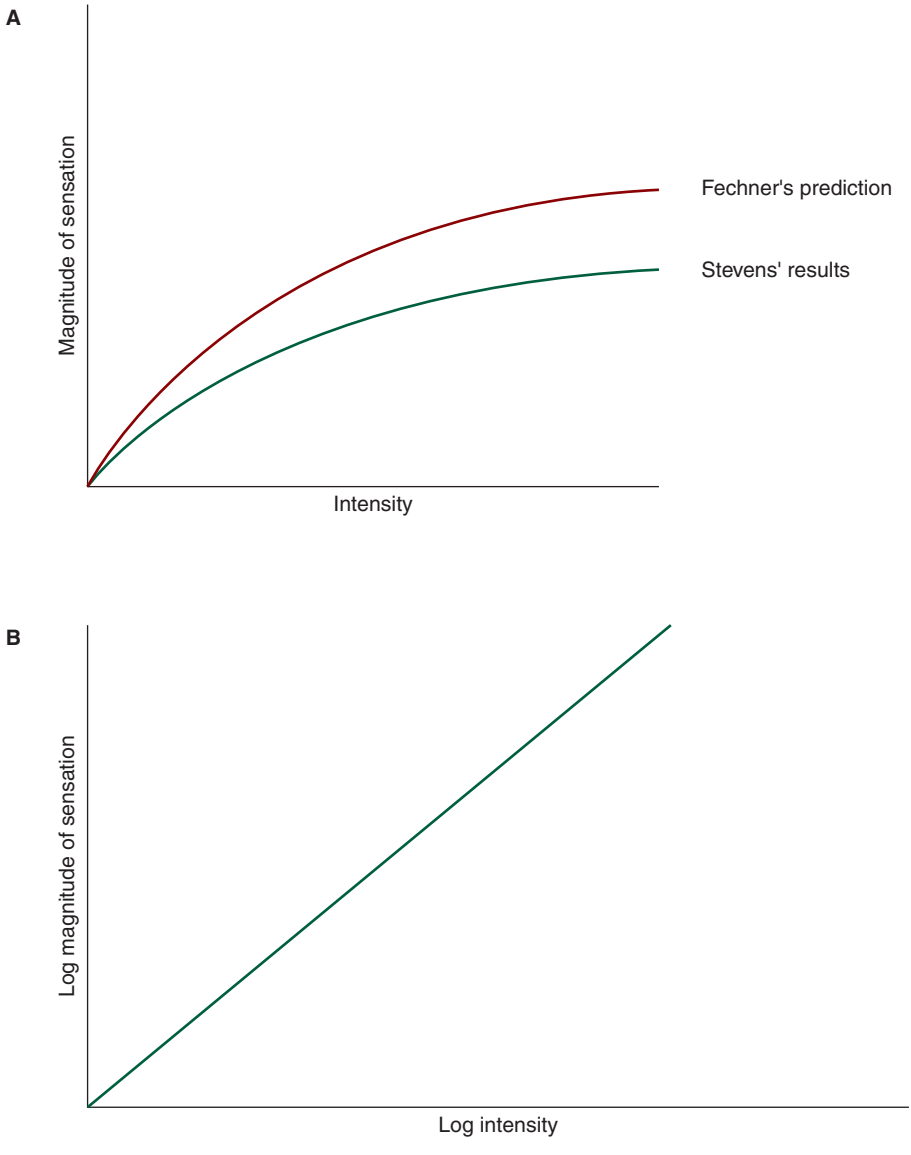

Figure 11–15. **A.** Fechner's log law and Stevens' power law plotted on linear coordinates. **B.** Stevens' power law plotted on log–log coordinates.

by Fechner. This means that a constant ratio of sensation is produced by a constant ratio of stimulation.

Stevens' law shows that there is a compression (i.e., saturation) of sensation as stimulus intensity increases. Consider a three-way light bulb that can be set at 0, 50, 100, or 150 W. The physical difference between each of these settings is the same (50 W), yet turning the light on (going from 0 to 50 W) is much more noticeable than adjusting the setting from 100 to 150 W.

SUMMARY

The determination of thresholds plays a large role in vision research and clinical eye care. The specific methodology used to determine a threshold may significantly affect the value obtained.

There is frequently a trade-off between the ease of determining a threshold and its validity. For clinical purposes, the repeatability (reliability) of a threshold measurement may be more important than its validity. Consequently, psychophysical procedures that are not unduly taxing for patients, such as the staircase method or method of limits, may provide clinically useful information.