Chapter 13

Separating Discrimination and Decision in Detection, Recognition, and Matters of Life and Death

John A. Swets

Editors' Introduction

This chapter explains signal detection theory (SDT) and illustrates the remarkable variety of problems to which it can be applied. When it was first developed (by the author of this chapter, among others), SDT revolutionized the way we think about the performance of sensory tasks, by explaining how performance depends not only on sensory information, but also on decision processes. The theory also provided ways to disentangle these two aspects of performance—to decompose or separate the underlying operations into sensory and decision processes and to decide whether the decision process is optimal, given the sensory information. Now, after four decades of research, we are led to the surprising conclusion that many tasks we perform, in domains ranging from memory recall to airplane maintenance, are analogous to sensory detection, and can be analyzed within the framework of this theory.

SDT asserts that performance in a discrimination or detection task must be divided into at least two stages. In the first stage, information about some situation is collected; in the second stage, this "signal" is evaluated for decision making. The signal provided by the first stage is often "noisy," which is to say, mixed with irrelevant material, and the second stage must evaluate the noisy signal provided by the first stage. To take a simple example, if an observer tries to decide whether she hears a faint sound, the message reaching her brain may be contaminated by noise, such as the variable sounds of her own pulse and breathing. One consequence of the noise is that decisions will sometimes be wrong. But the observer has some control over the errors that she makes. To use John Swets's terminology, there are two types of errors: false positives (e.g., asserting you heard something when there was nothing there) and false negatives (e.g., asserting you did not hear anything when there really was a sound). SDT explains how an observer can reduce the chance of one type of error, but only at the cost of increasing the chance of the other. (Can you see how a jury verdict might be a false positive or a false negative error, and how trying to reduce one type of error will affect the chance of the other?) SDT also predicts how observers will choose to balance the two types of errors.

SDT has its origins in work on noisy communication systems. Devices such as radars, radios, and TVs are all susceptible to electrical interference (one type of noise) and the engineering problem was how to determine when there was a "signal" (e.g., a radar image of a missile) within the obscuring noise. The big insight for psychology was that all communications systems, whether they be sensory systems, messages within the brain, or messages between people, have to deal with noise, particularly when the signal is weak. Early studies on perception showed that in audition and vision, the message that reached the brain was indeed noisy. Later studies showed that the retrieval of a weak memory could also be

described as an attempt to find the signal (memory) in the noise. Still other studies have shown that a radiologist examining an X-ray for evidence of cancer or an airplane technician examining a plane for evidence of stress cracks faces a similar situation, as Swets describes in this chapter. Other research shows that SDT can be applied to other important social questions, such as the reliability of blood tests for AIDS. Unfortunately, too few people yet appreciate the importance and broad applicability of SDT.

The work that Swets has done on many practical problems exemplifies the deep contributions that psychology can make. Swets discusses how a doctor examines an X-ray for evidence of cancer. If you have ever seen an X-ray, you know that it presents a vague shadowy image. The doctor's task is to make a decision on the basis of this vague image. This example illustrates a property of many decision-making situations. There may be several tell-tale signs of cancer in the X-ray, and the doctor must combine this information. Because this is often difficult to do reliably, Swets and his colleagues have developed computer programs to help doctors in this situation. This application makes use jointly of the strengths of humans and machines, and is therefore especially interesting in the context of cognitive science. And SDT can make important contributions to many other practical decision-making situations. It does not surprise us that in the 1994 White House policy report *Science in the National Interest*, Swets's work on signal detection theory and its applicability in an array of high-stakes decision-making settings was selected to illustrate the importance of basic behavioral science research.

Although there is a difference in terminology between Swets's discussion of the decision problem in detection and Wickens's discussion of the testing of statistical hypotheses (chap. 12, this volume), you will discover strong similarities.

Chapter Contents

13.1 Introduction

13.1.1 Detection, Recognition, and Diagnostic Tasks

Detection and recognition are fundamental tasks that underlie most complex behaviors. As defined here, they serve to distinguish between two alternative, confusable stimulus categories. The task of *detection* is to determine whether a specified stimulus (of category A, say) is present or not. For example, is a specified weak light (or specified weak sound, pressure, aroma, etc.) present or not? If not, we can say that a "null stimulus" (of category B) is present. The task of *recognition* is to determine whether

a stimulus known to be present is of category A or category B. For example, is this item familiar or new? The responses given in these tasks correspond directly to the stimulus categories: the observer says "A" or "B."

The task of *diagnosis* can be either detection or recognition, or both. In the cases of detection and recognition, the focus of this chapter will be on tasks devised for the psychology laboratory, as in the study of perception, memory, and cognition. In the case of diagnosis, the focus here will be on practical tasks, such as predicting severe weather, finding cracks in airplane wings, and determining guilt in criminal investigations. As a specific example of diagnosis, is there something abnormal on this X-ray image, and, if so, does it represent a malignant or a benign condition? Diagnoses are often made with high stakes and, indeed, are often matters of life and death.

In the tasks of primary interest, an organism, usually a human, makes observations repeatedly or routinely and each time makes a two-alternative choice based on that observation. Though considered explicitly here only in passing, the ideas of this chapter apply as well to observations (or measurements) and choices made by machines.

13.1.2 The Tasks' Two Component Processes: Discrimination and Decision

Present understanding of these tasks acknowledges that they involve two independent cognitive processes—one of discrimination and one of decision. In brief, a *discrimination* process assesses the degree to which the evidence in the observation (for example, perceptual, memorial, or cognitive evidence) favors the existence of a stimulus of category A relative to B. A *decision* process, on the other hand, determines how strong the evidence must be in favor of alternative A (or B) in order to make response A (or B), and chooses A (or B) after each observation depending on whether or not the requisite strength of evidence is met. We may think of the strength of evidence as lying along a continuum from weak to strong and the organism as setting a cutoff along the continuum—a "decision criterion," such that an amount of evidence above the criterion leads to a response of A and an amount below, to a response of B.

The observed behaviors in such tasks need to be separated or "decomposed," so that the discrimination and decision processes can be evaluated separately and independently. We want to measure the *acuity* of discrimination—how well the observer assesses the evidence—without regard to the appropriateness of the placement of the decision criterion; and we want to measure the *location* of the decision criterion—whether strict, moderate, or lenient, say—without regard to the acuity of discrimination. One reason to decompose is that an observed change in behavior may

reflect a change in the discrimination or the decision process. Another reason is that certain variables in the environment or in the person will have an influence on observed behavior through their effect on the discrimination process while other variables will be mediated by the decision process. Often we want to measure what is regarded as a basic process of discrimination, as an inherent capacity of the individual, in a way that is unaffected by decision processes that may vary from one individual to another and within an individual from one time to another. But as we shall also see, there are instances in which the decision process is the center of attention.

13.1.3 Diagnosing Breast Cancer by Mammography: A Case Study

The detection, recognition, and diagnostic tasks, and the decomposition of their performance data into discrimination and decision processes, are illustrated here by the diagnostic task that faces the radiologist in interpreting X-ray mammograms. Radiological interpretations assess the strength of the evidence indicative of breast cancer and provide a basis for deciding whether to recommend some further action. For our purposes, we shall consider the X-rays as belonging to either stimulus category A, "cancer," or stimulus category B, "no cancer"; and the corresponding response alternative to be a recommendation of surgery to provide breast tissue for pathology confirmation (i.e., a biopsy) or a "no action" recommendation because the breast is deemed "normal" as far as cancer is concerned.

13.1.3.1 Reading a Mammogram

It will help here to be concrete about how mammograms are interpreted visually (how they are "read")—that is, what perceptual features of the image are taken as evidence for cancer. And later in the chapter, we shall see how perceptual studies can improve both the acuity of radiologists in assessing those features and their ability to combine the assessments into a decision.

Radiologists look for ten to twenty visible features of a mammogram that indicate, to varying degrees, the existence of cancer. A perceptual feature is a well-defined aspect or attribute of a mammogram or of some entity within the mammogram. They fall into three categories: (1) the presence of a "mass," which may be a tumor; (2) the presence of "calcifications," or sandlike particles of calcium, which in certain configurations are indicative of cancer, and (3) "secondary signs," which are changes in the form or profile of the breast that often result indirectly from a cancer. Though all masses, calcifications, and secondary signs are abnormal, "malignant" abnormalities indicate a cancer, while "benign" abnormalities

do not. Thus the diagnostic task in mammography is one of detection (is there an abnormality present?) followed by recognition (is a present abnormality malignant or benign?).

Figure 13.1 illustrates some relevant features. Figure 13.1a shows a mass, seen as a relatively dark area, located at the intersection of the horizontal and vertical (crosshair) lines shown at the left and top of the breast. This mass has an irregular shape and an irregular border formed of spiked projections. These two features, of irregular mass shape and irregular border, are highly reliable signs of malignancy. The lower part of the breast image in figure 13.1a (above the vertical line at the bottom) shows some calcifications. These particular calcifications are probably benign because, compared to malignant ones, they are relatively large and scattered.

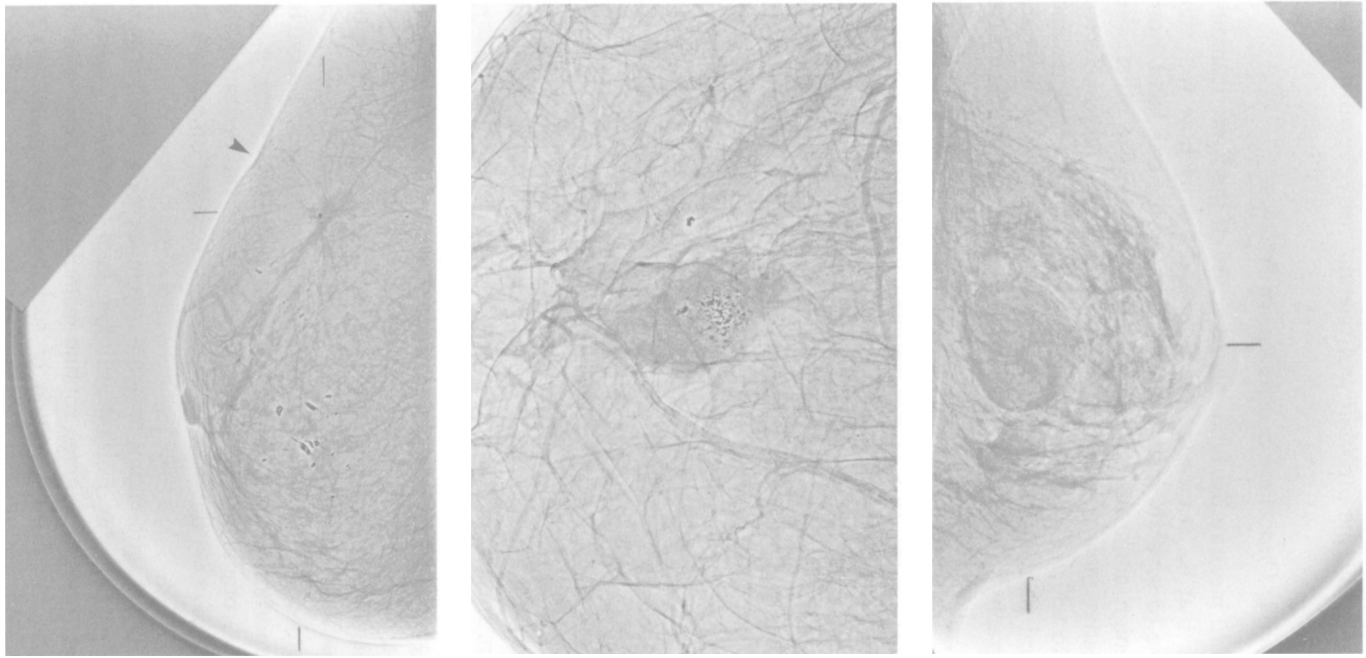
The arrow at the top left of figure 13.1a points to two kinds of secondary signs: a slight indentation of the skin and an increased darkness of the skin that indicates a thickening of the skin. Both are indicative of a malignancy.

In figure 13.1b, the mass in the center of the image is likely malignant because it has an indistinct or fuzzy border, indicating (as spiked projections do) a cancerous process spreading beyond the body of the tumor itself. This mammogram also shows some calcifications—which can occur inside of a mass, as they do here, or outside of a mass. Because these calcifications are relatively small and clustered, they suggest a malignancy.

The mass of figure 13.1c is benign and is, specifically, a relatively harmless cyst. A cyst has a characteristically round or oval shape and a clear and smooth border.

I hasten to mention that figure 13.1 gives exceptionally clear examples of malignant and benign abnormalities, to suit a teaching purpose; in practice, these perceptual features may be very difficult to discern. I wish also to draw a conceptual point from figure 13.1 that is fundamental to detection, recognition, and diagnostic tasks: observers must often combine many disparate pieces of information into a single variable, namely, the degree to which the evidence favors one of the two alternatives in question, category A relative to category B.

We can also think of this degree-of-evidence variable as indicating the *probability* that the stimulus is from category A. Then the observer who must choose between A and B will set a cutoff, or criterion value, along an evidence continuum viewed as a probability continuum—in effect, along a scale from 0 to 100. A cutoff at 75, say, means that the probability that the stimulus is an A must be 0.75 or greater (and that the stimulus is a B, 0.25 or less) for the observer to choose A. As indicated earlier and developed in more detail later, the evidence may be complex—it may contain many variables, or many "dimensions"—but, for purposes of a two-alternative A or B response, it is best to boil the evidence down to



a b c

Figure 13.1
Three illustrative mammograms, showing perceptual features as described in the text.

one dimension, namely, the probability of one alternative relative to the other.

13.1.3.2 Decomposing Discrimination and Decision Processes

There is a need to measure the fundamental acuity of the X-ray mammogram technique, that is, to measure precisely and validly how well this technique is able to separate instances of cancer, on the one hand, from instances of benign abnormality or no abnormality, on the other. We desire a quantitative measure of acuity that is independent of (unaffected by) the degree to which any or all radiologists are inclined to recommend a biopsy. Several parties wish to know in general terms how accurate X-ray mammography is so that it can be fairly compared to alternative diagnostic techniques, for example, physical examination (palpation), and the other available imaging techniques of ultrasound, computerized axial tomography ("CAT scans" or CT), and magnetic resonance imaging (MRI or MR). Hospital administrators and insurers, as well as physicians and patients, wish to use a technique that is "cost-effective," one that provides the best balance of high acuity and low cost. They need to appreciate that the acuity of diagnostic imaging techniques is fundamentally determined and set by the limitations of the technology as well as the perceptual abilities of the interpreter, whereas the decision criterion may tend to vary somewhat from one technique to another, and, indeed, can be adjusted by agreement. Moreover, agencies that certify individual radiologists for practice must know how acute the mammography technique is in each practitioner's hands, irrespective of decision tendencies.

Similarly, there is a need to know quantitatively how individual radiologists set their respective decision criteria, and how the profession generally sets its criterion, for recommending biopsy. A very lenient criterion—requiring only a little evidence to recommend biopsy (e.g., 5 on a 100-point scale) might be adopted in order to identify correctly, or "find," a large proportion of existing cancers. And, in fact, radiologists do set very lenient criteria in reading mammograms, with the idea that early detection of cancer reduces the risk of fatality. There are constraints, however, on how lenient the decision criterion can be. A lenient criterion will serve to find a large proportion of existing cancers, but, at the same time, it will lead to many recommendations of biopsy surgery on noncancerous breasts and thus increase the number of patients subjected unnecessarily to such surgery.

Radiologists read mammograms in two different settings, which require different placements of the decision criterion. In a "screening" setting, nonsymptomatic women are given routine mammograms (every year or every few years), and the proportion of such women actually having cancer is low, about 2 in 100 (Ries, Miller, and Hankey 1994). In a

"referral" setting, on the other hand, patients have some symptom of cancer, perhaps a lump felt in the breast. Among such patients, the proportion having cancer is considerably higher, about 1 in 3. I suggest later that a rather strict criterion is appropriate to the screening situation and a rather lenient criterion is appropriate to the referral situation.

It is clear, in any case, that biopsy surgery is expensive financially and emotionally so that unnecessary surgery needs to be curtailed. In fact, a large number of unnecessary biopsy recommendations can be unmanageable as well as undesirable. As the government health agencies advise more women to undergo routine, annual mammograms, and as more women comply, the number of pathologists in the country may not be large enough to accommodate a very lenient biopsy criterion. One way to measure the criterion in this case is by the fraction of breast biopsies that turn out to confirm a cancer: the "yield" of biopsy. In the United States the yield varies from about 2/10 to 3/10; approximately 2 or 3 of 10 breasts biopsied are found to have cancer (Sickles, Ominsky, and Sollitto 1990). England's physicians generally use a stricter criterion; their biopsy yield is about 5/10 (unpublished data from the UK National Breast Screening Centers, 1988–1993).

13.1.4 Scope of This Chapter

Although, in using mammography as a case study, I have tried with continual references to make new terms concrete, it will be necessary to treat the detection, recognition, and diagnostic tasks in formal terms, both to reflect their generality and to show how their performance data can be analyzed into discrimination and decision processes. Section 13.2 shows how two variables considered in the previous discussion of mammography—the proportion of cancerous breasts recommended for biopsy and the proportion of noncancerous breasts recommended for biopsy—are the basis for separating and measuring the two cognitive processes. More generally, the variables will be considered as the proportion of times that response A is given when stimulus A is present and the proportion of times that response A is given when stimulus B is present. To show the interplay of these variables in defining measures of acuity and the decision criterion, section 13.2 takes an excursion into a theory of signal detection that is based on the statistical theory of decision making. Section 13.3 then shows how both the theoretical ideas and the measures of discrimination and decision performance can be represented simply and compactly in a single graph.

Section 13.4 presents briefly some examples of successful separation of the two cognitive processes—examples taken from psychological tasks of perception and memory and from the practical tasks of polygraph lie

detection, information retrieval, and weather forecasting. With that additional motivation, section 13.5 returns to theory and measurement in order to reinforce the main concepts via a dice game that you are invited to play as a calculational exercise.

Section 13.6 briefly describes the theory of how several observations may be combined for each decision—much as the radiologist examines several perceptual features of a mammogram—in order to increase discrimination acuity. Section 13.7 then shows how the radiologists can be given certain aids to help them attend to the most significant perceptual features, to assess those features better, and to better merge those individual feature assessments into an estimate of the probability that cancer is present; and how these aids improve performance by simultaneously and substantially increasing the proportion of cancers found through biopsy while decreasing the proportion of normal breasts recommended for biopsy. Ways of setting and monitoring the radiologist's decision criterion are also discussed.

Section 13.8 treats briefly another practical example, that of human inspectors using certain imaging techniques to detect cracks in airplane structures. Data are presented on the state of the art that dramatically illustrate the need for separating discrimination and decision processes, in order to increase acuity and to set appropriate decision criteria—a need that remains to be appreciated in the materials-testing field.

Finally, section 13.9 gives a historical overview, describing how in the 1950s the relevant theory was taken into psychology from statistics, where it applied to testing statistical hypotheses (Wald 1950), via engineering, where it applied to the detection of radar and sonar signals (Peterson, Birdsall, and Fox 1954), to replace a century-old theory of an essentially fixed decision criterion, equivalent to sensory and memory thresholds (Green and Swets 1966). The diverse diagnostic applications of the theory, growing from the 1960s on, were based originally on psychological studies showing the validity of the theory for human observers in simple sensory tasks (Tanner and Swets 1954; Swets, Tanner, and Birdsall 1961).

13.2 Theory for Separating the Two Processes

13.2.1 Two-by-Two Table

The statistical theory for separating discrimination and decision processes is based on a two-by-two table, in which data from a task with two stimuli and two responses appear as counts or frequencies in cells of the table. As shown in table 13.1, the stimulus alternatives (cancer and normal) are represented at the top of the table in two columns, and the response

Table 13.1
The two-by-two table of stimulus (truth) and response (decision), showing the four possible decision outcomes.

		Stimulus (Truth)		
		Category A Positive	Category B Negative	
Response (Decision)	Category A Positive	(1) true positive (TP)	(2) false positive (FP)	
	Category B Negative	(3) false negative (FN)	(4) true negative (TN)	

alternatives (recommendation of biopsy and of no action) are represented at the side, in two rows. In general terms, as indicated in the table 13.1, both the stimulus and response alternatives can be called either "positive" (cancer exists; a biopsy recommendation is made) or "negative" (the patient is normal; no action is recommended). The convention is to refer to the stimulus of special interest (in our example, cancer) as "positive," even when that stimulus produces negative affect. (Colloquially, when the response is "positive," the mammogram, or other medical test, is also said to be "positive").

To acknowledge some terminology that has been implicit in this discussion, and needed beyond psychological studies, the two "stimulus" categories (A and B) are generally regarded as two alternative "states of the world." They may be conditions or events that follow, instead of precede, the "response" made to an observation. For example, the relevant states of the world follow the "response" in weather forecasting. And similarly, the "response" is more generally called a "decision"; the decision is a choice between two alternatives that may follow or, instead, anticipate, the occurrence of one or the other alternative. Establishing that one or the other of the states of the world actually exists relative to a particular decision (e.g., confirmation by biopsy) is said to provide the

"truth" relative to that decision. And so we can ask, "Which stimulus occurred?" or "What is the truth?" For convenience, we shall use mostly the "stimulus-response" terms.

There are four possible stimulus-response "outcomes," as shown in table 13.1. When the positive response coincides with the positive stimulus—for example, when a response to make a biopsy is followed by the pathologist's confirmatory determination of cancer—the outcome falls in the cell labeled "1". It is called a "true positive" (TP). Cell 2 represents the coincidence of the negative stimulus and a positive response—for example, when a biopsy is recommended for a normal patient; this outcome is called a "false positive" (FP). Proceeding, there are "false negative" (FN) and "true negative" (TN) outcomes, as indicated in cells 3 and 4, respectively. In FN no action is recommended even though cancer exists, and in TN no action is recommended and none is necessary. Cells 1 and 4 represent correct (true) responses; cells 2 and 3 represent incorrect (false) responses.

The counts or raw frequencies of the four possible coincidences of stimuli and responses are denoted a, b, c, and d in table 13.2, for cells 1, 2, 3, and 4, respectively. As shown, the two column sums are $a + c$ and $b + d$, and the two row sums are $a + b$ and $c + d$. The total number of counts is $N = a + b + c + d$. The proportion of positive stimuli for which a positive response is made is $a/(a + c)$ and is denoted here the "true positive proportion" (TPP). Similarly, we have $b/(b + d)$ or the "false positive pro-

Table 13.2
The two-by-two table with cell entries indicating frequencies of stimulus-response outcomes, to provide definitions of the four relevant proportions, as shown.

Response (Decision)	Stimulus (Truth): Positive	Stimulus (Truth): Negative	
Positive	(1) a	(2) b	$a + b$
Negative	(3) c	(4) d	$c + d$
	$a + c$	$b + d$	$N = a + b + c + d$

$TPP = a/(a + c)$ $\quad$ $FNP = c/(a + c)$
$FPP = b/(b + d)$ $\quad$ $TNP = d/(b + d)$

portion" (FPP). The remaining two possibilities are $c/(a+c)$ or "false negative proportion" (FNP), and $d/(b+d)$ or "true negative proportion" (TNP). Note that the proportions defined for each column add to 1.0—$a/(a+c)$ and $c/(a+c)$ in the left column and $b/(b+d)$ and $d/(b+d)$ in the right column—and thus just two of the four proportions (one from each column) contain all of the information in the four. As suggested earlier, the two column proportions to be used here are those defined by cells 1 and 2, namely, TPP and FPP; these are the two proportions when a positive response occurs. In our example, recommendations of biopsy when cancer exists give the TPP and those when no cancer exists give the FPP.

Finally, note that the proportions of positive and negative stimuli are, respectively, $(a+c)/N$ and $(b+d)/N$. Let us denote them $P(S+)$ and $P(S-)$. The proportions of positive and negative responses are $(a+b)/N$ and $(c+d)/N$, respectively, and are here denoted $P(R+)$ and $P(R-)$. In psychological experiments, the proportions of positive and negative stimuli can be set as desired (e.g., each at .50), whereas in real diagnostic tasks, they are determined by the actual occurrences of positive and negative stimuli in a given diagnostic setting, and may be very extreme, for example, .01 (cancer present in only 1 percent of the stimuli) and .99 (normal).

13.2.1.1 Change in Discrimination Acuity

Table 13.3 shows hypothetical data that we shall take as a baseline to consider the differential effects of a change in acuity and a change in the

Table 13.3
Hypothetical data to be taken as a baseline

Response (Decision)	Stimulus (Truth): Positive	Stimulus (Truth): Negative	
Positive	30	20	50
Negative	20	30	50
	50	50	100

TPP = .60 FPP = .40
P(S+) = .50 P(S−) = .50
P(R+) = .50 P(R−) = .50

Table 13.4
Hypothetical data indicating a change in discrimination acuity, relative to table 13.2.

		Stimulus (Truth)		
		Positive	Negative	
Response (Decision)	Positive	40	10	50
	Negative	10	40	50
		50	50	100
		TPP = .80	FPP = .20	
		P(S+) = .50	P(S−) = .50	
		P(R+) = .50	P(R−) = .50	

decision criterion. The proportions of positive and negative stimuli and responses, indicated by frequencies in the marginal cells of the columns (for the stimuli) and rows (for the responses), are all .50 (50/100). The summary measures of performance are TPP = .60 (30/50) and FPP = .40 (20/50).

Table 13.4 illustrates a change (specifically, an improvement) only in discrimination acuity: relative to table 13.3, the correct TP and TN decisions increase from 30 to 40 (TPP increases from .60 to .80) while the incorrect FP and FN decisions decrease from 20 to 10 (FPP decreases from .40 to .20). Acuity is greater because the proportions of true decisions of both kinds increase while the proportions of false decisions of both kinds decrease. Meanwhile, the marginal frequencies are unchanged; in particular, $P(R+)$ and $P(R-)$ remain at .50. Hence, there has been no change in the tendency toward a positive response, which we shall see below reflects no change in the decision criterion.

13.2.1.2 Change in the Decision Criterion

Table 13.5 shows, relative to table 13.3, a change in the decision criterion. We can say that the criterion has become more lenient because both TPP and FPP have increased, the former from .60 to .80 and the latter from .40 to .60. Overall, the proportion of positive responses has increased, $P(R+)$ changing from .50 to .70, consistent with a change to a more lenient criterion for making a positive response.

Table 13.5
Hypothetical data indicating a change in the decision criterion, relative to table 13.2.

Response (Decision)	Stimulus (Truth): Positive	Stimulus (Truth): Negative	
Positive	40	30	70
Negative	10	20	30
	50	50	100

TPP = .80 FPP = .60
P(S+) = .50 P(S−) = .50
P(R+) = .70 P(R−) = .30

13.2.1.3 Separation of Two Processes

We would like to know if table 13.5 relative to table 13.3 indicates a change in acuity as well as a change in the decision criterion. I shall defer the discussion of quantitative particulars of that question with the promise that an analysis of the relative changes in TPP and FPP will provide the answer.

We can see now that considering TPP alone to be a measure of discrimination acuity, as has often been done in psychology and elsewhere, will not provide the answer. That quantity increased from .60 to .80 in both tables 13.4 and 13.5; although the increase in TPP in table 13.4 reflects an acuity change only, as far as we know from the tables the increase in TPP in table 13.5 is due partly, and perhaps entirely, to a change in the decision criterion. Nor will looking only at the overall proportion of correct decisions—namely, $P(C) = (a + d)/N$—provide the answer. One might be tempted to infer that acuity has not changed from table 13.3 to table 13.5 because $P(C) = .60$ in both ($30 + 30$ and $40 + 20$, respectively), but such an inference is generally not justified. $P(C)$ is not a reliable or valid measure of acuity; both TPP and FPP need to be considered if the changes in the decision criterion are to be partialed out to leave a pure measure of acuity.

13.2.2 Statistical Decision and Signal Detection Theories

Statistical decision theory and signal detection theory have much in common. Here I present the more complete detection theory to show how

discrimination and decision effects can be separated in detection, recognition, and diagnostic tasks.

13.2.2.1 Assumptions About an Observation

Signal detection theory incorporates three basic assumptions about an observation: (1) it can be represented as a value along a single dimension—which we shall call the "decision variable," x—that reflects the likelihood of stimulus A relative to stimulus B; (2) observations of stimuli from either category A or category B will vary from one observation to another in the value of x they yield; and (3) values of x for observations from one category will overlap those from the other category. Let us consider these assumptions briefly.

ASSUMPTION 1: The observation is one-dimensional. The x value might represent an observer's confidence that stimulus A is present as opposed to B, as it does in mammography. Or it might be the amount of pressure in an eye as measured by an ophthalmologist screening for glaucoma. In a simple sensory task, the idea of a single dimension may seem plausible: detecting a spot of light might depend only on the rate of neural impulses in certain brain cells. However, no matter how many dimensions an observation may have—variations in a spot of light, for example, in hue, saturation, brightness, shape, duration, and so forth—the assumption of detection theory is that the only thing that counts is the likelihood of A relative to B implied by the several dimensions taken together. If there are only two possible responses, then, for purposes of making a decision, the observation need have only one dimension; indeed, it should be reduced to one dimension if the best decisions are to be made.

ASSUMPTION 2: Observations of stimuli from either stimulus category will vary. It is clear that X-rays of breasts with either benign or malignant lesions will vary from one patient to another in the apparent likelihood of malignancy. Some will show some particular signs of cancer and some will show other particular signs, and more or less clearly. Similarly, samples from each stimulus category in a sensory detection experiment will vary from trial to trial. For example, when observers try to detect the occurrence of a faint tone within a background of white noise, the noise is inherently variable and the tone can also vary because the tone generator is not perfectly stable. Even without such a noise background, essentially in quiet, the stimuli will appear variable to the observer because of natural variations in the observer's physiological, sensory system. Thus, auditory sensations affecting tone detection may result from the movements of blood within the ear and visual sensations affecting light detection may result from variations in blood pressure within the eye. (See Wickens, chap. 12, this volume, for more discussion of sample variability and sample distributions.)

ASSUMPTION 3: Values of the observation x from the two categories will overlap each other. A background of white noise in a tone detection task has, by definition, energy at all audible frequencies, including the frequency of the tone signal, and hence when the observer is attending narrowly to frequencies around the tone, the noise can produce observations that sound like the signal. A malignant lesion may give little or no positive evidence in an X-ray. A nonliar in a polygraph exam may have a physiological reaction as large as most liars. The overlap of observations will vary from small to large and discrimination acuity will vary accordingly. If there is no overlap, then no detection, recognition, or diagnostic problem exists. This is because all x values produced by stimuli from category A will be larger than any x values produced by stimuli from category B, and an observer should have no difficulty discriminating between the two stimulus categories.

13.2.2.2 Distributions of Observations

Look ahead, please, to figure 13.7, which shows a representation of the three assumptions as a pair of histograms labeled "0" and "3" on a decision variable on the x-axis. You can ignore there the particular names of the decision variable, which we have generally termed x, and of the two histograms. The histogram distributions may look more familiar if you imagine vertical lines from each tick mark along the horizontal axis at the bottom up to the dashed and solid lines, respectively, of the two distributions; then observations at each value of the decision variable are represented by vertical bars. The height of a given bar represents the probability that its value of x will occur. We see in accordance with assumption 2 that the observations from each stimulus category vary—from 2 through 12 for the stimulus category shown on the left ("category B" as represented by the histogram labeled "0") and from 5 through 15 for the category on the right ("A" or the "3" histogram). In accordance with assumption 3, the observations from the two categories overlap one another, such that the values 5 through 12 occur from both categories. There is surety in this case for the extreme values 2 to 4 and 13 to 15, but such surety may not be evident in all cases.

To develop the analytical tools of signal detection theory, it is convenient to consider the distributions in a different form, namely, as continuous probability distributions rather than as discrete histogram distributions. Figure 13.2 shows a representation of the three assumptions as continuous probability distributions on the decision variable x. Again, category B gives rise to values of the observation x according to the distribution on the left; category A, according to the distribution on the right. Each value of x will occur with a probability, when B is present, that is represented by the height of the B distribution at the particular value of x. And the

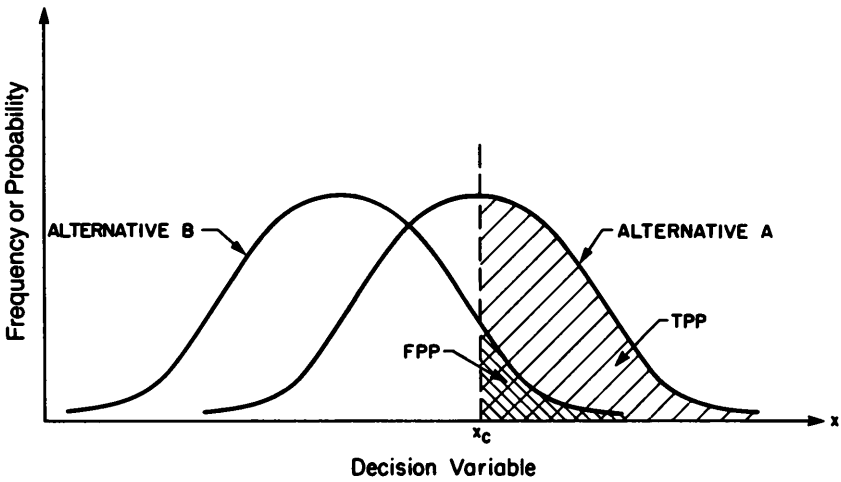


Figure 13.2
Probability distributions of observations, or of the decision variable x, for stimulus alternatives A and B, with an illustrative decision criterion x_c, and its corresponding areas or probabilities FPP and TPP.

height of the A distribution, as one moves along the curve, gives the probability that each value of x will occur when A is present.

Readers familiar with the testing of statistical hypotheses will recall a picture similar to that of figure 13.2, with the null hypothesis on the left and an alternative hypothesis on the right. In signal detection theory, the two distributions are referred to as distributions of "noise alone" and "signal plus noise," respectively. This terminology arises for historical reasons because with an electronic radar or sonar "signal," a stimulus from category A is always viewed against a background of random interference or static called "noise," that is, a stimulus from category B. This noise may be generated in the environment or in the detection device. In the case of a human detector, as mentioned, noise results from the physiological variability inherent in sensory systems and in the nervous system in general. Thus, even though a stimulus from B is sometimes called a "null stimulus" in a detection task, it is nevertheless a stimulus that impinges on the organism, and one that may mimic, and be confused with, a stimulus from A, or the signal. The occurrence of a signal adds something ("energy," say) to the noise background; in general, values of x are larger for A or the signal plus noise than for B or noise alone, and the distribution of x for A is displaced to the right of the distribution of x for B. A good deal of data from the tasks of interest here indicates that these probability distributions can reasonably, as well as conveniently, be considered as nor-

mal (Gaussian) distributions, that is to say, as having the specific form of the bell-shaped distribution that is portrayed in the figure. Ordinarily, however, despite their portrayal in figure 13.2, the distributions will have different spreads or variances (e.g., the signal may add a variable quantity, rather than a constant quantity, to the noise, and the signal distribution will then have a wider spread). Clearly, discrimination acuity will depend on the separation of the two distributions: roughly speaking, the less overlap, the less confusable are the stimuli. If, for example, the eye pressure test for glaucoma is perfectly acute, with pressure varying from 30 to 50 physical units when glaucoma is present and from 10 to 29 physical units in normal eyes, then there is no overlap and the two categories of disease and health will never be confused. If, however, the pressure test is not very acute, for example, if glaucoma is associated with measurements between 25 and 50 units and normal eyes with measurements between 10 and 35 units, then overlap exists between 25 and 35 units and the disease and health categories will often be confused.

13.2.2.3 The Need for a Decision Criterion

In the face of such variability in observations and confusability in stimulus categories, a rational decision maker will strive for whatever consistency is possible. A basic kind of consistency is always to give the same response to a given value of the decision variable. In other words, the decision maker will attempt to adopt, at least approximately, some particular value of x as the decision criterion—call it x_c—such that values of x greater than x_c always lead to response A and values of x less than x_c always lead to response B. In our mammography example, the radiologist will choose some level of confidence that cancer exists (e.g., 5 on a 100-point scale) and use it as consistently as possible for the decision criterion.

Figure 13.2 shows a somewhat conservative decision criterion x_c as a vertical line on the decision variable and represents the probabilities of true positive outcomes (TPP) and false positive outcomes (FPP) that result from that criterion. To see how this is so, consider how an observer who is being cautious about giving response A when response B is present might adopt the criterion shown, where most observations will be to the left of the criterion, each one occurring with a relative frequency or probability represented by the height of the B distribution at a given value of x. But alternative B also produces values of x greater than x_c—that is, to the right of x_c. The probability that a *given* value of x greater than x_c will occur is represented again by the height of the distribution at that value. However, the probability that any value of x greater than x_c will occur is given by considering the probabilities of those several values of x relative to the total probability of all values of x. As I justify in a later discussion, the total area under the curve is taken to represent a probability of 1.0.

The probability that any value of x greater than x_c will occur is equal to the proportion of the total area under the curve that lies to the right of x_c. Specifically, FPP is equal to the proportion of the area under curve B to the right of x_c and TPP is equal to the proportion of the area under curve A to the right of x_c, as shown by hatch lines. Both probabilities will increase for more lenient criteria—as x_c moves to the left—and decrease for stricter criteria—as x_c moves to the right. (Though not labeled in figure 13.2, the proportions of area to the left of the decision criterion x_c represent the two decision probabilities that are complementary to FPP and TPP, respectively: the true negative proportion under curve B and the false negative proportion under curve A.)

13.2.2.4 Decision Criterion Measured by the Likelihood Ratio

One general way to measure the location of the decision criterion—independent of a particular decision variable, be it a mammographer's confidence or the opthamologist's physical measure of pressure in the eye—is by the quantity called "likelihood ratio" (*LR*). The *LR* at any value of x of the decision variable is the likelihood that this value of x came from distribution A relative to the likelihood that it came from distribution B. In terms of figure 13.2, it is defined as the ratio (at that value of x) of the height of the A distribution to the height of the B distribution. Thus, for example, the *LR* is 1.0 where the two curves cross. As seen in figure 13.2, the *LR* for a decision criterion increases (toward infinity) as the criterion moves to the right and decreases (toward zero) as the criterion moves to the left. (If the two distributions have unequal spreads, the two curves will cross a second time out at one or the other tail, but we shall ignore such end effects for present purposes.) As one specific example, observe that in figure 13.2 the illustrative criterion x_c is set at the midpoint of the right-hand distribution, where the height of A happens in this picture to be a little more than twice the height of B. More precisely, the *LR* at that point is 2.5. Other measures of the decision criterion have been considered; an advantage of *LR*, as we shall see next, is that it facilitates definition of the best, or the "optimal," criterion for any specific task. Let us denote a criterion value of *LR* as LR_c.

13.2.2.5 Optimal Decision Criterion

In most tasks, particularly in diagnostic tasks, an observer/decision maker will want to choose a location of the decision criterion that is best for some purpose. In the mammography example, one desires an appropriate balance between the proportions of false positive and true positive responses, FPP and TPP. This is because FP outcomes have significant costs and TP outcomes have significant benefits; similarly with the other two stimulus-response outcomes: false negative (FN) outcomes have costs and true

negative (TN) outcomes have benefits. Further, the proportions of the positive and negative stimuli, which we have denoted $P(S+)$ and $P(S-)$, will affect the location of the best criterion. We shall discuss this relation later, but you can see now that if $P(S+)$ is high—if, for example, in a certain breast-cancer referral setting, most of the patients seen have a malignancy—one would do best to reflect that high $P(S+)$ in a high proportion of positive responses, $P(R+)$, and a lenient criterion toward the left of figure 13.2 is needed to produce a high $P(R+)$. Conversely, a low $P(S+)$—as in mammography screening of nonsymptomatic women—is best served by a low $P(R+)$, which requires a strict criterion toward the right of figure 13.2.

An optimal decision criterion can be defined quantitatively in various ways. One very useful definition of the optimum is based, as discussed, on the prior probabilities of the two stimuli, $P(S+)$ and $P(S-)$, and the benefits and costs of the four decision outcomes, TP, FP, FN, and TN, shown in table 13.1. This criterion is called the "expected value" criterion because it maximizes the "mathematical expected value" of a decision—or the net result of the benefits and costs that may be expected on average when this criterion is used for many decisions (Peterson, Birdsall, and Fox 1954; Green and Swets 1966). Specifically, if we multiply the benefit or cost of each of the four possible outcomes by the probability of that outcome (a/N, b/N, etc., in table 13.2), and add these four products, we obtain the expected (or average) value of the decision. (In this calculation, costs must be taken as negative.) It is desirable to maximize that value over many decisions because then the total benefit relative to the total cost is greatest.

Although a fair amount of algebra is required, which we shall bypass, it can be shown that the expected value is maximized when a criterion value of LR, that is, LR_c, is chosen such that

$$LR_c = \frac{P(S-)}{P(S+)} \times \frac{\text{benefit(TN)} - \text{cost(FP)}}{\text{benefit(TP)} - \text{cost(FN)}}.$$

That is, the expected value optimal criterion is specified by an LR_c defined as a ratio of the prior probabilities and a ratio of benefits and costs. (Because the costs are negative, one adds the absolute values of the benefits and costs.) How benefits and costs may be assigned to decision outcomes will be discussed later.

In this way, any set of prior probabilities, benefits, and costs determines a specific criterion value x_c, in terms of LR_c, that is best for that set of variables. It is best or optimal because it maximizes the payoff to the decision maker.

Note that the negative alternative is represented in the numerator of the probability part of the equation, $P(S-)$, and also in the numerator of

the benefit-cost part. Similarly, the denominators in both the probability and benefit-cost parts represent the occurrence of the positive alternative. Thus, for a fixed set of benefits and costs, the optimal LR_c will be relatively large—and the criterion will be relatively strict—whenever $P(S-)$ is appreciably greater than $P(S+)$. That is, one should not make the positive decision very readily if the chances are great that the negative alternative will occur. If, instead, the prior probabilities are constant and the benefits and costs vary, the optimal LR_c will be large and the optimal criterion will be relatively strict when the numerator of the benefit-cost part of the equation is large relative to its denominator, that is, when more importance is attached to being correct in the event the negative alternative occurs. Such might be the case when a surgical technique in question has substantial risks and its chances of a satisfactory outcome are low. Conversely, the optimal LR_c will be small and the optimal criterion will be lenient when the benefit-cost denominator is large relative to its numerator, that is, when it is more important to be correct when the positive alternative occurs. Such is the case in deciding whether to predict a severe storm.

13.2.2.6 A Traditional Measure of Acuity

If a decision criterion value x_c is placed midway between two distributions and the separation between the two distributions is increased, the TPP will get closer and closer to 1.0 while the FPP is getting closer to 0. Thus, as can be seen in figure 13.2, one way to measure discrimination acuity is to measure the difference, or distance, between the midpoints or means of the two distributions. In some tasks, the distributions are available themselves from data; in tasks for which they are not, this measure can be inferred from measured values of FPP and TPP. Usually, the difference between the means is divided by the standard deviation of one of the distributions—as a measure of the spread of the distribution—and therefore is measured in the units of that standard deviation. (See Wickens, chap. 12, this volume, for a discussion of the standard deviation of a distribution.) If the two distributions have the same standard deviation, which I have said in practice they usually do not, then this measure is called d' (read "d-prime"). I mention d' here because it is ingrained in psychological uses of signal detection theory. Indeed, the LR_c criterion value has been termed β ("beta") and the phrase "d-prime and beta" is used often as shorthand to signify separation of the discrimination and decision processes in psychology. Other discrimination measures have been defined and used that are variants of d', which are appropriate when the two distributions differ in spread. Section 13.3 defines a measure I believe to be preferable, and incidentally one that is commonly used in diagnostic tasks.

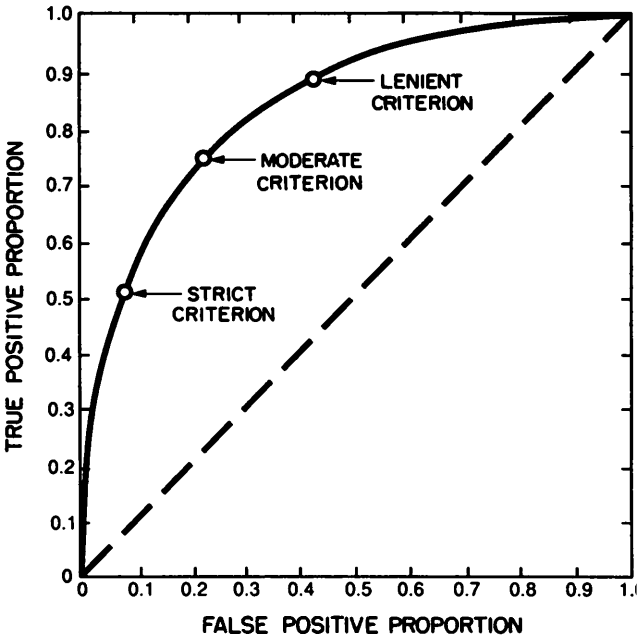


Figure 13.3
The relative operating characteristic (ROC), with three nominal decision criteria.

13.3 The Relative Operating Characteristic

A simple and compact, graphical way of quantitatively separating discrimination and decision processes is the "relative operating characteristic" (ROC). The ROC is a plot of TPP versus FPP, for a given acuity, as the decision criterion varies (Peterson, Birdsall, and Fox 1954). Given a strict (high) criterion corresponding to an x_c criterion toward the right side of figure 13.2, both proportions are near 0; given a more lenient (low) criterion corresponding to an x_c value to the left in figure 13.2, they both approach 1.0. Figure 13.3 shows an idealized ROC as a curve of decreasing slope running from 0 to 1.0 on each axis. It identifies three arbitrarily selected (TPP, FPP) pairs along the curve, each corresponding to a different possible decision criterion.

Note that if the ROC fell along the dashed, diagonal line, it would represent zero acuity: everywhere along this line TPP is equal to FPP, which is a result that can be obtained by pure guesswork or chance in choosing between categories A and B, without making an observation. As acuity increases, the ROC moves away from the diagonal, in a direction leftward and upward toward the upper left corner of the graph, where acuity is

perfect: TPP = 1.0 while FPP = 0. Hence a suitable measure of acuity will reflect the distance of the ROC from the diagonal line. For a particular level of acuity, a given decision criterion produces a particular data point along the ROC's curve and so a suitable measure of the criterion will represent the location of that point along the curve. The purpose of the ROC is to measure discrimination acuity independent of any decision criterion by displaying discrimination data at all possible decision criteria.

13.3.1 Obtaining an Empirical ROC

To obtain an empirical ROC for any observer or device at a particular level of acuity, sufficient data points (each corresponding to a different decision criterion) are obtained along the curve to fit the curve adequately, that is, to determine quite reliably just where the curve lies. Five points, each based on 100 or so observations, are usually thought to be sufficient. One way to obtain these points is to vary from one group of trials to another some variable or variables that will induce the observer to adopt a different criterion for each group of trials. Thus one could vary the prior probabilities of the stimuli A and B, or the benefits and costs (rewards and penalties) of the decision outcomes, or both, and define one ROC data point for each particular case. In such an experiment, the observer makes a choice between the two alternative responses, A and B in our general terms. In a signal detection problem, the response is either "yes" (a signal is present) or "no" (no signal is present) and the two-response method is often called the "yes-no method."

A more efficient way of obtaining data points on an empirical ROC is the "rating method," in which the observers rate their confidence (e.g., on a six-category scale) that stimulus A is present. Here, the observer chooses among multiple responses, six in this example, and the experimenter regards these different confidence responses as resulting from the simultaneous adoption of a set of different criteria. Specifically, if an observer establishes five different decision criteria, the decision variable is divided into six regions, each corresponding to one of the six possible responses. Then, in analysis, the experimenter treats different responses as representing different decision criteria. In figure 13.2, you could picture five vertical lines (corresponding to five values of x_c) spread across the decision variable. To illustrate, let us follow the convention that a rating of 1 indicates the highest confidence that alternative A is present, and 6, the lowest. Thus a rating of 1 corresponds to a very strict criterion (an x_c to the right in figure 13.2) while a rating of 6 corresponds to a very lenient criterion (an x_c at the left in figure 13.2.) The data analyst first takes only ratings of 1 to indicate a "yes" (or positive judgment) and calculates a data point (FPP and TPP) based on them. This point represents the strictest criterion used by the observer. Next, the analyst takes ratings of

both 1 and 2 to indicate a positive judgment and calculates a data point based on them. This is the second strictest criterion used by the observer. And so on for ratings 1, 2, and 3, ratings 1, 2, 3, and 4, and, finally, for ratings 1, 2, 3, 4, and 5—the rating of 6 is never included as a positive response, because then all responses would be positive and the trivial data point at the upper right corner (FPP = 1.0, TPP = 1.0) would result. In this way, progressively more rating categories are treated as if they were positive responses and progressively more lenient criteria are measured. In performing under the rating method, the observer adopts several criteria simultaneously—rather than successively in different groups of trials—and thereby provides an economy in data collection.

13.3.2 A Measure of the Decision Criterion

Although the ROC of figure 13.3 is shown with three possible decision criteria, in theory, the criterion can be set anywhere along the decision variable. Hence, if the decision variable is scaled finely enough to be essentially continuous, the ROC will be essentially continuous. One might, for example, ask an observer to make a probability estimate that stimulus A is present, that is, to use a 100-category rating scale, and thereby achieve an approximately continuous ROC.

As remarked earlier, a criterion corresponds to, and can be measured by, a value of likelihood ratio. Recall that the criterion value LR_c is the ratio of the heights of the two probability distributions at any given value of the decision variable, x_c (figure 13.2). It can be shown that the value of LR_c is also equal to the slope of the ROC at the data point that is produced by that criterion LR_c. (Strictly, it is equal to the slope of a line tangent to the ROC at the data point.) Thus very strict criteria to the right in figure 13.2, with high values of LR_c, produce a point on the ROC where the slope is steep—at the lower left of the graph. A moderate criterion, LR_c near 1.0, produces a point near the middle of the ROC. A very lenient criterion, to the left in figure 13.2, yields an ROC point where the slope is approximately flat, near 0, at the upper right. This slope measure of the decision criterion is denoted S (rather than LR_c) in our further discussion. Two illustrative values of S are illustrated in figure 13.4: $S = 2$ for a relatively strict criterion and $S = 1/2$ for a relatively lenient criterion. The calculation of the slope measure is illustrated with data in section 13.5.

13.3.3 A Measure of Discrimination Acuity

Figure 13.5 shows ROCs representing three possible degrees of discrimination acuity. As mentioned, the range of possible ROCs runs from the dashed diagonal line (where TPP = FPP and hence acuity is zero and decisions are correct only by chance) to a curve that follows the left and top axes (where acuity is perfect, TPP = 1.0 for all values of FPP). Thus it

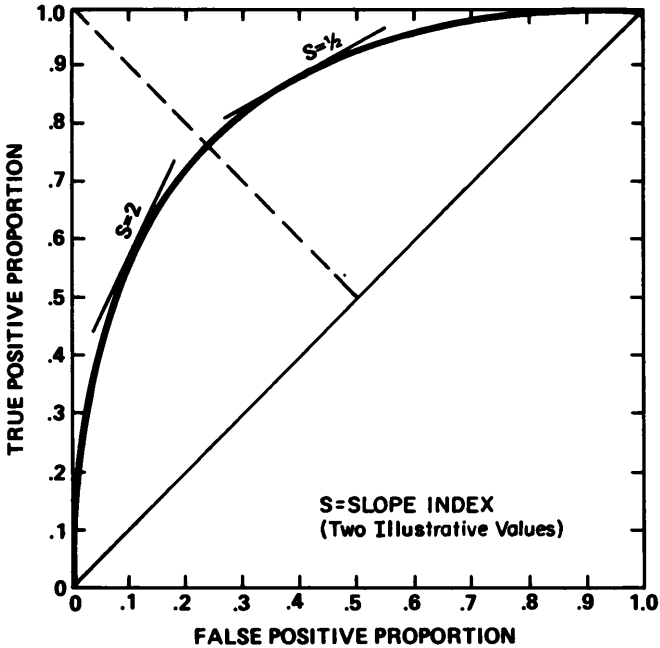


Figure 13.4
The relative operating characteristic (ROC) with two illustrative values of the slope measure of the decision criterion, *S*.

is evident that the proportion of the graph's area that lies beneath the ROC is a possible measure of acuity. This area measure ranges from a low value of .50 (for an ROC lying along the "chance" diagonal, where half of the graph's area is beneath the ROC) to a high value of 1.0 (for an ROC running along the left and upper axes of the graph and subtending all of its area). When normal distributions are assumed, as in section 13.2 above, the area measure is denoted A_z, because the units along the horizontal axis of the normal distribution are called "z scores." Figure 13.6 shows some illustrative values of A_z.

It may help to provide some intuitive grasp of the values of this area measure to know that it is equal to the proportion of correct choices in a "paired-comparison" task—in which alternatives A and B are presented *together* in each observation and the observer says which is which. For example, a radiologist might be shown two X-rays and would have to say which shows the malignant condition. If the radiologist were correct 95 percent of the time in such a paired-comparison task, the yes-no or rating method would yield the top curve ($A_z = .95$) in figure 13.6 (Green and Swets 1966).

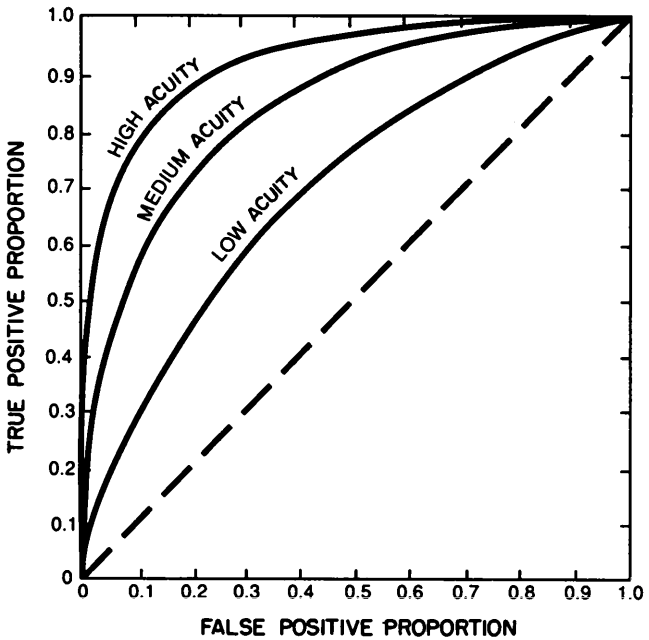


Figure 13.5
The relative operating characteristic (ROC) at three nominal levels of discrimination acuity.

A measure of acuity drawn from the ROC is independent of any decision criterion that might be adopted; it reflects all possible decision criteria. I emphasize now that such a measure is also independent of the prior probabilities of the two stimulus alternatives that may inhere in any particular situation because it is based on the quantities FPP and TPP, which are independent of the prior probabilities—as shown in table 13.2 and related text. For example, changing $P(S+)$ will change the column sum $a + c$ in table 13.2, but does not change TPP $= a/(a + c)$. Hence an ROC acuity measure is a general measure of the fundamental capacity for discrimination and it represents the full range of situations in which that particular discrimination might be called for. It is not specific to, or dependent on, any of one them.

13.3.4 Empirical Estimates of the Two Measures

I merely note that it is possible to obtain *graphical* estimates of the criterion measure S and of an area measure of acuity, as when successive data points in the ROC space are connected by straight lines. In this case, the slopes of connecting lines between points give the criterion measure S for each successive point (specifically, the slope of the line connecting two

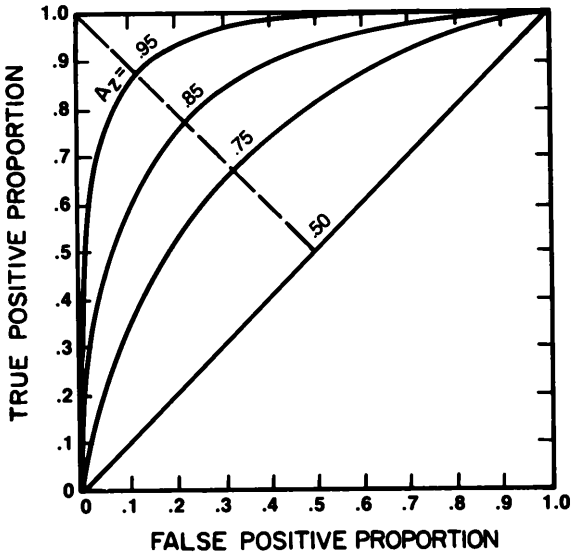


Figure 13.6
The relative operating characteristic (ROC) at some illustrative values of the measure of discrimination acuity, A_z.

points gives S for the higher point). The area beneath that ROC can be measured graphically by adding up the areas in the trapezoids formed by dropping vertical lines from the data points. In practice, however, a computer program is used to fit smooth curves to ROC data and to calculate the two measures along with estimates of their statistical variability (Swets and Pickett 1982).

It should be clear now how the ROC indicates observed changes in behavior to be discrimination effects or decision effects, or both. If the entire curve moves in the direction of the dashed diagonal line in figure 13.6, that is, moves to contain more or less area beneath it, then there is a discrimination effect, whether or not there is a decision effect as given by the measure S. If a data point (or set of points in the rating method) moves along a given curve, then there is a decision effect. The quantitative measures give the sizes of such effects.

13.4 Illustrations of Decomposition of Discrimination and Decision

There follow five examples in which the ROC analysis was used to separate the two behavioral processes of interest, two examples from psychology and three from practical diagnostic tasks.

13.4.1 Signal Detection during a Vigil

Outside the laboratory, signal detection must often be accomplished during a long observation period with infrequent signals occurring at random times. Such is the case in military contexts, where the problem may be to detect the approach of an enemy plane, and in industrial inspection, where the problem may be to detect defective products on an assembly line. Since such infrequent, random signals have been studied in the laboratory, beginning in about 1950, it has been found consistently that the true positive proportion falls off noticeably in only a half hour of observation—a finding suggesting that enemy planes and faulty products will likely be missed (Mackworth 1950; Mackie 1977). Investigators asked whether this finding could reflect a decrement in discrimination acuity due to fatigue or inattention, even though the time course of the decrement was very short. Hundreds of studies over several years examined variables thought to affect fatigue and alertness, including work-rest cycles, intersignal interval, irrelevant stimulation, incentives, knowledge of results, introversion-extroversion, temperature, drugs, age, and sex (Buckner and McGrath 1963; Broadbent 1971). At least five theories were proposed to account for a decrement in acuity (Frankmann and Adams 1962).

Then, led by James Egan (Egan, Greenberg, and Schulman 1961) and Donald Broadbent (Broadbent and Gregory 1963), investigators began to ask whether the drop in the TPP reflected instead a change in the decision criterion. Might observers be setting a stricter criterion as time progressed—that is, requiring stronger evidence to say that a signal was present—conceivably because their estimate of the prior probability of signal occurrence was going down as they experienced signals at a lower-than-expected rate? If so, the FPP (as well as the TPP) should be decreasing over time. False positive responses were then tallied and found to be in fact decreasing. Another hundred or so studies have since shown that for most stimulus displays, the principal, and sometimes only, change over a vigil is in the decision criterion.

Under some rather special conditions, changes in acuity are regularly found, as well as changes in the criterion. For example, a change in acuity is found when a discrimination must be made between two stimulus alternatives presented successively, which puts a greater demand on memory than a discrimination between simultaneous stimuli. A high rate of stimulus occurrence also produces a decline in acuity, perhaps because it requires continuous observation, which may be difficult to maintain (Parasuraman and Davies 1977).

Given the tendency for the criterion to change as the main effect, however, task design has shifted to controlling it: Military commanders would like their observers to use consistently a criterion that reflects realistically

the prior probability of attack and the benefits and costs of alternative decision outcomes; similarly, manufacturers would like their inspectors to employ a criterion that satisfies management's objectives for the quality of the product (Swets 1977).

13.4.2 Recognition Memory

In a typical laboratory recognition memory task, the subject is asked to say whether each of a series of items (e.g., a word) was presented before ("old") or not ("new"). Applying signal detection theory to obtain a pure measure of "memory strength," analogous to acuity, depends on the assumption that all items lie along a continuum of strength, with the strength of each item being determined by conditions of memorizing and forgetting (Murdock and Duffy 1972). The subject sets a decision criterion on the strength continuum to issue a response of "old," and the investigator attempts to separate actual phenomena of memory from decision or response processes that may vary within and across subjects for reasons independent of memory (Egan 1958).

Several experimental effects that were presumed to reflect memory strength were later shown to be effects of the decision criterion instead (Swets 1973). These effects include the apparently better recognition of more common words—in fact, there is evidence that recognition memory is better for uncommon words (Broadbent 1971); the differences in recall of familiar and unfamiliar associations (McNicol and Ryder 1971); the buildup of false responses of "old" during a continuous recognition task (Donaldson and Murdock 1968); effects of interpolated learning (Banks 1969); changes in semantic or association context from acquisition to recall (DaPolito, Barker, and Wiant 1971); and gender differences (Barr-Brown and White 1971). Effects due both to memory and decision processes were found for meaningfulness of items (Raser 1970), serial position (Murdock 1968), and the similarity of distractor items (Mandler, Pearlstone, and Koopmans 1969). A study of elderly patients, including both demented and depressed individuals, showed that an apparent memory loss that seemed similar for both types of patients was, in fact, a true memory impairment for demented individuals, but rather a criterion (confidence or caution) effect for depressed individuals (Miller and Lewis 1977).

13.4.3 Polygraph Lie Detection

Attempts to detect individuals who are lying about certain events—by examining the various physiological measures (such as heart rate) that are taken by a "polygraph" machine—are increasingly being made both in court cases and in employment settings where security is important (Saxe,

Dougherty, and Cross 1985). According to a recent article, the need to separate discrimination acuity and the decision criterion is simply not appreciated in this field:

> Despite lengthy congressional hearings conducted during the preparation of the Employee Protection Act of 1988, no one ever explained the distinction between *accuracy* [acuity] and a *decision criterion* to the legislators. Thus, the policy makers never learned about a polygrapher's predilection to err on the side of false positives or false negatives, constancy in the location of his or her decision criterion, or ability to change his or her decision criterion. Individual differences among polygraphers with respect to their decision criteria thus remain unknown, and persons appearing before a polygrapher are—unknowingly—up against the "luck of the draw." (Hammond, Harvey, and Hastie 1992, 84).

Studies have shown polygraphers to have widely different acuities and also quite different decision criteria for accusing a person of lying (Shakhar, Lieblich, and Kugelmass 1970; Szucko and Kleinmuntz 1981). However, there is a general tendency toward a lenient criterion for concluding that an individual is lying, probably because accusations hold a chance of eliciting a confession. In this respect, polygraphers may care less about their technique's acuity if it is effective in eliciting a confession now and then. As a consequence, the ratio of persons falsely accused to those truly accused is often quite high—in some studies, as high as 20 to 1 (Szucko and Kleinmuntz 1981). As attempts are made to expand the domain of polygraphy, we can ask, should the criterion for accusation be different for security screening in the workplace than for criminal cases? Specifically, how do the prior probabilities, and especially the benefits and costs, differ? Polygraphers find it rather easy to change their decision criteria; as in mammography, they look for certain perceptual features of the recording of physiological variables and can adjust the decision criterion explicitly by requiring the presence of more or fewer of these features.

13.4.4 Information Retrieval

Conventional library systems manually retrieve documents from shelves and facts from documents by means of familiar card indexes based on cataloguing methods. Computer-based retrieval systems scan documents electronically, looking for various properties of the text—for the appearance of key words, say—and give the documents a relevance score to represent how likely they are to contain wanted information. In both cases, a decision criterion must be established: how likely must the document be to satisfy the specified need for information in order to proceed

to retrieve it for further examination? The decision criterion might be specified for computer systems in terms of the number of key words a document must contain or in terms of how high its relevance score must be. Here again, librarians considering alternative systems for use should know the relative acuities of the systems—how well they separate the wheat from the chaff—independent of any particular decision criterion. Users will want to set a criterion to individual likings when the system is in operation. In examining the retrieval performance of several alternative systems with the relative operating characteristic (ROC) technique, I found that their various approaches to the scoring of relevance led to only negligible differences in acuity (Swets 1969). As examples of different approaches, one system might consider only specified key words, whereas a second might also consider synonyms of the specified key words; a third system might scan the full text, and a fourth, just the abstract.

Incidentally, the typical acuities found suggest that the retrieval problem is more difficult than we might have thought. Consider a file of 3,000 documents and assume that for each query to the file 10 documents are actually relevant. To retrieve 3 of the 10 relevant documents, one must accept, on average, 3 false positives, that is, 3 unwanted, irrelevant documents. To retrieve 6 of 10 relevant items, a more lenient criterion must be effected, and then 30 false positives will occur. A user desiring to retrieve 9 of the 10 relevant documents will have to set a very lenient criterion and then will have to winnow the 9 relevant documents from the approximately 300 retrieved, some 290 of which will be just "noise."

13.4.5 Weather Forecasting

Ian Mason (1982a) first applied signal detection theory to predicting various kinds of weather events. Weather forecasting represents a diagnostic problem similar to information retrieval, medical diagnosis, materials testing, and so forth, in that we want to measure acuity independent of the decision criterion when alternative systems (or "models") for forecasting are being evaluated prior to selecting one of them for use, and then want to adjust the criterion to an appropriate level when the selected system is operating in actual practice. Mason showed that predictions of various kinds of weather give data that are fitted well by the ROC. He showed further (1982b) that the several existing (non-ROC) measures of acuity used in this field confound discrimination acuity and the decision criterion and hence give unreliable measures of acuity.

Weather forecasters have recognized the operational problem of setting decision criteria in recent years by giving probability estimates for weather events, thus permitting the forecast user to set his or her own decision

criterion. The orange grower can decide, in terms of personal costs and benefits, whether the likelihood of frost is high enough to set out the smoke pots. You and I can decide whether or not to carry an umbrella.

13.5 Computational Example of Decomposition: A Dice Game

A deep understanding of the theory discussed above comes quite easily by way of a game (Swets 1991a). In the game, three dice are thrown repeatedly. Two are ordinary dice with 1 through 6 spots on the six sides; the third die has 3 spots on each of three sides and blanks on the other three sides. For each throw, you are given the total number of spots showing—your observation—and you must say whether the unusual die showed a 3 (positive event) or a 0 (negative event). You win $1 on true positive and true negative outcomes and lose $1 on false positive and false negative outcomes. I assume that you will set a decision criterion at some particular number of spots and consistently respond "3" when and only when that number is met or exceeded. The optimal criterion is the one that maximizes your payoff and you will want to determine that number.

13.5.1 Distributions of Observations

Consider the initial steps in determining the optimal decision criterion. You need first to calculate the distributions of observations under each stimulus alternative: the probability of each of the possible totals 2 through 12 when the third die shows a 0 and the probability of each of the possible totals 5 through 15 when the third die shows a 3. So you construct table 13.6 to show the number of different ways in which each total 2 through 12 can occur on two ordinary dice—relevant to throws on which the third die shows a 0. Reading along the positive diagonals from lower left to upper right, you note that a total of 2 can occur on the two ordinary

Table 13.6
Possible throws when third die shows

		Number of spots showing on first die					
		2	3	4	5		
Number of spots showing on second die	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Table 13.7
Possible throws when third die shows 3

		Number of spots showing on first die					
		1	2	3	4	5	6
Number of spots showing on second die	1	5	6	7	8	9	10
	2	6	7	8	9	10	11
	3	7	8	9	10	11	12
	4	8	9	10	11	12	13
	5	9	10	11	12	13	14
	6	10	11	12	13	14	15

dice in only one way (1, 1), a total of 3 can occur in two ways (1, 2, and 2, 1), and so on. Because there are 36 different kinds of throws possible, the probability of a 2 (which can occur one way) is 1/36 (or .028), the probability of a 3 (which can occur two ways) is 2/36 (.056), and so forth. The appropriate table for the throws on which the third die shows a 3 is obtained by adding 3 to each cell entry in table 13.6; these values are shown in table 13.7.

You will need table 13.8, which lists the *number of ways* in which each total can occur when a 0 shows (column 2) and when a 3 shows (column 3) and also the *probability* of each total given a 0 (column 4) and given a 3 (column 5). Just as the numbers of ways a given total can occur (for 0 and 3, respectively) add up to 36, so the probabilities (for each of the two types of throw) add up to 1.

Now you can construct the two probability distributions—as histograms because your observations are of integer values only; that is, they are discrete rather than continuous. These distributions are shown in figure 13.7, where the distribution for throws of 0 is on the left (dashed line) and the distribution given a 3 is on the right (solid line). In earlier terminology, the totals of the two ordinary dice constitute the "noise" of signal detection theory, and the third die either adds a signal to the noise or does not. This dice game provides more surety than many more realistic tasks in that certain observations very clearly and definitely come from one or the other stimulus alternative. Specifically, in this case, there is no uncertainty or decision problem for the totals 2 to 4 (which definitely indicate a 0) or the totals 13 to 15 (which definitely indicate a 3).

Following up on our earlier discussion of continuous probability distributions in figure 13.2, picture vertical lines in figure 13.7 extended from each tick mark on the horizontal axis up to the dashed and solid lines, so that each total of the three dice (under each of the alternatives 0 and 3) is represented by a vertical bar. If the bars are assumed to have a width of

Table 13.8
Probabilities of the various totals, the likelihood ratio, and the ROC coordinates.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total spots showing	Ways if 0	Ways if 3	Probability if 0	Probability if 3	Likelihood ratio	FPP	TPP
2	1	0	.028	.000	0	1.0	1.0
3	2	0	.056	.000	0	.973	1.0
4	3	0	.083	.000	0	.917	1.0
5	4	1	.111	.028	.25	.834	1.0
6	5	2	.139	.056	.40	.723	.973
7	6	3	.167	.083	.50	.584	.917
8	5	4	.139	.111	.80	.417	.834
9	4	5	.111	.139	1.25	.278	.723
10	3	6	.083	.167	2.00	.167	.584
11	2	5	.025	.139	2.50	.084	.417
12	1	4	.028	.111	4.00	.028	.278
13	0	3	.000	.083	∞	0	.167
14	0	2	.000	.056	∞	0	.084
15	0	1	.000	.028	∞	0	.028

Note: Columns 2 and 3, respectively, are the numbers of ways each total number of spots as listed in column 1 can occur when a 0 is thrown, and when a 3 is thrown, on the third die; columns 4 and 5 are the corresponding probabilities for each total listed in column 1; column 6 is the likelihood ratio for each total listed in column 1, namely, the ratio of column 5 to column 4; the FPP in column 7 is the probability of observing a total number of spots that is at least as great as the value in column 1, given that a 0 shows on the third die; similarly, the TPP in column 8 is that probability given a 3 showing on the third die.

one unit, then the area of a bar (as well as its height) reflects the probability of its corresponding total of the three dice. Because the probabilities for each histogram add up to 1, the total area under each histogram equals 1. Hence we can add the areas of the bars to the right of a decision criterion, for the left and right histograms, to give the probabilities FPP and TPP, respectively. This operation may help you to understand the way areas yield FPP and TPP for the continuous distributions of figure 13.2.

13.5.2 The Optimal Decision Criterion for the Symmetrical Game

What total of the three dice do you choose for your decision criterion? The answer is straightforward for this symmetrical game—"symmetrical" in the sense that the chances of a 0 and 3 are equal (there are three of each on the third die) and all benefits and costs are equal. Here we need only to determine which totals are more likely when a 3 occurs on the third die

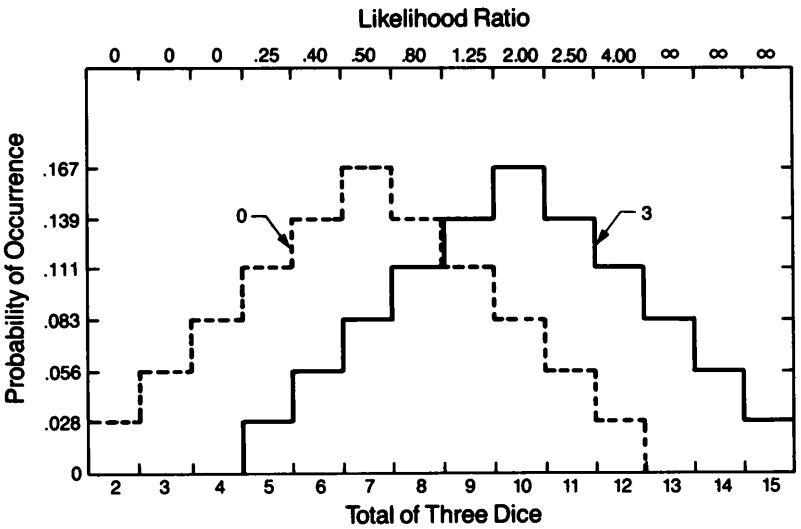


Figure 13.7
Histograms of the probabilities of occurrence of the totals of three dice (for the odd die showing 0 on three sides and 3 spots on three sides), with the likelihood ratio for each total.

than when a 0 occurs and, conversely, which totals are more likely for a 0 than a 3. According to table 13.8, totals of 9 and greater are more likely for a 3 than for a 0, and totals of 8 or less are more likely for a 0 than for a 3. If you think of the optimal criterion then as 8.5, you can note in figure 13.7 that it occurs exactly in the middle of the pair of histograms, that is, where they "cross" each other. This result is intuitively proper for a symmetrical game.

Would you rather play the game when the third die has 4 spots on three sides instead of 3 spots? For this new version of the game, the histograms are further apart and so your acuity will be higher, and the decisions should be better and produce a higher payoff. Specifically, the right-hand histogram moves one unit to the right. (Now the total of 5 is a sure sign of a 0, along with totals of 2, 3, 4, and the new possible total of 16 is a sure sign of a 4, along with totals of 13, 14, 15.) Because the optimal criterion also moves one unit to the right, to 9.5, here you say "4" for totals ≥ 10.

13.5.3 The Optimal Decision Criterion in General

Returning to the first version of the game, with 0 or 3 spots, what now if the third die shows 3 spots on four sides, rather than three? Or on two sides rather than three? How do such changes in the prior probabilities affect the optimal criterion? Given our earlier discussion (in section

13.2.2.5), you respond that having 3 spots on four sides leads to a more lenient criterion, whereas having 3 spots on two sides leads to a stricter criterion (than 8.5).

Further, consider a change in benefits and costs: say that the TN and FP outcomes (possible when the third die shows a 0) remain at +$1 and −$1, respectively, but that the TP and FN outcomes (possible when the third die shows a 3) become +$5 and −$5, respectively. Because ±$1 applies when a 0 occurs on the third die and ±$5 applies when a 3 occurs, it is more important to be correct when a 3 occurs, and you will want to say "3" more often, by setting a lower criterion number of total spots (lower than 8.5). To determine just what that criterion number is, we need to employ the formula for the optimal, expected value criterion presented earlier in terms of the likelihood ratio criterion, LR_c, in section 13.2.2.5. It is an "expected value" (EV) criterion you want because you want to maximize your average payoff over many plays of the game and that is what the EV is.

Again, to avoid a fair amount of algebra, we shall go through the operations of this example only in summary. First you multiply the probability of each outcome by its value (benefit or cost) to obtain its EV, and then you add the two EVs for each possible decision (0 or 3) to get each decision's EV. At this point, were you to make some algebraic substitutions and collect and rearrange some terms, you would find that if you choose 3 for those totals having a higher EV for 3 than 0 (and choose 0 for the other totals), you will be saying "3" whenever the *LR* of a total is greater than LR_c, where

$$LR_c = \frac{P(0)}{P(3)} \times \frac{\text{benefit (TN)} - \text{cost(FP)}}{\text{benefit(TP)} - \text{cost(FN)}},$$

which is identical to the formula given in section 13.2.2.5. (As noted before, a cost has a negative value; hence the absolute values of benefit and cost are added.) For the game at hand, where the benefits and costs are ±$1 and ±$5, you say "3" when the *LR* exceeds

$$\frac{.5}{.5} \times \frac{1+1}{5+5} = \frac{2}{10} = .20.$$

13.5.4 The Likelihood Ratio

To use this result, you need to know the value of the *LR* for each total number of spots. Table 13.8 shows the values of the *LR* in column 6; they are the ratios of values in column 5 (probability of a given total if 3 occurs) to those in column 4 (probability of a given total if 0 occurs)—just as we defined the *LR* earlier in terms of heights of continuous distributions (section 13.2.2.4). You note in table 13.8 that with an *LR* criterion of .20,

you say "3" for totals of 5 or more—as it happens here whenever the probability of 3 exceeds zero.

The *LRs* listed in table 13.8 are printed along the top of figure 13.7 where they indicate the ratios of heights of the histogram bars (the ratio of 3's heights to 0's heights) at each three-dice total. If you draw a vertical line there, to represent $LR = .20$, at the tick mark between 0 and .25, and extend it down to the bottom axis, you see again that an *LR* criterion of .20 leads to saying "3" whenever the total is 5 or more (greater than 4.5).

13.5.5 The Dice Game's ROC

To construct the game's ROC, we must know the probabilities for TP and FP for each decision criterion. I designate these probabilities as I did their proportions, TPP and FPP. We can calculate the relevant probabilities by applying the ideas used in the rating method introduced in section 13.3.1. They are shown in columns 7 and 8 of table 13.8. Note that in this table, the criteria from top to bottom (2 spots to 15 spots) run from lenient to strict, and thus, to maintain the convention of section 13.3.1, where we started cumulating probabilities at the strictest criterion, we begin cumulating from the bottom of the table. For illustration, consider first the probabilities of various numbers of total spots showing when a 0 occurs on the third die—in column 4—and the corresponding values of FPP for criteria set at those different numbers of total spots—in column 7. Reading from the bottom of column 4, we see that the probabilities of total spots equaling 15 or more, 14 or more, and 13 or more are all .000. For 12 or more spots, the probability in column 4 is .028, and thus the value of FPP in column 7 is .028. For the criterion at 11 or more spots, we begin to cumulate; we add the probability of .056 in column 4 to the probability of .028 in column 7 to determine the value of .084 in column 7. Moving up, the probability of .083 in column 4, for a criterion of 11 or more spots, is cumulated with the .084 in column 7 to show .167 in column 7, for a criterion of 10 or more spots.

Look next at TPP in column 8, which is achieved by cumulating over the probabilities of a given number of total spots showing when a 3 occurs on the third die, as shown in column 5. For 15 or more total spots, the probabilites in both column 5 and column 8 are .028. Move up in column 5 to the next row and cumulate .056 with the .028 in column 8 to calculate the TPP of .084, for a criterion set at 14 or more total spots showing, and so on.

In this way, by using the rating method described in section 13.3.1, we determine values of FPP and TPP for each possible decision criterion in the dice game, and thus can plot the game's ROC. Plotting ROC points at the FPP and TPP coordinates given in columns 7 and 8 of table 13.8 gives points marked by *x*s along the ROC shown in figure 13.8. Thus, for exam-

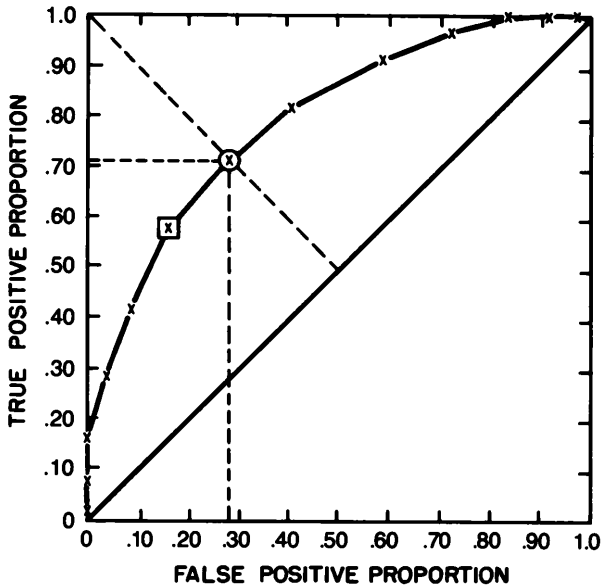


Figure 13.8
Relative operating characteristic (ROC) for the dice game.

ple, the symmetrical criterion at 8.5 total spots yields the point FPP = .278 and TPP = .723 as circled in figure 13.8. We can see that this criterion is symmetrical because its point lies at the negative diagonal of the figure where the error probabilities are equal: FPP = .278 and the false negative proportion FNP is 1 − TPP or 1 − .723 = .278 (within rounding error).

Note that the slope measure S of the criterion is obtained from the slope of the line connecting that criterion's data point and the data point to its left. For example, the *LR* criterion that yields the circled point is 1.25, corresponding in table 13.8 to a total greater than 8.5. And the slope of the line connecting the circled point to the squared point is 1.25. Recall that a slope, here S, is calculated as the ratio of the increment on the y-axis (Δy) to the increment on the x-axis (Δx). Working from table 13.8, $\Delta y = .723 - .584$ or .139 and $\Delta x = .278 - .167$ or .111, and hence the ratio $\Delta y/\Delta x = .139/.111$ or 1.25.

13.5.6 The Game's Generality

The dice game makes concrete the random distributions of observations that occur for the two different stimulus alternatives. As described so far, the game's two distributions have the same spread. The "signal plus noise" distribution has the same variation as the "noise alone" distribution,

because the signal (a throw of 3 on the third die) adds a constant value (i.e., 3) to the variable noise-alone observation (the total of the two ordinary dice). However, as mentioned, data in most detection, recognition, and diagnostic tasks yield distributions with different spreads. A small change in the dice game serves to represent this case. For example, consider a version of the game in which the third die has 2 spots on one side, 3 on a second side, and 4 on a third side. (Your task is still to say whether the third die shows spots or a blank.) In this case, there is variation in the signal in addition to that in the noise-alone distribution, and the signal-plus-noise distribution will have a larger spread than the noise distribution. In particular, the spread of totals when the third die shows some spots will range from 4 to 16, instead of from 5 to 15.

The dice game is a general representation of detection, recognition, and diagnostic tasks in other respects as well, as noted in the preceding discussion. Thus it illustrates calculation of an ROC based on the multiple responses of the rating method, the adoption of a decision criterion in terms of the likelihood ratio, the concept of an optimal (expected value) criterion, and the measurement of that decision criterion by the slope of the ROC at a particular point. An area measure of discrimination acuity is appropriate.

13.6 Improving Discrimination Acuity by Combining Observations

This brief section is included because it is basic to applications of signal detection theory and, specifically, is central to the case study of diagnosing breast cancer considered in the following section. The idea is that the more observations that enter into each decision, the greater will be the acuity afforded by those decisions (Green and Swets 1966).

Just how much discrimination acuity will increase as more observations are considered in each decision depends heavily on how *correlated* or redundant the observations are. Two diagnostic systems that are completely redundant—that is, are always in complete agreement—clearly offer no more acuity together than either one alone. At the other extreme, diagnostic systems that provide totally different information give the greatest potential for increased acuity. As an example, you might expect that repeated mammograms of a given patient taken by the same machine and same technician and read by the same radiologist would provide more correlated information than ones taken and read under diverse conditions. Pursuing this example, you would expect that a physical examination of the breast and a mammogram would provide less correlated information than two mammograms. The same is true for a computerized tomography or ultrasound scan of the breast along with a mammogram, relative to two images taken by one technique. These three techniques

differ physically in the way anatomical and physiological information is acquired and displayed, and are correspondingly uncorrelated.

The largest increase in acuity that can result from combining observations is when the systems that produce them are completely independent, or when the variability of a given system makes successive observations from it completely independent. According to statistical theory, acuity is then proportional to the square root of the number of systems considered or the number of times a sample of observations is replicated. Hence a doubling of systems or samples increases acuity by $\sqrt{2}$, or 1.4. Thus a substantial increase in acuity (40 percent) results from doubling the number of systems or the sample size (Green and Swets 1966).

We would like a way to predict the gain in acuity that results from observations correlated to any specified degree. Then, in any particular setting with any amount of measured or estimated correlation among observations, we could determine whether the likely gain in acuity would outweigh the cost of acquiring the additional observations. For example, we could anticipate the gains in acuity of having mammograms read twice by the same radiologist or by different radiologists. We further wish to handle observations that are more or less informative in their own right, that is, observations that by themselves afford more or less acuity of discrimination. In sum, if we have two observers looking for a given signal, we wish to predict their combined acuity if one observer is partially redundant with the other and one is inherently not as acute as the other, by particular amounts.

Theory exists to make such predictions (Metz and Shen 1992) and figure 13.9 is presented here to illustrate one of the variables, namely, the effect of correlation—but not the other variable, namely, the effect of different acuities because that is too complicated to show. The figure shows along the horizontal axis the factor by which the number of observations is increased—moving, for example, from one to two to three to four observers, diagnostic tests in a clinic, or perceptual features in a mammogram. On the vertical axis is shown the theoretical prediction of percent gain in acuity. Separate curves are shown for six degrees of correlation (as indicated by r, the correlation coefficient) ranging from zero correlation (i.e., independent observations) to perfect correlation (i.e., additional observations provide no additional information). When there are more than two observations, the correlations listed are averages for all possible pairs of observations.

It can be seen that predicted acuity increases sharply for uncorrelated observers or tests (where $r = 0$), by 100 percent when increasing from one to four. Any correlation between observers or tests or features acts to reduce the possible gain in acuity from the additional three observations; for example, for $r = .20$, the maximum gain is about 50 percent. For high

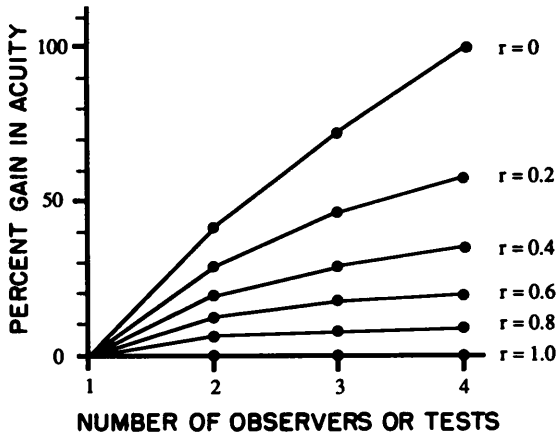


Figure 13.9
Obtainable gain in discrimination acuity from additional observers or tests as a function of their intercorrelations (from Metz and Shen 1992).

correlations, additional observers or tests or features are hardly worth the costs of employing them.

Data my colleagues and I obtained in a mammographic setting are very close to these predictions (as reported by Metz and Shen 1992). For example, on the scale of figure 13.9, and with an average correlation between pairs of observers of .55, the gain in acuity from using two observers rather than one was 10 percent in one test and 12 percent in another. These differences in acuity correspond to differences in A_z of .024 and .029, and to differences in TPP at a value of FPP = .10 of .064 and 0.84.

These results are relevant to the study discussed in the next section, in which the objective is to increase the acuity of mammography in detecting breast cancer. With an imaging test such as mammography, radiologists must know which perceptual features to look for, how much the acuity of diagnosis increases with the number of perceptual features observed, and how to merge the observations of features in a way that is sensitive to their correlations and their relative contributions to acuity. Radiologists should not, for example, "double count," that is, give full weight to a feature that contains information fully redundant with the information of another feature previously taken into account.

13.7 Enhancing the Interpretation of Mammograms

We consider in this section some ways of enhancing the interpretation of X-ray mammography images in the diagnosis of breast cancer. This study

was carried out with my psychologist colleagues David Getty and Ronald Pickett, and with radiologist and mammography specialist Carl D'Orsi. It shows how the radiologist's discrimination acuity can be increased and illustrates the types of factors that are involved in choosing an appropriate decision criterion for a positive judgment. It has implications for how performance and outcomes can be substantially improved in a wide range of diagnostic tasks of importance to individuals and society (Getty et al. 1988).

The practical significance of enhancing discrimination acuity in mammography is indicated by the fact that approximately 175,000 new cases of breast cancer were diagnosed last year in the United States. Further, early detection can reduce the number of deaths caused by breast cancer, now standing at approximately 45,000 each year. On the other hand, a large number of breast biopsies is performed on patients without cancer; as mentioned earlier, that proportion is approximately 7 or 8 out of every 10 breast biopsies done in the United States (counting all breast biopsies done for whatever reason). Thus large benefits would accrue from increasing TPP and decreasing FPP (D'Orsi et al. 1992).

13.7.1 Improving Discrimination Acuity

In mammography as practiced today, the "error rate" is thought to be about 15 percent, meaning that the two error proportions FPP and FNP = .15 (so TPP = .85). To increase discrimination acuity beyond that level, we developed two aids for the radiologist. The first is an aid to perception or discrimination and is here called a "reading aid." This aid is a checklist of the types of observations the radiologist should make in reading every image—or of the various perceptible features the radiologist should attend to, as discussed earlier in section 13.1.2.1. Each perceptual feature on the list is accompanied by a numerical scale that elicits from the radiologist a quantitative judgment about the extent to which it occurs, or the radiologist's confidence that it is present, or a measurement by ruler. The scale permits a more finely graded representation of perceptual information than the binary (present or absent) judgment that is typically made.

The second aid, a "merging aid," is a computer program that merges the several scale values for the features of each case into the single value they imply of the final decision variable, namely, the probability of cancer. In our earlier terminology, the computer supplies a value of the decision variable x that the observer can compare to a decision criterion value x_c. In the merging, the program takes into account the relative importance or relative "weight" that each feature should have in estimating the overall probability of cancer. That weight depends on how diagnostic the feature

is and also on the degree to which it is correlated with other features, as discussed in section 13.6. Radiologists can use the computer's probability estimate as advisory information in making their own estimate.

The intent of this approach is to retain for the human observer a task that computers do not (at present) do well: detecting and recognizing subtle perceptual aspects of an image—and to delegate to the computer a task that humans do not do as well: remembering and optimally combining several pieces of quantitative information. Several previous studies have shown that an explicit formula for combining information gives greater diagnostic acuity than an individual's intuitive combination (Davies and Corrigan 1974; Ackerman and Gose 1972; Gale et al. 1987). Our approach further compensates for human frailties by supplying the checklist, which forces the mammogram reader to be comprehensive in every image viewed. In this study, we did not separate the magnitudes of the effects of the two aids, which would require the reader to give an overall confidence judgment before as well as after learning the computer-based estimate of cancer. Another study we made of the interpretation of magnetic resonance images of the prostate gland, however, was designed to separate the two effects (Seltzer et al. 1997).

Before these two aids could be specifically designed, we needed to determine carefully the perceptual features that should be included, and their relative importances in diagnosis. In our approach, we begin with the advice of mammography experts and proceed to certain perceptual tests and statistical analyses. Our approach is systematic, as contrasted with the usual, almost haphazard, way in which several separate investigators publish articles over a span of years about specific features they believe to be useful in medical imaging. Our approach is carried out in a short time, important in an era in which new medical imaging techniques appear at a rapid rate and are soon modified, and does not await an individual's motivation to collate and organize individual features in a review article or book.

13.7.1.1 Determining Candidate Perceptual Features

We took two steps to determine which perceptual features should be examined as candidates for the final list to be implemented in the two aids. The first step was a very direct approach. It consisted of a series of interviews with five specialists in mammography in which they were asked to mention all possibly useful features. The intent was to create an exhaustive list, so that no feature of possible relevance was overlooked; this step yielded about 50 features. In several instances, no doubt, the same feature appeared with different names and noticeably different features probably occurred that were highly correlated or conveyed a good deal of redundant information.

The second step was a highly indirect approach, where the mammography specialists made quantitative, nonverbal judgments, as described below, about what they saw in a representative set of mammograms and these judgments were analyzed by a statistical procedure to reveal the perceptual features on which they were based. The objective of this second step was to reveal any features possibly affecting the radiologists' perception that were not previously verbalized and also to reduce and refine the list of features that came from the interviews.

Specifically, in the second step, 24 mammograms were selected to represent typical variations in both cancerous and normal cases and the five mammography specialists (and three psychologists specialized in perception) made judgments about the visual similarity of each of the 276 possible pairs of the 24 mammograms, on a 10-point scale. To these judgments, we applied a statistical analysis—called "multidimensional scaling" (MDS)—that assumes that the degree of judged similarity between two mammograms reflects the distance between them in a hypothetical perceptual (Euclidean) space of several dimensions (Shepard 1964; Shiffman, Reynolds, and Young 1981). In this space, similar mammograms cluster together and others diverge in various directions (and to various extents) depending on the nature (and extent) of the perceived dissimilarity. The MDS analysis determines the several dimensions of the space implied by the total set of similarity judgments and locates each mammogram in that space; that is, it gives the coordinate value of each mammogram on each dimension. We investigators then regarded these dimensions as candidate features for the checklist.

The concepts and techniques of multidimensional scaling, and of similar indirect approaches to determining features, cannot be discussed here in any detail. To appreciate something of this approach, you can imagine that a set of ratings of the distances between all U.S. state capitals could be analyzed to produce a familiar map, in two dimensions, of the United States. (The point of using ratings of distances, rather than actual distances, in this state capital analogy is to reflect the statistical variation in human judgments of similarity.) The MDS analysis of distances between state capitals would produce the east-west and north-south dimensions and give the coordinate values (here, x and y) of each capital on these dimensions. A similar, two-dimensional analysis of mammogram similarities/distances would show how two of the main features of mammograms (e.g., irregular shape and irregular border of a mass) serve to place all mammograms in a two-dimensional space. Those with high values on both dimensions (high x and y, upper right quadrant) would likely be malignant and those with low values on both dimensions (lower left quadrant) would likely be benign. Similarly, a table of straight-line distances between all pairs of a set of stars could be used to construct

a three-dimensional space—a volume in which each star could be located at its position relative to the others.

MDS analysis generally yields up to four to five dimensions for a given set of ratings. In the present case, each reader made three similarity ratings for each of the 276 pairs of mammograms in the selected set, one for each of the three major qualitative categories of mammographic features (masses, calcifications, and secondary signs). With four or so dimensions produced by each of the three sets of ratings, the total analysis produced about a dozen dimensions, and hence a dozen potential features. The intent of using a judgment of similarity, that is, a global judgment about image features, within a qualitative category, is to permit intuitive and unlabeled aspects of the image to have an influence.

A virtue of MDS analysis is that it yields independent dimensions, that is, the perceived value of one potential feature (for a given mammogram) is not affected by, and hence not correlated with, variation in the perceived value of another potential feature (as scaled for that mammogram.) Conversely, the features given by the specialists directly in interviews are not assuredly independent. Also, the MDS analysis assigns weights to the dimensions according to how well they represent the total set of similarity judgments; those weights are an indication of the dimensions' potential importances as diagnostic features.

13.7.1.2 Reducing the Set of Features and Designing the Reading Aid

With the MDS results in hand, the specialists were consulted as a consensus group to help select a set of the most promising features. The original fifty features they verbalized were available for their consideration along with the features suggested by the MDS dimensions. The intent was to select a set of features of small and workable size, while keeping all features that were likely to be more than minimally predictive of malignancy. The further intent was to determine what to name those features and how to scale them.

To investigate the features possibly arising from the MDS analysis, the twenty-four mammograms the specialists had judged were arrayed in order for them along each perceptual dimension in turn, according to the mammograms' coordinate values on the given dimension. So the specialists could see and discuss what perceptual property seemed to be varying on each dimension and agree on what it was best called. Taking some examples from figure 13.1, the two masses in figures 13.1a and 13.1c would appear at the two ends of a particular array, which could be called "roughness/smoothness of border." The calcifications in figures 13.1a and 13.1b would appear at the ends of other arrays, which could be called, respectively, "clustering of calcifications," "size of calcifications," and

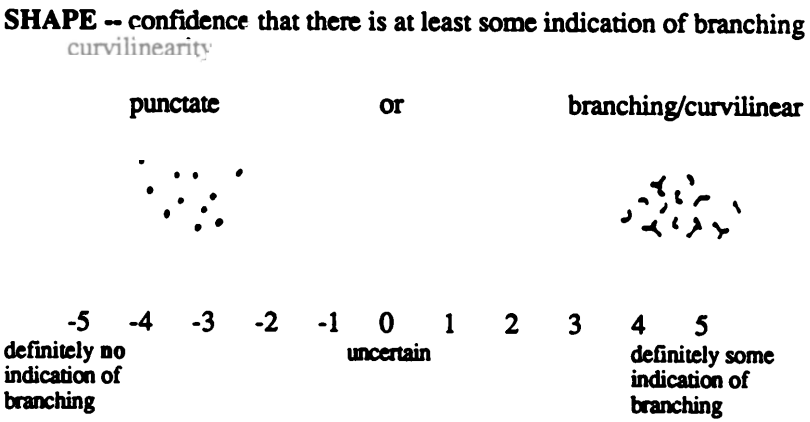


Figure 13.10
Illustration of a scale of a perceptual feature on a mammogram. Branching or curvilinear calcium elements are indicative of cancer (at right), whereas punctate elements are indicative of a benign abnormality.

"shape of calcifications." Less extreme instances of these dimensions would spread out in between. Pursuing the last of these examples, curvilinear calcifications versus noncurvilinear or "punctate" calcifications would be a strong determination of visual similarity and hence would define a dimension or feature. The investigators and specialists devised a numerical scale for each such feature, as exemplified for curvilinear and punctate calcifications in figure 13.10. For other types of features, the scale was a reader's confidence that a particular feature is present or, alternatively, a physical measure of extent—for example, the length and width of a mass or group of calcifications.

As mentioned, the MDS procedure reveals certain features to be highly enough correlated with others so that not all need be retained. This information was used, along with the relative importances of the features, to reduce the set of candidate features to about thirty. The features described in connection with figure 13.1 were confirmed as important by the MDS analysis and the experts' consensus. Mammography is a well-established imaging technique of many years' standing, and the features revealed by our MDS analysis were among those mentioned earlier in interviews. Hence the MDS analysis served mainly to help us select among the features previously mentioned by mammographers. However, in our applications of MDS to newer techniques, for example, magnetic resonance imaging of the prostate gland, MDS analysis has revealed previously unsuspected features. The same was true in a study of photographs of the lens of the eye typically made to classify types of cataracts; features

newly revealed by MDS analysis provided a more sensitive classification than that previously available.

13.7.1.3 Determining the Final List of Features and Their Weights

We investigators believed that the set of thirty features that remained after the consensus conference was too large to be workable and probably contained several features that would not carry much weight in a diagnosis, that is, would not be very predictive of malignancy. We therefore made a second reading study and applied another statistical technique to reduce the feature set to just the most important ones. We had in mind a set of a dozen or so features. We wanted to know quantitatively how predictive each of these features was so that we could assign a weight to each, that is, a number reflecting its importance in diagnosis. We looked ahead to a merging aid for the radiologist who, in reading mammograms, would assign a scale value to each feature, which would then be multiplied by the weight of that feature, such that the resulting products could be linearly combined to calculate a probability of malignancy.

In the second reading study, the specialists applied the scales of the thirty features that remained after the consensus conference to the images of 100 new cases for which the diagnostic truth was known, including 50 cancers proven by biopsy and 50 noncancers established by one year of follow-up. The statistical procedure of "discriminant analysis" was then used to determine which minimal set of features was necessary and sufficient to be retained in the final checklist, and to determine their relative weights. This analysis shows specifically which features with which weights serve to maximize discrimination acuity, where acuity is expressed by a measure closely akin to the ROC area measure (A_z) defined earlier. Twelve features were selected for the final set, to be used in creating a checklist for radiologists. This final set of features included those with high enough weights to contribute significantly, in our view, to discrimination acuity and excluded those making only a minimal contribution.

13.7.1.4 The Merging Aid

The same discriminant analysis was the basis for the merging aid, the computer program that linearly combined the twelve scale values given to it by the radiologist for the twelve checklist features, appropriately weighted, for each case read. The computer's output, as mentioned, was an estimate of the probability of cancer that radiologists could use in making their own estimate. Readers might report a higher or lower probability value for cases where they thought there was something in the mammogram that was not captured adequately by the feature scales.

The concepts and details of how discriminant analysis determines a final set of features for a checklist and then serves as a merging aid are the

subjects of entire books and manuals (e.g., Lachenbruch 1975), but I will attempt to capture briefly the intuitive essentials. Recall that the mammography specialists gave a rating on each of thirty features (in a "master" set) for each of 100 mammographic cases known to have cancer or not. A simple way to determine which of the features are quantitatively the most useful in distinguishing cancer from noncancer cases would be to calculate from those ratings an ROC measure of acuity for each feature independently. One problem with this approach is that there is no way to determine which of the thirty features constitute an appropriate smaller set. Another problem with this approach is that there is no clear way to merge a set of feature values into a single decision variable, namely, the probability of cancer, especially because the features cannot be assumed to be independent. The form of discriminant analysis we used addresses these problems by implementing a stepwise feature selection procedure, in which features are selected in order of their acuity after taking their intercorrelation into account. That is, the procedure first selects the most discriminating feature and gives a measure of its diagnostic acuity, then selects the feature that is second most discriminating, when only acuity beyond that offered by the first is considered, and measures the *additional* (uncorrelated) acuity the second feature provides, and so on. This particular second feature is not necessarily the one with the second highest acuity when features are considered independently. The procedure can be terminated when additional features are adding only insignificantly to total diagnostic acuity. If you think of each mammogram as being represented by a point in a multidimensional space (as in MDS analysis) that has the final set of features (say, twelve) as dimensions, then the mammograms indicating cancer collect in one cloudlike region of the space and those indicating no cancer collect in another. The discriminant analysis determines the direction, or vector, through the space that best separates the two clouds. This determination serves to reduce the number of dimensions of the space to just one dimension, namely, the line represented by the vector. The equation for that vector is a linear function of the dimensions, and the coefficients of the equation represent the relative weights of the features in diagnosis. The equation is called a "discriminant function."

This function constitutes the merging aid when individual mammogram cases are read. Scale values assigned by the reader to the features of a given case are multiplied by their respective weights (the function's coefficients) and added together to yield a discriminant score. In effect, the mammogram being read is represented as a point along the discriminant vector. This point or score is then translated into a probability that the mammogram at hand reflects the presence of cancer, and this probability

is the advisory estimate that is given as an aid to the mammogram reader (Getty et al. 1988).

13.7.1.5 Experimental Test of the Effectiveness of the Aids

Six other radiologists, experienced but not specialized in mammography, read a different set of cases (58 cancer, 60 noncancer) to provide an evaluation of the effectiveness of the two aids. They read the cases first in their usual manner without the aids ("standard condition") and then, after six months, with them ("enhanced condition"). In both instances, the readers gave their degree of confidence, on a five-category scale, that a cancer was present. The rating method described in section 13.3.1 was used to construct an ROC for each reader in each condition. A group ROC for each condition was obtained by pooling the rating data of the individual readers.

Figure 13.11 shows the ROCs for standard and enhanced conditions for the group of six readers. The lower curve, for the standard reading, has an acuity measure (as defined in section 13.3.3) of $A_z = .81$. The upper curve, for the enhanced reading, has an $A_z = .87$. All six readers showed an

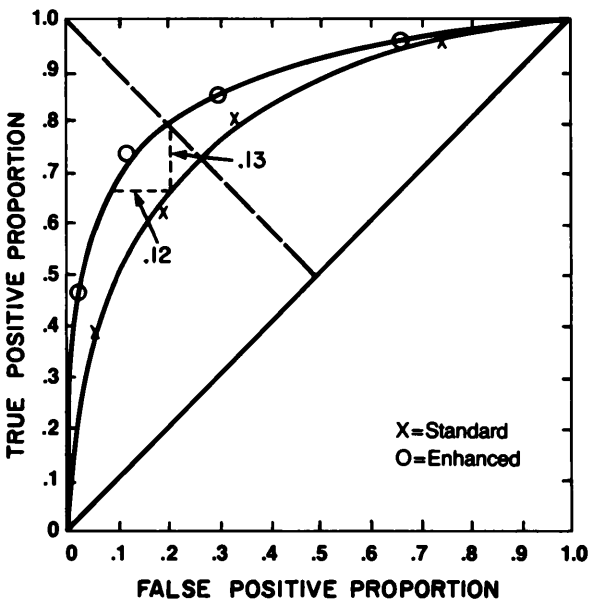


Figure 13.11
Empirical relative operating characteristics (ROCs) for standard and enhanced readings of mammograms. Dashed lines illustrate gains from enhancements in either the true positive or false positive proportion.

effect in the same direction, of about the same size; thus the finding is highly unlikely to arise from chance or random variation.

As a control condition in the experiment, to help assure that the observed difference was due to the aids, three of our readers reread the cases in the standard mode just before making the enhanced readings and the other three reread the cases in the enhanced manner six months after the first enhanced reading. The A_z values for rereadings in the standard mode were essentially the same as the A_z values for the original standard reading, indicating that the passage of six months did not bring an increase in A_z. Likewise, the A_z values in the first and second enhanced readings were almost identical, indicating that the enhancement produced by the aids would be sustained over time. Incidentally, the aids brought the nonspecialists' performance up to that of the specialists. That improvement may be even better than it looks, because our measure of the specialists' performance, while ostensibly in a standard condition, was obtained after their experience in the interview, MDS study, and consensus conference, and we believe this measure is higher than it would be had it been taken prior to that intensive experience.

13.7.1.6 Clinical Significance of the Observed Enhancement

More informative for purposes of this study than the overall acuity measures are the specific gains in TPP, reductions in FPP, or both that the aids produced. Consider a decision criterion for recommending a biopsy that yields a point on the "standard" ROC at FPP = .20, TPP = .67, as indicated in figure 13.16 where the short dashed lines meet. It can be seen that for the same value of FPP = .20, the "enhanced" ROC shows TPP = .80—an improvement in TPP of .13, or an increase of 19 percent. Turning instead to the gain in acuity afforded by the aids as a reduction in FPP, it can be seen that for TPP = .67, the "enhanced" FPP = .08; this value of .08 is .12 less than the "standard" FPP of .20 (a decrease of 60 percent). The indication here is that, for 100 patients with cancer, the aids will permit detection of cancers in about 13 additional patients (an increase from 67 to 80); alternatively, for 100 patients without cancer, they will permit avoidance of an unnecessary biopsy in about 12 additional patients (a decrease from 20 to 8). Other balances are possible in making use of the observed gain in acuity, for example, a simultaneous increase in TPP and decrease in FPP of approximately .06 each (a 9 percent increase in the TPP and a 30 percent decrease in the FPP).

Following this indication of the value of our feature-based approach to mammography, we have begun to design a computer-based tutorial program for training new radiologists and for the continued education of practicing radiologists (Getty et al. 1992). We consider the possibility—

given that the perceptual feature approach does not depend on a knowledge of underlying anatomy and pathology—that paramedics can be trained and assisted in reading mammograms with good acuity, a result of especial value in areas or countries where radiologists are in short supply.

13.7.2 Optimizing the Decision Criterion

Choosing an appropriate decision criterion in almost any practical setting is difficult, and how to do so in mammography is by no means clear. The aim of the next few paragraphs is to indicate that decomposing the discrimination and decision processes is critical and that concepts of the optimal criterion provide a structure for analyzing the problem and for moving toward a consensus.

13.7.2.1 The Expected Value Optimum

Can the expected value definition of the optimal decision criterion be applied in this medical setting? As discussed in section 13.2.2.4, one would need to estimate the prior probability of cancer in the population of patients and to assign numerical benefits and costs to the four possible diagnostic outcomes. Data are available on prior probabilities (called "prevalences" in medicine) for two major populations, the population involved in a screening program for women in general and the population of patients who appear at a referral center because they have a high risk for breast cancer or because a physical examination suggested the possibility of a lump. The prior probability of cancer in screening populations is about .02; in referral populations, about .32. This implies a sixteen fold difference in the optimal *LR* or slope measure *S* and, for any set of benefits and costs, would call for a large difference in the decision criterion adopted in the two settings, with a far more lenient criterion being appropriate for the referral setting. There have been several attempts, as discussed in the literature on medical decision making, to develop ways to assign benefits and costs to diagnostic outcomes (e.g., Weinstein and Fineberg 1980). However, such an assignment is always very difficult when human lives are evaluated, and individual opinions may differ widely. One way to soften the demands of the expected value definition is to quantify its overall ratio of benefits and costs without assigning all four individual values: benefit (TP), cost (FP), benefit (TN), and cost (FN). One could say, for example, "I would five times rather be right when condition A exists (when cancer is present) than when condition B exists (when cancer is not present)." Combining just that ratio with the ratio of prior probabilities is enough to specify an optimal criterion.

13.7.2.2 The Optimal Criterion Defined by a Particular False Positive Proportion

Another way to bypass explicit benefits and costs is to select the criterion that satisfies a specified limit on FPP. This definition of the criterion is used in testing statistical hypotheses, where the limit on FPP (the "type I error", or "level of significance") is usually .05 (or .02 or .01). With this criterion, benefits and costs are considered only tacitly in arriving at the proportion of false positive errors that can be tolerated. Anecdotally, this tolerable FPP is often thought to be around .10 in medical contexts.

13.7.2.3 Societal Factors in Setting a Criterion

The problem of choosing an appropriate decision criterion is illuminated by looking directly at the values of FPP and TPP that are attainable. Take the top ROC in figure 13.5 as an approximation to the performance of mammography (perhaps a little generous). The FPP = .10 just mentioned allows a TPP = .80. There is a strong desire in medical circles to increase that TPP—to .90, say—because missing 2 of 10 patients with cancer seems like too many. But TPP = .90 brings along a FPP = .20. And a further increase to TPP = .95 carries an increase to FPP = .35. Consider further those last two ROC operating points: TPP can be increased from .90 to .95 at a price of increasing FPP from .20 to .35. Consider them in connection with a screening population, wherein, as I mentioned, about 2 percent of the patients have a breast cancer. In a sample of 5,000 patients, 100 will have cancer and 4,900 will not; hence, at the more lenient criterion, 95 instead of 90 of the 100 cancers will be detected. Meanwhile, the number of unnecessary biopsies will increase by 735—from $4{,}900 \times .20 = 980$ up to $4{,}900 \times .35 = 1{,}715$. Detecting 5 more cancers at a price of 735 more unnecessary biopsies does not make the more lenient criterion an obviously better balance of TPP and FPP than that of slightly stricter criterion. Choosing between those or any two criteria would seem to require some quantitative analysis of costs and benefits. Then again, the wishes of individual patients and their physicians may dominate, and both patients and physicians tend to read and interpret the odds very conservatively. When physicians are asked to say how low a probability of cancer will lead them to recommend biopsy, they indicate a range in the neighborhood of .05.

That the yield of biopsy is 20 to 30 percent in the United States suggests that the decision criterion is set with the desires of individual patients and physicians in mind, rather than with a focus on what might be regarded as a cost-effective approach for society as a whole. England's yield of about 50 percent reflects a noticeably stricter criterion, one possibly more mindful of society's requirements. Societal strictures may be

effected by regulation or by the availability of services. As I observed earlier, if the U.S. population were to follow the national guidelines for mammographic examinations, thus requiring exams in much larger numbers, and if the current criterion for biopsy were retained, there would probably not be enough surgeons and pathologists available to supply the resulting number of recommended biopsies. At that point, the issue of choosing an appropriate decision criterion would take on a salience that is not apparent now.

13.7.3 Adapting the Enhancements to Medical Practice

Now that we have a way to increase diagnostic acuity in mammography and a way to think about the decision criterion for recommending treatment, how might these results be put into practice? In a project just underway, my colleagues and I are working on a computer system that will embed these enhancements in a network of several related workstations. I am pleased to say that the radiologists with whom we work have been very cooperative. Some may regret systematization at the expense of art in reading images, but radiologists are scientifically and quantitatively minded and most can be expected to adopt a system that provides a demonstrated gain in acuity.

Another benefit that derives from the present approach is a standardization of radiologists' reports to referring clinicians and surgeons. There is now a good deal of clamor nationally among associations representing the recipients of those reports, and several government agencies have worked with the American College of Radiology to bring about a standardized vocabulary for the prose reports. In the system we are building, the radiologist's key presses (or spoken words) to indicate scale values to the computer will be converted automatically to corresponding, appropriate sentences in the report (Swets et al. 1995). In fact, the perceptual features as named in the study described above are the elements of a standard vocabulary or lexicon currently espoused by the American College of Radiology (Kopans and D'Orsi 1992).

Though no data exist, I believe that standardizing perception in mammography (the scaling of features) will give better results than standardizing language alone. The time saved for the radiologist by automatic report generation may compensate effectively for whatever additional time is required by explicitly scaling perceptual features in reading an image. Moreover, the reports can be delivered to their intended recipients as they are generated. Reports can also be stored in a database, along with subsequent outcome data, which may afford several benefits. Individual radiologists can be assisted to greater acuity by tutorials tailored to the knowledge they need, for example, how better to scale particular features.

A database can also assist in standardizing decision criteria across the radiologists in a given hospital, say, by reporting the yields of biopsies they recommend. On another tack, digital imagery will permit radiologists to secure second opinions from anywhere, in feature-by-feature detail, and to resolve differences of opinion by reference to quantitatively scaled features. Finally, there is a possibility that the use of a standard, approved procedure will reduce a practitioner's liability in suits for malpractice.

The next section briefly describes another possible application to the interpretation of diagnostic images. But note that images are not essential to this approach. Neither, for that matter, is a two-alternative response. One might, for example, develop features related to a psychiatric patient —the patient's nonadaptive behavior, risk level for dangerous behaviors or chemical dependencies, social resources, and so on—and, in turn, develop a system that combines ratings of those features to enable a choice among several graded treatment possibilities, ranging across standard outpatient treatment, more frequent visits, and acute inpatient care.

13.8 Detecting Cracks in Airplane Wings: A Second Practical Example

The possibilities for improving discrimination acuity and adopting an appropriate decision criterion, as demonstrated for mammography, exist in similar diagnostic fields. Consider an example from materials testing, wherein the imaging techniques of ultrasound and eddy current are used to visualize flaws in metal structures. The specific example is the detection of cracks in airplane wings by maintenance personnel. Incidentally, a diagnostic extension of this simple detection task that involves recognition as well requires a characterization of the severity of a flaw in order to project the time course of its developing into something still more serious.

13.8.1 Discrimination Acuity and the Decision Criterion

There is clearly a lot at stake here: a false negative decision can lead to a catastrophic accident and dramatic loss of lives. Still, a false positive decision takes a plane out of service unnecessarily at substantial costs in dollars and convenience. On balance, as in mammography, the costs and benefits seem to suggest a lenient criterion for declaring a flaw. On the other hand, the prior probability of a dangerous flaw, although increasing with an aging fleet of airplanes, is still very low, and low prior probabilities, even with moderate criteria, tend to produce a large number of false positives, possibly so large as to be unworkable. Thus attention to both discrimination acuity and the decision criterion is required.

13.8.2 Positive Predictive Value

The idea that airplane fleet operators cannot just proceed with a lenient decision criterion, accepting many false positives in order to reduce false negatives to near zero, may need support. An important concept in this connection and many others is the predictive value of a positive decision. The *positive predictive value* (PPV) is the proportion of positive responses that are true. In the two-by-two array of table 13.2, it is defined as $a/(a + b)$; that is, it is a TP proportion based on the occurrence of the response (row sum), rather than the TPP based on the occurrence of the stimulus (column sum) that I have emphasized thus far. It is also referred to as the "inverse TP proportion," because it goes backward from response to stimulus. Note that I introduced this concept in speaking of the yield of the biopsy procedure in the mammography study. The point is that low prior probabilities and lenient decision criteria conspire to produce a low PPV, such that the observer's decisions often "cry wolf" when there is no wolf.

I have looked at the values of PPV that can be expected in another aircraft setting, where specialized sensors (detectors) give in-flight warnings to pilots about various imminent dangers—including collision, engine failure, ground proximity, and wind shear. Even when it is assumed that the detectors are extremely acute and the decision criteria are very strict, the PPV can easily be lower than .05, so that only 1 in 20 warnings is valid (Getty et al. 1995). Such a low predictive value can be inappropriate, for example, in a setting in which a simple "fly-around" by a plane in final approach to a major airport—a fly-around in response to a wind shear warning, say—can put air traffic control under additional strain for several hours, and thus endanger many other aircraft in addition to the one taking the first evasive action. In short, the cost of a false positive decision is not negligible. In fact, pilots come to ignore warnings from unreliable systems when they have other important things to do.

An example from medical diagnosis of the effect of low prior probabilities on PPV comes from testing for the human immunodeficiency virus (HIV) of AIDS. For low risk populations—for example, blood donors or a company's employees—the prior probability of HIV is about .003 (Bloom and Glied 1991). Ordinarily, in such a setting, a positive outcome on a typical screening test is followed by a more conclusive (and expensive) confirmatory test (Schwartz, Dans, and Kinosian 1988). Using the College of American Pathologist's estimates of test performance (TPP and FPP) for the best screening and confirmatory tests (Bloom and Glied 1991), I calculate that after a positive result on *both* tests, the PPV is .13. Hence, 6 of 7 individuals diagnosed as positive in this manner are told they have the HIV when they do not in fact have the HIV (Swets 1992). As another

example, the issue of *Newsweek* magazine that appeared as I write (19 April, 1994) mentions a study in which 65 exploratory surgeries for ovarian cancer were performed for every cancer found by the surgery—to many of us, no doubt, an incredibly low yield for a risky surgical procedure. The prior probability (lifetime risk) of ovarian cancer was said to be .015; even for the category of patients at special risk for ovarian cancer because that disease has occurred in the family, the prior probability is only .05.

13.8.3 Data on the State of the Art in Materials Testing

The U.S. Air Force kindly gave me access to the data from its major study of detecting cracks in airplane wings. In the study, 148 metal specimens with and without flaws were carried to 17 air force bases, where they were inspected, in total, by 121 technicians using ultrasound imaging and 133 technicians using an eddy current imaging technique.

The performance of each technician in this study is depicted by a point on an ROC graph in figures 13.12 and 13.13. The impact of just a glance at these graphs is the main result: the variation across technicians is about

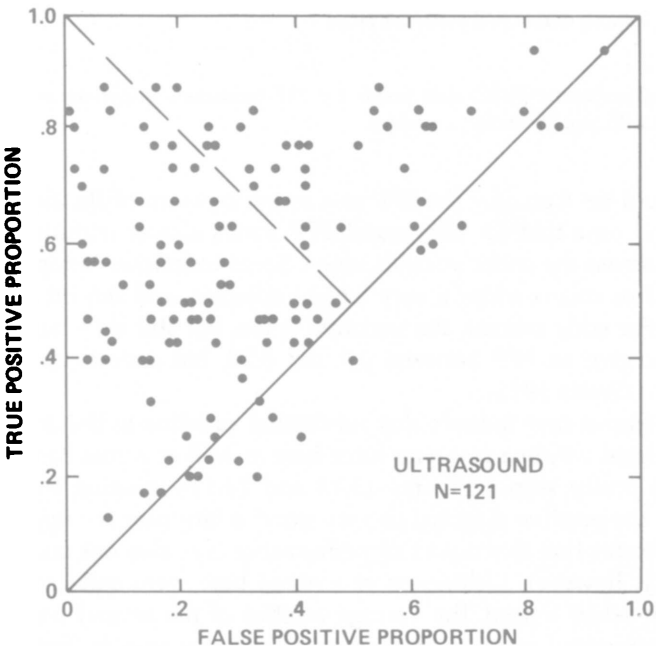


Figure 13.12
Relative operating characteristic (ROC) data points for 121 technicians inspecting metal specimens for cracks with an ultrasound technique.

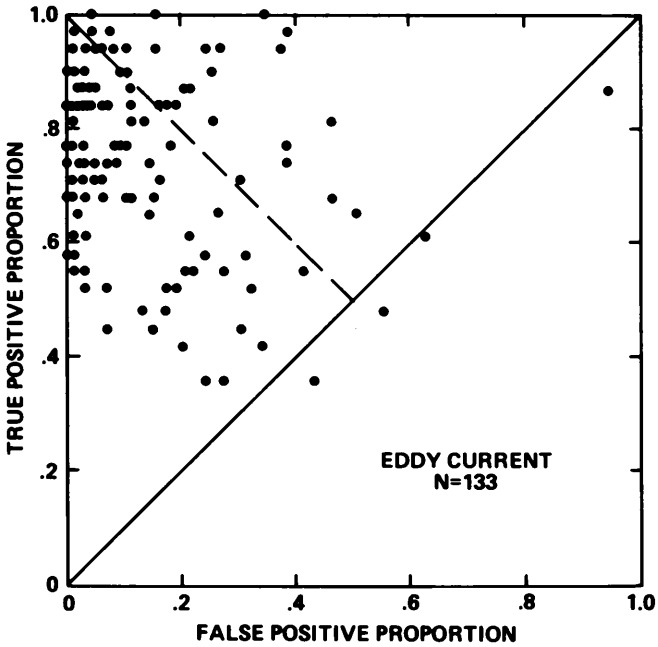


Figure 13.13
Relative operating characteristic (ROC) data points for 133 technicians inspecting metal specimens for cracks with an eddy current technique.

as large as it could be. Consider the FPP as a rough measure of the decision criterion and note that for ultrasound, FPP varies almost uniformly from 0.0 to 1.0, across the entire possible range. Some technicians adopt a very strict criterion, others adopt a very lenient criterion, and still others are in-between. For eddy current, the variation is less but still very large: most technicians give an FPP between 0.0 and 0.20, but several range from 0.20 to 0.50 (Swets 1992).

Analyses not shown here indicate that substantial variation in the decision criterion existed within a given air force base as well as across bases. The variation in acuity seen in figures 13.12 and 13.13—ranging from chance acuity at the positive diagonal to very good acuity near the upper left corner—indicates that this aspect of performance was also not under control. In acuity, however, technicians at a given base were quite consistent with each other; indeed, the average acuities of the several bases varied almost uniformly across the full range of possible acuities (Swets 1992). This result suggests that the inspection techniques used at the bases with high acuity could be analyzed and transferred to the others, much in the way demonstrated for mammography. Similar variation across techni-

cians in acuity has been found in other studies of materials testing, for example, in the examination by eddy current of steam generators in power plants (Harris 1991).

13.8.4 Diffusion of the Concept of Decomposing Diagnostic Tasks

As a counter to the example of medical radiology, where the ROC has been used extensively, it should be noted that the ROC approach has not been picked up in the materials-testing field. Over ten years, I carried the message along with the data of figures 13.12 and 13.13 to the Air Force and the Army, and also to the Federal Aviation Administration and a major commercial aircraft manufacturer, and further to the research laboratory of the U.S. electric power companies, to no avail. I am told that a tutorial article in the field's professional magazine (Swets 1983) has begun recently to have some impact (W. S. Brown, personal communication, December, 1995), but I have not been able to obtain the reports.

There are also blind spots within clinical medicine. For example, the several tests for screening for the HIV of AIDS have widely different acuities and decision criteria (Swets 1992). Moreover, each test's decision criterion is held constant across different settings, for example, whether used on donated blood or to diagnose individuals—two applications in which the cost of a false positive decision differs considerably. Further, those criteria remain fixed when screening different populations of individuals for which the prior probability differs greatly, say, from .003 (low-risk blood donors) to .45 (methadone clinic attendees).

However, again on the brighter side, the idea of using an ROC measure of acuity is making noticeable progress in weather forecasting, where it is overcoming resistance from advocates historically of other measures (Williams 1985; Brunet, Verret, and Yacowar 1987; Sarrazin and Wilson 1987). The ROC is generally accepted by evaluators in the field of information retrieval (Brookes 1968; Heine 1973). A laboratory of the National Aeronautics and Space Administration is now examining the application of signal detection theory and the ROC to in-flight warning systems (Getty et al. 1995). If I have any idea about the unevenness in the acceptance of the decomposition concepts, it is a matter of the size of the diagnostic problem, and perhaps the scientific versus technological bent of the people involved. University radiologists, laboratory meteorologists, and information retrieval methodologists do studies and write articles, and have the frame and peace of mind to want to do them right. Managing the nation's aircraft travel, power plants, and disease epidemics no doubt attracts individuals having a different frame of mind and gives them no peace of mind. It is all the more a shame, but we can easily imagine that technical matters that seem difficult or esoteric get blown off desks at the

Federal Aviation Administration, the Department of Energy, and the Food and Drug Administration (Swets 1991b).

13.9 Some History

For a century in psychology, beginning with the systematic study of sensory detection and recognition around 1850 (Fechner 1860), the dominant psychological position was that the two processes of discrimination and decision were essentially one—that the discrimination process led automatically to response A or response B. Decisions themselves were thought not to be involved; rather, the discrimination process included an automatic threshold device that determined response A or response B depending on whether a particular, fixed strength of evidence was exceeded or not, with the particular strength being determined by the physiological makeup of the organism. There was for some psychologists a concern that attitude could affect the observer's responses, but in general, it was thought that discrimination acuity could be measured without concern that the measures would be distorted or made unreliable by variation in a separate decision process.

The decomposition of discrimination and decision into two measurable processes was suggested and made possible about 1950 by a general theory of signal detection. The theory was developed by engineers Wesley Peterson and Theodore Birdsall for theoretical devices—called "ideal observers"—in the context of radar (Peterson, Birdsall, and Fox 1954). The discrimination problem they addressed was to determine which values of certain physical measurements of electronic waveforms indicated the presence of a signal that represents, for example, the approach of an enemy plane. The generality of signal detection theory is indicated by the similarity of this discrimination problem and that of mammography.

The decision part of signal detection theory was derived from statistical decision theory (Wald 1950), which is an extension of the theory of testing statistical hypotheses developed by Jerzy Neyman and Egon S. Pearson in 1933. The earlier one-process view of detection and recognition in psychology mirrored statistical hypothesis testing in the form popularized by R. A. Fisher at about the same time (Gigerenzer and Murray 1987). In determining, for example, whether the mean of population A differs from the mean of population B, a fixed decision criterion is assumed, akin to a fixed threshold, such that the probability of deciding $A \neq B$ when in fact $A = B$ (a "type I error") is no greater than .05 (or .02 or .01). In the broader decision theory, the placement of a variable decision criterion in any instance depends on the prior probabilities of $A \neq B$ and $A = B$ and

on the benefits and costs of correctly and incorrectly deciding that $A \neq B$ or that $A = B$. The use of Fisher's fixed criterion in statistical analysis (with prior probabilities and benefits and costs being only tacit) is common when making an inference from data, that is, a "knowledge claim." A variable criterion (with those situational variables being explicit) is more appropriate when, instead, some action is to be taken.

Psychologist Wilson P. Tanner, Jr., and I (1954), working with the signal detection theorists, proceeded to show that human observers in simple sensory detection tasks (detecting brief tones or spots of light) base their responses on a variable decision criterion that takes into account the prior probability that a "signal" is present and the benefits and costs associated with the various possible decision outcomes. Although the human observers did not match exactly the optimal expected value criterion, their criteria under different conditions of prior probabilities and benefits and costs were highly correlated with the optimum. In general, the empirical ROCs obtained from these observers looked like the theoretical curves shown in earlier sections (Swets, Tanner, and Birdsall 1961).

This inclusion of nonsensory factors in the theory of simple sensory tasks—"expectancy" and "motivation" in psychological terms—was no doubt part of the spirit of the times as cognitive psychology came into view. Gird Gigerenzer and Richard Murray (1987) have recently developed in some detail how perceptual and cognitive theory have followed statistical theory, how the human has been viewed as an intuitive statistician. In the human sensory experiments first following on signal detection theory, the decision criterion was actually manipulated by varying prior probabilities of the stimuli and benefits and costs of the stimulus-response outcomes. Without overt manipulation, you might suspect that decision criteria would vary from one observer to another based on individual impressions of implicit prior probabilities and implicit benefits and costs. For example, one observer might want to demonstrate acuity by maximizing the true positive proportion, while another might want instead not to issue a false positive response, perhaps because it might be viewed as an hallucination. Distinguishing such different tendencies by isolating decision effects is necessary if acuity is to be measured validly (Swets 1961; 1973; 1988).

James Egan (1958) soon extended the psychological coverage of signal detection theory to recognition memory, that is, to deciding whether an item presented (a word, say) was on a list shown earlier. Other psychologists then applied the theory to attention, imagery, reaction time, manual control, learning, conceptual judgment, personality, and animal behavior (Swets 1973). The first extension to a practical diagnostic problem was to information retrieval, where one wishes to single out those documents in a library that satisfy a particular need for information and to avoid the

irrelevant documents (Swets 1963). Lee Lusted (1968) showed the value of the theory's decomposition of diagnostic tasks in radiology, presaging hundreds of published applications in clinical medicine.

Some aspects of signal detection theory, without the variable decision criterion and ROC, were present in earlier psychological theory. For example, in the 1920s, the innovative psychometrician Louis Leon Thurstone (1927a,b) proposed stimulus distributions such as those in figure 13.2 as part of an extensive statistical theory of discrimination, calling them "discriminal processes." For Thurstone, however, the two stimuli were not polarized as positive and negative, and the (response) criterion was viewed in the theory as being fixed at the symmetrical location, where the two curves cross. Hence his theory did not go on to separate decision and discrimination processes.

The ROC was developed in signal detection theory, where it is called the "receiver operating characteristic," and is a unique contribution of that theory. I alert you to the tendency of statisticians to remark that the ROC is "essentially the *power curve* of statistics." Rather, the power curve is the S-shaped curve of TPP plotted against the difference between the means of two statistical hypotheses for a fixed, low value of FPP (e.g., .05). Thus, although the power curve includes the same three variables as the ROC, it shows how acuity and TPP vary together for a single FPP. It does not show the interplay of TPP and FPP for each level of acuity, that is, a variable decision criterion, and thus gives no insight into the separation of discrimination and decision processes. Indeed, what is essentially the power curve appears also in signal detection theory, where it is called the "betting curve," and it has thrived for over a century in psychology, where it is called the "psychometric function," without suggesting the impact of the ROC (Green and Swets 1966).

Finally, a few comments on method and terminology, to help the reader who may see the ideas of the chapter in psychological contexts. In early sections of this chapter, I mentioned three data collection methods, which I called the "yes-no," "rating," and "paired-comparison" methods. In sensory psychology, where the incentive is to minimize criterion effects in the simplest way, the frequently used paired-comparison method is called the "two-alternative forced-choice" (2AFC) method. In connection with the other two methods, both used to determine an ROC, I used the general terms *true positive* and *false positive proportions*. In psychology, where the legacy of signal detection theory is most familiar, these tend to be called, respectively, "hit" and "false alarm proportions." The three methods existed long before signal detection theory (SDT) was developed, but tended to give different results. SDT related the three methods quantitatively and rationalized them by showing theoretically that they give, under fixed conditions, the same value of a common measure of dis-

crimination acuity (Swets, Tanner, and Birdsall 1961). Empirical demonstrations of the validity of this theoretical prediction were important in establishing the general validity of the concepts of signal detection theory.

Suggestions for Further Reading

Three textbooks on signal detection theory and ROC analysis are those of Green and Swets (1966), Egan (1975), and Macmillan and Creelman (1991). A collection of articles treats the range of subjects of this chapter (Swets 1996). A pair of articles shows representative ROCs from several psychological and diagnostic tasks and develops the implications of their form for alternative measures of discrimination acuity (Swets 1986a,b).

The examples given here of experimental decomposition of discrimination and decision processes are treated further as follows: vigilance (Davies and Parasuraman 1982; See et al. 1995); recognition memory (Murdock 1974); polygraph lie detection (Shakhar, Lieblich, and Kugelmass 1970; Swets 1988); information retrieval (Swets 1963, 1969); and weather forecasting (Swets 1986b, 1988).

On combining observations to achieve greater acuity, see Green and Swets (1988), Metz and Shen (1992), Seltzer et al. (1992), and Swets (1984). Articles describing the enhancement of accuracy of mammography are by Getty et al. (1988), Swets et al. (1991), and D'Orsi et al. (1992). A textbook treatment of evaluation methods in diagnostics is given by Swets and Pickett (1982). Measures of diagnostic acuity in several settings are given by Swets (1988); choosing the best decision criterion in diagnostics is discussed by Swets (1992).

For the background and history of signal detection theory in engineering, statistics, and psychology, see Swets (1973, 1996) and Gigerenzer and Murray (1987).

Problems

13.1 Figure 13.2 shows probability distributions corresponding to stimuli A and B, respectively, and a decision criterion x_c at the midpoint (mean) of the A distribution. The likelihood ratio of this criterion is 2.5. The value of TPP resulting from the criterion is .50; the value of FPP is .15; as represented by hatched and crosshatched areas. What would TPP, FPP, and LR_c be for a criterion at the midpoint of the B distribution?

13.2 The table below presents illustrative rating scale data. Plot the relative operating characteristic (ROC) for these data. Estimate by a graphical method the slope measure S of the decision criterion for each data point.

	Stimulus	
Response	Abnormal (positive)	Normal (negative)
1, very likely abnormal	350	25
2, probably abnormal	50	50
3, possibly abnormal	40	75
4, probably normal	35	150
5, very likely normal	25	200

13.3 What are the expected payoffs (average over many decisions) of the diagnostic systems performing as in table 13.3 (system A), table 13.4 (system B) and table 13.5 (system C), if the benefits and costs are TP = 10, FP = −8, FN = −4, and TN = 2? What is the optimal likelihood ratio criterion, LR_c, for each system?

13.4 Suppose that the ROCs of figure 13.6 give the acuities of three weather-forecasting systems for predicting frost that are available to a fruit grower whose produce is susceptible to damage by frost. System A ($A_z = .75$) is the daily newspaper; system B ($A_z = .85$) is the National Weather Service's prediction given by telephone; and system C ($A_z = .95$) is a commercial service that focuses on frost and precise local areas. Suppose further that the optimal decision criterion for a particular grower is $LR_c = 1.0$. How would you express the gains in forecast accuracy that may be achieved in moving from system A to B to C?

13.5 Suppose that the data in tables 13.3, 13.4, and 13.5 were obtained in an evaluation of three alternative detection systems designed to warn a pilot of the possibility of midair collision with another aircraft. The frequencies in the rows give the numbers of warnings issued by the systems (positive response) and not issued (negative response) when the present condition was truly dangerous (left column) or not truly dangerous (right column).

Call the three systems A (table 13.3), B (table 13.4), and C (table 13.5). Which system would you choose if you were required to put one of them into routine operation? What is the "positive predictive value" for each system that is implied by the data; that is, after a large number of warnings by each, what proportion of warnings would have truly signified a dangerous condition?

References

Ackerman, L. V., and Gose, E. E. (1972) Breast lesion classification by computer and Xeroradiograph. *Cancer* 30 (4), 1025–1035.

Banks, W. P. (1969). Criterion change and response competition in unlearning. *Journal of Experimental Psychology* 82, 216–223.

Barr-Brown, M., and White, M. J. (1971). Sex differences in recognition memory. *Psychonomic Science* 25, 75–76.

Bloom, D. E., and Glied, S. (1991). Benefits and costs of HIV testing. *Science* 252, 1798–1804.

Broadbent, D. G. (1971). *Decision and stress*. London: Academic Press.

Broadbent, D. G, and Gregory, M. (1963). Vigilance considered as a statistical decision. *British Journal of Psychology* 54, 309–323.

Brookes, B. C. (1968). The measures of information retrieval effectiveness proposed by Swets. *Journal of Documentation* 24, 41–54.

Brunet, N., Verret, R., and Yacowar, N. (1987). Comparison of MOS and perfect PROG systems in producing numerical weather element forecasts. Tenth Conference on Probability and Statistics in Atmospheric Sciences, Edmonton, Canada. Boston, MA: American Meteorological Society, 12–17.

Buckner, D. N., and McGrath, J. J. (Eds.) 1963. *Vigilance: A symposium*. New York: McGraw-Hill.

DaPolito, F., Barker, D., and Wiant, J. (1971). Context in semantic information retrieval. *Psychonomic Science* 24, 180–182.

Davies, D. R., and Parasuraman, R. (1982). *The psychology of vigilance*. New York: Academic Press.

Davies, R. M., and Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin* 81, 95–106.

Donaldson, W., and Murdock, B. B., Jr. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology* 76, 325–330.

D'Orsi, C. J., Getty, D. J., Swets, J. A., Pickett, R. M., Seltzer, S. E., and McNeil, B. J. (1992). Reading and decision aids for improved accuracy and standardization of mammographic diagnosis. *Radiology* 184, 619–622.

Egan, J. P. (1958). *Recognition memory and the operating characteristic*. Technical Note AFCRC-TN-58-51. Indiana University, Hearing and Communication Laboratory. See Green,

D. M., and Swets, J. A. (1966/1988). *Signal detection theory and psychophysics.* New York: Wiley.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.

Egan, J. P., Greenberg, G. Z., and Schulman, A. I. (1961). Operating characteristics, signal detectability, and the method of free response. *Journal of the Acoustical Society of America* 33, 993–1007.

Frankmann, J. P., and Adams, J. A. (1962). Theories of vigilance. *Psychological Bulletin* 59, 257–272.

Fechner, G. T. (1860). *Elemente der Psychophysik.* Leipzig, Germany: Breitkopf and Hartel. English translation of Volume 1 by H. E. Adler, D. H. Howes, and E. G. Boring (Eds.), *Elements of Psychophysics.* New York: Holt, Rinehart, and Winston, 1960.

Gale, A. G., Roebuck, E. J., Riley, P., and Worthington, B. S. (1987). Computer aids to mammographic diagnosis. *The British Journal of Radiology* 60, 887–891.

Getty, D. J., Pickett, R. M., D'Orsi, C. J., Freeman, B. F., and Swets, J. A. (1992). Computer assisted instruction in radiology. Bolt Beranek and Newman Inc., Final report under NIH Grant No. 5 RO1 CA45574-03.

Getty, D. J., Pickett, R. M., D'Orsi, C. J., and Swets, J. A. (1988). Enhanced interpretation of diagnostic images. *Investigative Radiology* 23(4), 240–252.

Getty, D. J., Swets, J. A., Pickett, R. M., and Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied* 1(1), 19–33.

Gigerenzer, G., and Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Green, D. M., and Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley. Reprint (with updated topical bibliography), Los Altos Hills, CA: Peninsula, 1988.

Hammond, K. R., Harvey, L. O., Jr., and Hastie, R. (1992). Making better use of scientific knowledge: Separating truth from justice. *Psychological Science* 3, 80–87.

Harris, D. H. (1991). Eddy currents steam generator data analysis performance. Paper presented at the American Society of Mechanical Engineers International Joint Power Generation Conference. San Diego, CA.

Heine, M. H. (1973). The inverse relationship of precision and recall in terms of the Swets model. *Journal of Documentation* 29, 81–84.

Kopans, D., and D'Orsi, C. J. (1992). ACR system enhances mammography reporting. *Diagnostic Imaging,* September, 125–132.

Lachenbruch, P. (1975). *Discriminant analysis.* New York: Hafner.

Lusted, L. (1968). *Introduction to medical decision making.* Springfield, IL: Thomas.

Macmillan, N. A., and Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Mackie, R. R. (Ed.) 1977. *Vigilance: Relationships among theory, operational performance, and physiological correlates.* New York: Plenum.

Mackworth, N. H (1950). *Research on the measurement of human performance.* Medical Research Council Special Report Series No. 268. London: His Majesty's Stationery Office.

Mandler, G., Pearlstone, Z., and Koopmans, H. S. (1969). Effects of organization and semantic similarity on recall and recognition. *Journal of Verbal Learning and Verbal Behavior* 8, 410–423.

Mason, I. (1982a). A model for assessment of weather forecasts. *Australian Meteorological Magazine* 37, 75–81.

Mason, I. (1982b). On scores for yes/no forecasts. Ninth Conference on Weather Forecasting and Analysis, Seattle, WA. Boston, MA: The American Meteorological Society, 169–174.

McNichol, D., and Ryder, L. A. (1971). Sensitivity and response bias effects in the learning of familiar and unfamiliar associations by rote or with mnemonic. *Journal of Experimental Psychology* 90, 81–89.

Metz, C. E., and Shen, J-H. (1992). Gains in accuracy from replicated readings of diagnostic images: Prediction and assessment in terms of ROC analysis. *Medical Decision Making* 12, 60–75.

Miller, E., and Lewis, P. (1977). Recognition memory in elderly patients with depression and dementia: A signal detection analysis. *Journal of Abnormal Psychology* 86(1), 84–86.

Murdock, B. B., Jr., (1968). Serial order effects in short-term memory. *Journal of Experimental Psychology*, Monograph Supplement 76, 1–15.

Murdock, B. B., Jr. (1974). *Human memory: Theory and data*. Hillsdale, NJ: Erlbaum.

Murdock B. B., Jr., and Duffy, P. O. (1972). Strength theory and recognition memory. *Journal of Experimental Psychology* 94(3), 284–290.

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, Series A, 289–337.

Parasuraman, R., and Davies, D. R. (1977). A taxonomic analysis of vigilance performance. In R. R. Mackie (Ed.), *Vigilance: Relationships among theory, operational performance, and physiological correlates*, 559–574. New York: Plenum.

Peterson, W. W., Birsdsall, T. G., and Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory* PGIT-4, 171–212. Reprinted by R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Readings in mathematical psychology*, 167–211. New York: Wiley, 1963.

Raser, G. (1970). Meaningfulness and signal detection theory in immediate paired-associate recognition. *Journal of Experimental Psychology* 84, 173–175.

Ries, L. A. G., Miller, B. A., Hankey, B. F., et al. (1994). Cancer statistics review 1973–91: Tables and graphics. NIH Publication No. 94-2789, National Cancer Institute.

Sarrazin, R., and Wilson, L. J. (1987). Comparison of MOS and perfect PROG probability of precipitation forecasts using the signal detection theory model. Tenth Conference on Probability and Statistics in the Atmospheric Sciences, Edmonton, Canada. Boston, MA: American Meteorological Society, 95–100.

Saxe, L., Dougherty, D., and Cross, T. (1985). The validity of polygraph testing. *American Psychologist* 40(3), 355–366.

Schwartz, J. S., Dans, P. E., and Kinosian, B. P. (1988). Human immunodeficiency virus test evaluation, performance, and use: Proposals to make good tests better. *Journal of the American Medical Association* 259, 2574–2579.

See, J. E., Howe, S. R., Warm, J. S., and Dember, W. N. (1995). A meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin* 117(2), 230–249.

Seltzer, S. E., Getty D. J., Tempany, C. M. C., Pickett, R. M., Schnall, M. D., McNeil, B. J., and Swets, J. A. (1997). Staging prostate cancer with MR imaging: A combined radiologist-computer system. *Radiology* 202, 219–226.

Seltzer, S. E., McNeil, B. J., D'Orsi, C. J., Getty, D. J., Pickett, R. M., and Swets, J. A. (1992). Combining evidence from multiple imaging modalities: A feature-analysis method. *Computerized Medical Imaging and Graphics* 16(6), 373–380.

Shakhar, C. B., Lieblich, I., and Kugelmass, S. (1970). Guilty-knowledge technique: Application of signal detection measures. *Journal of Applied Psychology* 54(5), 409–413.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology* 1, 54–87.

Shiffman, S. S., Reynolds, M. L., and Young, F. W. (1981). *Introduction to multidimensional scaling*. New York: Academic Press.

Sickles, E. A., Ominsky, S. A., Sollitto, R. A., et al. (1990). Medical audit of a rapid throughput mammography screening practice: Methodology and results of 27,114 examinations. *Radiology* 175, 323–327.

Swets, J. A. (1961). Is there a sensory threshold? *Science* 134 (3473), 168–177.

Swets, J. A. (1963). Information retrieval systems. *Science* 141 (3577), 245–250.

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation* 20, 72–89. Reprinted in B. Griffith (Ed.), *Key papers in information science*, 349–366. White Plains, NY: Knowledge Industry, 1980.

Swets, J. A. (1973). The relative operating characteristic in psychology. *Science* 182 (4116), 990–1000.

Swets, J. A. (1977). Signal detection theory applied to vigilance. In R. R. Mackie (Ed.), *Vigilance: Relationships among theory, operational performance, and physiological correlates*, 705–718. New York: Plenum.

Swets, J. A. (1983). Assessment of NDT Systems: I. The relationship of true and false detections; II. Indices of performance. *Materials Evaluation* 41, 1294–1303.

Swets, J. A. (1984). Mathematical models of attention. In R. Parasuraman and R. Davies (Eds.), *Varieties of attention*, 183–242. New York: Academic Press.

Swets, J. A. (1986a). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin* 99(1), 100–117.

Swets, J. A. (1986b). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin* 99(2), 181–198.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* 240(4857), 1285–1293.

Swets, J. A. (1991a). Normative decision making. In J. Baron and R. V. Brown (Eds.), *Teaching decision making to adolescents*, 273–296. Hillsdale, NJ: Erlbaum.

Swets, J. A. (1991b). The science of high-stakes decision making in an uncertain world. Science and Public Policy Seminar of the Federation of Behavioral, Psychological, and Cognitive Sciences, Rayburn House Office Building, Washington, D.C., September 6.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist* 47(4), 522–532.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: Erlbaum.

Swets, J. A., Getty, D. J., Pickett, R. M., D'Orsi, C. J., Seltzer, S. E., and McNeil, B. J. (1991). Enhancing and evaluating diagnostic accuracy. *Medical Decision Making* 11(1), 9–18.

Swets, J. A., Getty, D. J., Pickett, R. M., Seltzer, S. E., D'Orsi, C. J., Frenna, T., Freeman, B. F., Mateer, M., and Hermann, L. (1995). Increasing the accuracy of mammogram interpretation. Bolt Beranek and Newman Inc., Annual report to U.S. Army Medical Research and Materiel Command, Contract No. DAM017-94-C-4082.

Swets, J. A., and Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Swets, J. A., Tanner, W. P., Jr., and Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review* 68, 301–340.

Szucko, J. J., and Kleinmuntz, B. (1981). Statistical versus clinical lie detection. *American Psychologist* 36(5), 488–496.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review* 34, 273–286.

Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology* 38, 368–389.

Tanner, W. P., Jr., and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review* 61, 401–409.

Wald, A. (1950). *Statistical decision functions*. New York: Wiley.

Weinstein, M. C., and Fineberg, H. V. (1980). *Clinical decision analysis*. Philadelphia, PA: Saunders.

Williams, G. M. (1985). An evaluation of precipitation probability forecasts using signal detection theory. Ninth Conference on Probability and Statistics in Atmospheric Sciences, Virginia Beach, VA. Boston, MA: American Meteorological Society, 214–217.

About the Author

John A. Swets is Chief Scientist–Information Sciences at BBN Corporation in Cambridge, Massachusetts, a multidisciplinary think tank involved in acoustics, engineering psychology, human-computer interaction, and computer science (e.g., development of the Internet). He chairs the National Research Council Commission on Behavioral and Social Sciences and Education, and he is also a lecturer at Harvard Medical School. He writes:

> When I entered graduate school in 1950, I wanted to measure how observers were influenced by instructions when trying to detect faint stimuli (e.g., lights or tones). Through a fellow graduate student, Wilson P. Tanner, Jr., I met two graduate students in electrical engineering who were working on the problem of interpreting the output of unreliable or "noisy" electronic sensing devices. Tanner and I thought that their work, based on statistical decision theory, might be an appropriate model for a cognitive theory of stimulus detection and recognition. Our work showed how a signal detection theory based on these ideas applies to human decision making, as described in this chapter. Others have shown that the theory applies to animals as well.
>
> After teaching for eight years at the University of Michigan and at MIT, I moved to BBN because I could do research of my choice full-time. I continued research on detection processes and wrote a book with David M. Green, *Signal Detection Theory and Psychophysics* (1966). With colleagues at BBN I developed the application of signal detection theory to the identification of complex stimuli and then to an array of diagnostic tasks, such as the X-ray cancer diagnosis problem described in the chapter. It has been a career path I would follow again without hesitation.
>
> I jumped at the chance to write this chapter because it was to be addressed to undergraduates, who are told in most textbooks only about the old concept of fixed sensory thresholds. And I would like undergraduates to appreciate that signal detection theory is a general cognitive theory that goes far beyond sensory and memory processes. It is applicable to tasks of detecting or identifying objects, or discriminating between them, not only in psychological studies but in real life—be the objects lights, tones, words in memory, road signs, handwriting samples, faces in a lineup, conceptual categories, political statements, personality disorders, brain malfunctions, job applicants, or weather patterns.