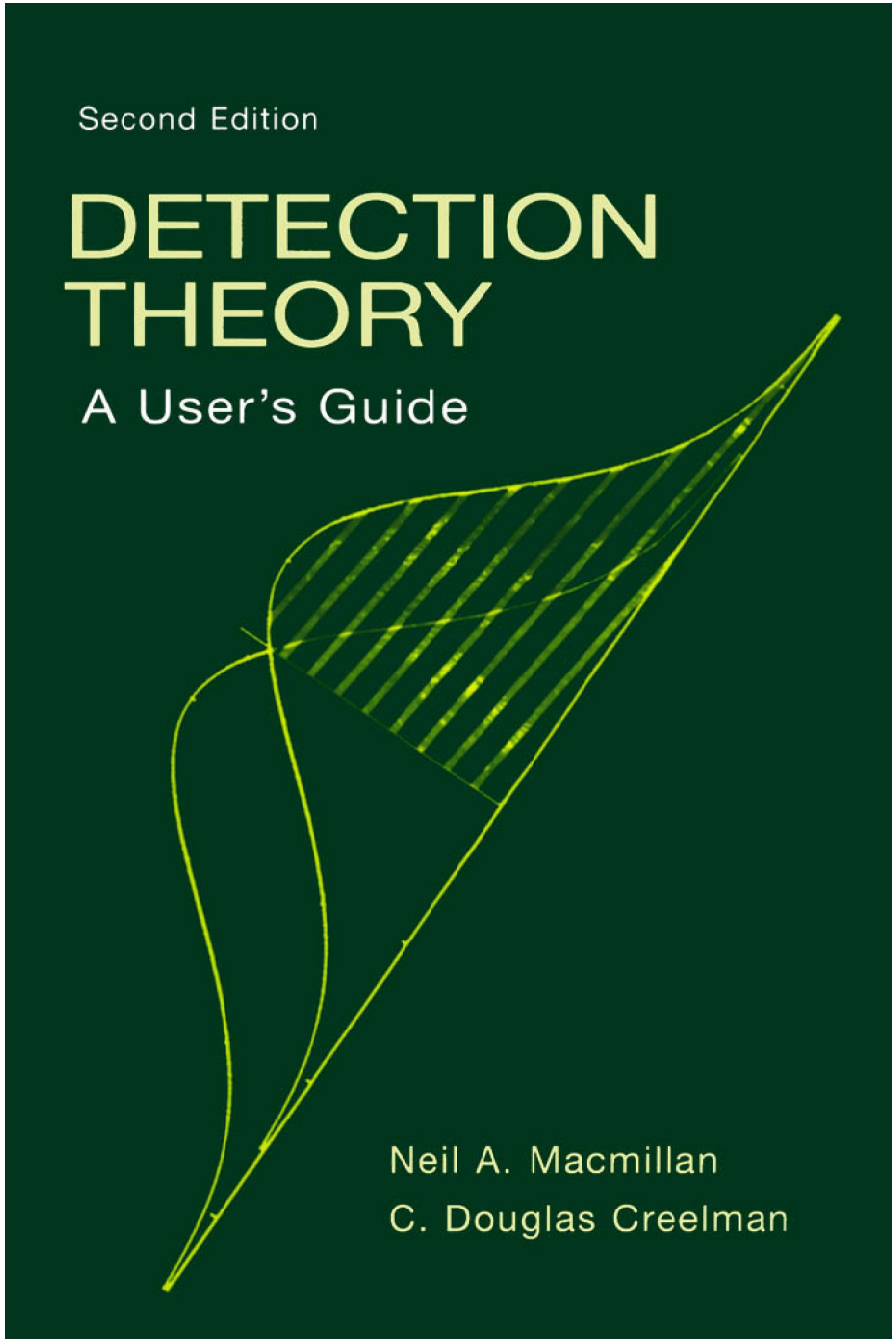

Detection Theory:
A User's Guide

(2nd edition)

This page intentionally left blank

I

Basic Detection Theory and One-Interval Designs

Part I introduces the *one-interval design*, in which a single stimulus is presented on each trial. The simplest and most important example is a correspondence experiment in which the stimulus is drawn from one of two stimulus classes and the observer tries to say from which class it is drawn. In auditory experiments, for example, the two stimuli might be a weak tone and no sound, tone sequences that may be slow or fast, or passages from the works of Mozart and Beethoven.

We begin by describing the use of one-interval designs to measure *discrimination*, the ability to tell two stimuli apart. Two types of such experiments may be distinguished. If one of the two stimulus classes contains only the null stimulus, as in the tone-versus-background experiment, the task is called *detection*. (This historically important application is responsible for the use of the term *detection theory* to refer to these methods.) If neither stimulus is null, the experiment is called *recognition*, as in the other examples. The methods for analyzing detection and recognition are the same, and we make no distinction between them (until chap. 10, where we consider experiments in which the two tasks are combined).

In chapters 1 and 2, we focus on designs with two possible responses as well as two stimulus classes. Because the possible responses in some applications (e.g., the tone detection experiment) are "yes" and "no," the paradigm with two stimuli, one interval, and two responses is sometimes termed *yes-no* even when the actual responses are, say, "slow" and "fast." Performance can be analyzed into two distinct elements: the degree to which the observer's responses mirror the stimuli (chap. 1) and the degree to which they display bias (chap. 2). Measuring these two elements requires a theory; we use the most common, normal-distribution variant of detection theory to

accomplish this end. Chapter 4 broadens the perspective on yes-no sensitivity and bias to include three classes of alternatives to this model: threshold theory, choice theory, and "nonparametric" techniques.

One-interval experiments may involve more than two responses or more than two possible stimuli. As an example of a larger response set, listeners could rate the likelihood that a passage was composed by Mozart rather than Beethoven on a 6-point scale. One-interval rating designs are discussed in chapter 3. As an example of a larger stimulus set, listeners could hear sequences presented at one of several different rates. If the requirement is to assign a different response to each stimulus, the task is called *identification*; if the stimuli are to be sorted into a smaller number of classes (perhaps slow, medium, and fast), it is *classification*. Chapter 5 applies detection-theory tools to identification and classification tasks, but only those in which elements of the stimulus sets differ in a single characteristic such as tempo. Identification and classification of more heterogeneous stimulus sets are considered in Part II.

1

The Yes-No Experiment: Sensitivity

In this book, we analyze experiments that measure the ability to distinguish between stimuli. An important characteristic of such experiments is that observers can be more or less *accurate*. For example, a radiologist's goal is to identify accurately those X-rays that display abnormalities, and participants in a recognition memory study are accurate to the degree that they can tell previously presented stimuli from novel ones. Measures of performance in these kinds of tasks are also called *sensitivity measures: High sensitivity* refers to good ability to discriminate, *low sensitivity* to poor ability. This is a natural term in detection studies—a sensitive listener hears things an insensitive one does not—but it applies as well to the radiology and memory examples.

Understanding Yes-No Data

Example 1: Face Recognition

We begin with a memory experiment. In a task relevant to understanding eyewitness testimony in the courtroom, participants are presented with a series of slides portraying people's faces, perhaps with the instruction to remember them. After a period of time (and perhaps some unrelated activity), recognition is tested by presenting the same participants with a second series that includes some of the same pictures, shuffled to a new random order, along with a number of "lures"—faces that were not in the original set. Memory is good if the person doing the remembering properly recognizes the Old faces, but not New ones. We wish to measure the ability to distinguish between these two classes of slides. Experiments of this sort have been performed to compare memory for faces of different races, orientations (upright vs. inverted), and many other variables (for a review, see Shapiro & Penrod, 1986).

Let us look at some (hypothetical) data from such a task. We are interested in just one characteristic of each picture: whether it is an Old face (one presented earlier) or a New face. Because the experiment concerns two kinds of faces and two possible responses, "yes" (I've seen this person before in this experiment) and "no" (I haven't), any of four types of events can occur on a single experimental trial. The number of trials of each type can be tabulated in a stimulus-response matrix like the following.

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old	20	5	25
New	10	15	25

The purpose of this yes-no task is to determine the participant's sensitivity to the Old/New difference. High sensitivity is indicated by a concentration of trials counted in the upper left and lower right of the matrix ("yes" responses to Old stimuli, "no" responses to New).

Summarizing the Data

Conventional, rather military language is used to describe the yes-no experiment. Correctly recognizing an Old item is termed a *hit*; failing to recognize it, a *miss*. Mistakenly recognizing a New item as old is a *false alarm*; correctly responding "no" to an Old item is, abandoning the metaphor, a *correct rejection*. In tabular terms:

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	Hits (20)	Misses (5)	(25)
New (S_1)	False alarms (10)	Correct rejections (15)	(25)

We use S_1 and S_2 as context-free names for the two stimulus classes.

Of the four numbers in the table (excluding the marginal totals), only two provide independent information about the participant's performance. Once we know, for example, the number of hits and false alarms, the other two entries are determined by how many Old and New items the experimenter decided to use (25 of each, in this case). Dividing each number by the total in its row allows us to summarize the table by two numbers: The *hit rate* (H) is the proportion of Old trials to which the participant responded "yes," and the *false-alarm rate* (F) is the proportion of New trials similarly (but incorrectly) assessed. The hit and false-alarm rates can be written as conditional probabilities[1]

$$H = P(\text{"yes"} \mid S_2) \tag{1.1}$$

$$F = P(\text{"yes"} \mid S_1)\,, \tag{1.2}$$

where Equation 1.1 is read "The proportion of 'yes' responses when stimulus S_2 is presented."

In this example, $H = .8$ and $F = .4$. The entire matrix can be rewritten with response rates (or proportions) rather than frequencies:

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	.8	.2	1.0
New (S_1)	.4	.6	1.0

The two numbers needed to summarize an observer's performance, F and H, are denoted as an ordered (*false-alarm, hit*) pair. In our example, (F, H) = (.4, .8).

Measuring Sensitivity

We now seek a good way to characterize the observer's sensitivity. A function of H and F that attempts to capture this ability of the observer is called a sensitivity *measure*, *index*, or *statistic*. A perfectly sensitive participant would have a hit rate of 1 and a false-alarm rate of 0. A completely insensitive participant would be unable to distinguish the two stimuli at all and, indeed, could perform equally well without attending to them. For this observer, the probability of saying "yes" would not depend on the stimulus presented, so the hit and false-alarm rates would be the same. In interesting cases, sensitivity falls between these extremes: H is greater than F, but performance is not perfect.

[1] Technically, H and F are *estimates* of probabilities—a distinction that is important in statistical work (chap. 13). Probabilities characterize the observer's relation to the stimuli and are considered stable and unchanging; H and F may vary from one block of trials to the next.

Perfectly sensitive participant

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	1.0	0.0	1.0
New (S_1)	0.0	1.0	1.0

Interesting cases falls between these two extreme observers

H > FA but performance **is not perfect**

Completely insensitive participant

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	.5	.5	1.0
New (S_1)	.5	.5	1.0

The simplest possibility is to ignore one of our two response rates using, say, H to measure performance. For example, a lie detector might be touted as detecting 80% of liars or an X-ray reader as detecting 80% of tumors. (Alternatively, the hit rate might be ignored, and evaluation might depend totally on the false-alarm rate.) Such a measure is clearly inadequate. Compare the memory performance we have been examining with that of another group:

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old	8	17	25
New	1	24	25

Group 1 successfully recognized 80% of the Old words, Group 2 just 32%. But this comparison ignores the important fact that Group 2 participants just did not say "yes" very often. The hit rate, or any measure that depends on responses to only one of the two stimulus classes, cannot be a measure of sensitivity. To speak of sensitivity to a stimulus (as was done, for instance, in early psychophysics) is meaningless in the framework of detection theory.[2]

An important characteristic of sensitivity is that it can only be measured between two alternative stimuli and must therefore depend on both H and F. A moment's thought reveals that not all possible dependencies will do. Certainly a higher hit rate means greater, not less, sensitivity, whereas a higher false-alarm rate is an indicator of less sensitive performance. So a sensitivity measure should increase when either H increases or F decreases.

A final possible characteristic of sensitivity measures is that S_1 and S_2 trials should have equal importance: Missing an Old item is just as important an error as incorrectly recognizing a New one. In general, this is too strong a requirement, and we will encounter sensitivity measures that assign different weights to the two stimulus classes. Nevertheless, equal treatment is a good starting point, and (with one exception) the indexes described in this chapter satisfy it.

[2]The term *sensitivity is* used in this way, as a synonym for the hit rate, in medical diagnosis. *Specificity* is that field's term for the correct-rejection rate.

Two Simple Solutions

We are looking for a measure that goes up when H goes up, goes down when F goes up, and assigns equal importance to these statistics. How about simply subtracting F from H? The difference $H - F$ has all these characteristics. For the first group of memory participants, $H - F = .8 - .4 = .4$; for the second, $H - F = .32 - .04 = .28$, and Group 1 wins.

Another measure that combines H and F in this way is a familiar statistic, the proportion of correct responses, which we denote $p(c)$. To find proportion correct in conditions with equal numbers of S_1 and S_2 trials, we take the average of the proportion correct on S_2 trials (the hit rate, H) and the proportion correct on S_1 trials (the correct rejection rate, $1 - F$). Thus:

$$p(c) = \tfrac{1}{2}[H + (1 - F)] = \tfrac{1}{2}(H - F) + \tfrac{1}{2} \quad . \tag{1.3}$$

If the numbers of S_1 and S_2 trials are not equal, then to find the literal proportion of trials on which a correct answer was given the actual numbers in the matrix would have to be used:

★ $p(c)^*$ = (hits + correct rejections)/total trials . (1.4)

Usually it is more sensible to give H and F equal weight, as in Equation 1.3, because a sensitivity measure should not depend on the base presentation rate.

Let us look at $p(c)$ for equal presentations (Eq. 1.3). Is this a better or worse measure of sensitivity than $H - F$ itself? Neither. Because $p(c)$ depends directly on $H - F$ (and not on either H or F separately), one statistic goes up whenever the other does, and the two are monotonic functions of each other. Two measures that are monotonically related in this way are said to be *equivalent* measures of accuracy. In the running examples, $p(c)$ is .7 for Group 1 and .64 for Group 2, and $p(c)$ leads to the same conclusion as $H - F$. For both measures, Group 1 outscores Group 2.

A Detection Theory Solution

The most widely used sensitivity measure of detection theory (Green & Swets, 1966) is not quite as simple as $p(c)$, but bears an obvious family re-

Group 2

Stimulus-response matrix

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old	8	17	25
New	1	24	25

$$H - F = \frac{8}{25} - \frac{1}{25} = .32 - .04 = .28$$

$$p(c)^* = \frac{\text{(hits+correct rejections)}}{\text{Total}} = \frac{8 + 24}{25 + 25} = .64$$

Nel caso di totali di riga uguali:

$$p(c) = \frac{(8 + 24)}{2 \times (25)} = \frac{1}{2} \times \frac{(8 + 24)}{25} = \frac{1}{2} \times \left(\frac{8}{25} + \frac{24}{25}\right) = \frac{1}{2} \times [H + (1 - F)] = .64$$

Perfectly sensitive participant

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	1.0	0.0	1.0
New (S_1)	0.0	1.0	1.0

$$\boldsymbol{H - F} = 1 - 0 = 1$$

$$\boldsymbol{p(c)} = \frac{1}{2} \times [H + (1 - F)]$$

$$= \frac{1}{2} \times [1.0 + (1 - 0.0)] = 1$$

Completely insensitive participant

Stimulus Class	*Response*		
	"Yes"	"No"	Total
Old (S_2)	.5	.5	1.0
New (S_1)	.5	.5	1.0

$$\boldsymbol{H - F} = .5 - .5 = 0$$

$$\boldsymbol{p(c)} = \frac{1}{2} \times [H + (1 - F)]$$

$$= \frac{1}{2} \times [.5 + (1 - .5)] = \frac{1}{2}$$

semblance. The measure is called d' ("dee-prime") and is defined in terms of z, the inverse of the normal distribution function:

$$d' = z(H) - z(F)\,. \tag{1.5}$$

The z transformation converts a hit or false-alarm rate to a z score (i.e., to standard deviation units). A proportion of .5 is converted into a z score of 0, larger proportions into positive z scores, and smaller proportions into negative ones. To compute z, consult Table A5.1 in Appendix 5. The table makes use of a symmetry property of z scores: Two proportions equally far from .5 lead to the same absolute z score (positive if $p > .5$, negative if $p < .5$) so that:

$$z(1 - p) = -z(p)\,. \tag{1.6}$$

Thus, $z(.4) = -.253$, the negative of $z(.6)$. Use of the Gaussian z transformation is dominant in detection theory, and we often refer to normal-distribution models by the abbreviation *SDT*.

We can use Equation 1.5 to calculate d' for the data in the memory example. For Group 1, $H = .8$ and $F = .4$, so $z(H) = 0.842$, $z(F) = -0.253$, and $d' = 0.842 - (-0.253) = 1.095$. When the hit rate is greater than .5 and the false-alarm rate is less (as in this case), d' can be obtained by adding the absolute values of the corresponding z scores. For Group 2, $H = .32$ and $F = .04$, so $d' = -0.468 - (-1.751) = 1.283$. When the hit and false-alarm rates are on the same side of .5, d' is obtained by subtracting the absolute values of the z scores. Interestingly, by the d' measure, it is Group 2 (the one that was much more stingy with "yes" responses) rather than Group 1 that has the superior memory.

When observers cannot discriminate at all, $H = F$ and $d' = 0$. Inability to discriminate means having the same rate of saying "yes" when Old faces are presented as when New ones are offered. As long as $H \geq F$, d' must be greater than or equal to 0. The largest possible *finite* value of d' depends on the number of decimal places to which H and F are carried. When $H = .99$ and $F = .01$, $d' = 4.65$; many experimenters consider this an effective ceiling.

Perfect accuracy, on the other hand, implies an infinite d'. Two adjustments to avoid infinite values are in common use. One strategy is to convert proportions of 0 and 1 to $1/(2N)$ and $1 - 1/(2N)$, respectively, where N is the number of trials on which the proportion is based. Suppose a participant has 25 hits and 0 misses ($H = 1.0$) to go with 10 false alarms and 15 correct rejections ($F = .4$). The adjustment yields 24.5 hits and 0.5 misses, so $H = .98$ and $d' = 2.054 - (-0.253) = 2.307$. A second strategy (Hautus, 1995; Miller, 1996) is to add 0.5 to *all* data cells regardless of whether zeroes are present. This adjustment leads to $H = 25.5/26 = .981$ and $F = 10.5/26 = .404$. Rounding to two decimal places yields the same value as before, but d' is slightly smaller if computed exactly.

Essay: The Provenance of Detection Theory

Psychophysics, the oldest psychology, has continually adapted itself to the substantive concerns of experimentalists. In particular, detection theory is well suited to cognitive psychology and might indeed be considered one of its sources. No grounding in history is needed to use this book, but some appreciation of the intellectual strains that meet here will help place these tools in context.

The term *psychophysics* was invented by Gustav Fechner (1860), the 19th-century physicist, philosopher, and mystic. He was the first to take a mathematical approach to relating the internal and external worlds on the basis of experimental data. Some present-day psychophysicists directly pursue Fechner's interest in relating mental experience to the physical world, usually in simple perceptual experiments. Measuring the way in which the reported experience of loudness grows with physical intensity is a psychophysical problem of this sort; we consider a detection theory approach to this problem in chapter 5.

This book is part of a second Fechnerian legacy, also methodological, but more general than the first. Fechner developed, tested, and described experimental methods for estimating the *difference threshold*, or *just noticeable difference (jnd)*, the minimal difference between two stimuli that leads to a change in experience. Fechner's assumption that the jnd could be the unit of measurement, the fundamental building block or atom of experience, was central to Wundt's and Titchener's structuralism, the first experimentally based theory of perception. The analogy to 19th-century chemistry was close: Theory and experiment should focus on uncovering the basic units and the laws of combination of those units.

Fechner's methods were adopted and became topics of investigation in their own right; they still form the backbone of experimental psychology. Attempts to measure jnds led to two complications: (a) The threshold appeared not to be a fixed quantity because, as the difference between two stimuli increases, correct discrimination becomes only gradually more likely (Urban, 1908); and (b) different methods produced different values for the jnd.

The concept of the jnd survived the first problem by redefinition: The jnd is now considered to be the stimulus difference that can be discriminated on some fixed percentage of trials (see chaps. 5 and 11). Two early reactions to the problem of continuity in psychophysical data are recognizable in modern research (see Jones, 1974).

One line of thought retained the literal notion of a sensory threshold, building mechanical and mathematical models to explain the gradual nature of observed functions (see chap. 4 for the current status of such models). The threshold idea was congenial with early 20th-century behaviorist and operationist attitudes: Sensory function could be studied and measured without invoking unpopular notions of mental content (Garner, Hake, & Eriksen, 1956). The threshold, in this view, was a construct derived from data and did not have to relate to any internal and unobservable mental process. The solution to method dependence was merely to subscript thresholds to indicate the method by which they were obtained (Graham, 1950; Osgood, 1958).

The second response to the variability problem, instigated, according to Jones (1974), by Delboef (1883), substituted a continuum of experience for the discrete processes of the threshold; it is this view that informs most contemporary psychophysics. One approach to measuring such continuous experience was Stevens' (1975) magnitude estimation, which used direct verbal estimates. Detection measurement, in contrast, relies on underlying random variation or noise. Psychologists' realization of the importance of random variation dates at least to Fullerton and Cattell (1892), who invoked it in a rigorous quantitative way to account for inconsistency in response with repetitions of identical stimuli. Variability later served as the key building block for the pioneering work of Thurstone (1927a, 1927b) in measuring distances along sensory continua indirectly.

The idea of variability or noise as an explanatory concept also arose in engineering, with the development and evaluation of radar detection apparatus. Radar and sonar are limited in performance by intrinsic noise in the input signal. Any input from an antenna or sensor can be due to noise alone or to a signal of interest embedded in the background noise. Groups at the University of Michigan (Peterson, Birdsall, & Fox, 1954), MIT (van Meter & Middleton, 1954), and in the Soviet Union (Kotel'nikov, 1960) recognized that the physical noise that was mixed with all signals, and that could mimic signal presence, was a major limitation to detection performance.

Knowing that stimulus environments are noisy does not, in itself, tell an observer how best to cope with them. An approach to this problem was contributed by another applied science: statistical decision theory. Decision theorists pointed out that information derived from noisy signals could lead to action only when evaluated against well-defined goals. Decisions (and thus action) should depend not only on the stimulus, but on the expected outcomes of actions. The viewer of a radar display that might or might not

contain a blip, for example, should consider the relative effects of failing to detect a real bomber and of detecting a phantom before deciding on a response to that display.

W. P. Tanner, Jr., working with J. A. Swets at the University of Michigan, realized that these engineering notions could be applied to psychology and appropriated them directly into the psychophysical experiment (Tanner & Swets, 1954). By separating the world of stimuli and their perturbations from that of the decision process, detection theory was able to offer measures of performance that were not specific to procedure and that were independent of motivation. Procedure and motivation could influence data, but affected only the decision process, leaving measurable aspects of the internal stimulus world unchanged and capable of being evaluated separately.

According to detection theory, the observer's access to the stimuli being discriminated is indirect: An intelligent, not entirely reliable process makes inferences about them and acts according to the demands of the experimental situation. One might say that detection theory "deals with the processes by which [a decision about] a perceived, remembered, and thought-about world is brought into being from [an] unpromising beginning" (Neisser, 1967, p. 4). Neisser's landmark book linked perception and cognition into a unified framework after a hiatus of many decades. The constructionist (although not complicated) decision processes of detection theory mark it as an early example of cognitive psychology. The ideas behind detection theory are the everyday assumptions of behavioral experimenters in the cognitive era, and the theory itself is central to a wide range of research areas in cognitive science. Perhaps Estes' (2002) assessment is not an overstatement: "... [SDT is] the most towering achievement of basic psychological research of the last half century" (p. 15).

Summary

The results of a one-interval discrimination experiment can be described by a hit and a false-alarm rate, which in turn can be reduced to a single measure of sensitivity. Good indexes can be written as the difference between the hit and false-alarm rates when both are appropriately transformed. The sensitivity measure proposed by detection theory, d', uses the normal-distribution z transformation. The primary rationale for d' as a measure accuracy is that it is roughly invariant when response bias is manipulated; simpler indexes such as proportion correct do not have this property. The use of d' implies a model in which the two possible stimulus classes lead to normal distributions differing in mean, and the observer decides which class occurred by comparing an observation with an adjustable criterion.

Conditions under which the methods described in this chapter are appropriate are spelled out in Chart 2 of Appendix 3.

Problems

1.1. Suppose you are measuring the sensitivity of a polygraph ("lie detector"). What are "hits," "misses," "false alarms," and "correct rejections"?

1.2. The following tables give the number of trials in three conditions of a detection experiment on which participants responded "yes" or "no" to S_1 or S_2. (a) Calculate H and F. (b) Find $H - F$, $p(c)$, and $p(c)$*. For these data sets, can $H - F$ be greater than $p(c)$ in one case and the reverse ordering occur in another, or is one index *always* greater than the other?

(a)	"yes"	"no"
S_2	9	6
S_1	7	8

(b)	"yes"	"no"
S_2	55	45
S_1	5	25

(c)	"yes"	"no"
S_2	45	55
S_1	25	5

1.3 (a). In Problem 1.2(a), the numbers of S_1 and S_2 trials are equal, but in (b) and (c) they are not. Does this matter computationally? experimentally?
(b). Is it possible to calculate $p(c)$ for S_2 trials only? What would this statistic measure?

1.4. Compute d' for the following (F, H) pairs:
(a) (.16, .84), (.01, .99), (.75, .75).
(b) (.6, .9), (.5, .9), (.05, .9).