Signal Detection Theory

Michael S. Landy

Dept. of Psychology and Center for Neural Science

New York University

Landy, M. S. (in press). Signal detection theory. *Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences*, Volume 3, Chapter 24. New York: Cambridge University Press.

Abstract

Signal detection theory (SDT) is the primary performance model and data-analysis method for sensory experiments (in audition, vision, etc.)—detecting whether a stimulus is present or discriminating between two possible stimuli. SDT allows the experimenter to separately estimate discriminability (the observer's ability to discriminate the presence or absence of the stimulus) and bias (the observer's preference to respond "yes" or "no"). Applications of SDT are widespread—to sensory coding, memory, value-based decision-making, and analysis of the information content of neural spiking. The same analysis has also been used in applied settings to understand the performance of baggage screeners, disease diagnosis from medical images, and the efficacy of medical diagnostic tests. I review the theory and discuss applications to behavioral and neural responses. I review the application of SDT to discrimination judgments and metacognition—the observer's confidence in those judgments.

Keywords: signal detection, ROC curve, decision-making, metacognition, d-prime

Introduction

Suppose you are out on a foggy night and looking down the street ahead of you. You hear a sound that seems like footsteps and get the vague visual impression of someone walking toward you. Is there someone there or not? Signal detection theory suggests that, somewhere in your brain, you combine all the evidence for the presence of a person—a faint smell of perfume, the sound of footsteps, a faint outline that resembles a human figure—resulting in a single number that represents the strength of the evidence. That number will typically be small if no one is out there, and large if someone is there.

If one repeats this experience multiple times, the strength of the evidence will vary across occasions, *even if the circumstance (absence or presence) doesn't change*. That is, the strength of the evidence is random—on average higher when someone is

there and lower when no one is there. This stochasticity of the evidence can come from many sources, both external and internal. External randomness can be from the varying density of the fog, the variation in sound, smell and body outline across people and across viewpoints, and even from the randomness of the stimulus itself, such as the random number of photons arriving at the retina from a dim location in the scene. Randomness can be internal to the observer, such as the randomness of neural responses to repeated, identical stimuli. Signal detection theory provides a model of how observers derive a binary response (the person is there or isn't) using this noisy evidence.

Signal detection theory (Peterson et al., 1954; Van Meter & Middleton, 1954), also sometimes called sensory decision theory, grew out of statistical decision theory (all of which conveniently have the acronym SDT). It provides both a theory of how decisions under uncertainty are made, as well as a method for analyzing behavioral and neural data (Green & Swets, 1966; Macmillan & Creelman, 1991; Wickens, 2002). SDT is applicable to a wide variety of tasks, including sensory experiments (vision, audition, touch, proprioception, etc.), neuroeconomic experiments (where the intensive parameter is value), and experiments on memory (where the intensive parameter is the memory strength). It also applies in everyday life when making yes-or-no decisions based on uncertain evidence (e.g., Does this suitcase contain a weapon? Does this mammogram indicate breast cancer?).

Signal Detection Theory: Optimal Decision-Making

Measurement model

In standard signal detection theory, for a given decision (e.g., one trial of an experiment), there are two possible states of the world—either there is a signal ("S"), or there is no signal ("N"). It is the observer's task to determine whether the world is in state "S" or "N". In our example above, "S" represents a scene in which a person is approaching you. The observer makes an observation of that world state (views a visual display, listens to an auditory stimulus, etc.), resulting in a measurement—a single number (the "decision variable" x) that summarizes the evidence concerning the state of the world. We assume that x is typically larger when a signal is present (e.g., when someone is approaching) than when it is absent. The decision variable also varies from trial to trial even when the stimulus is fixed. For any given state of the world, we assume that the distribution of x is Gaussian (i.e., x is normally distributed) and that the variance of this distribution is fixed—not depending on whether a signal is present or not. Thus, the measurement model is (Fig. 1A):

$$p(x\,|\,N) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu_N)^2}{2\sigma^2}\right] \text{ and}$$
$$p(x\,|\,S) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu_S)^2}{2\sigma^2}\right]. \quad (1)$$

Signals typically lead to larger values of x than when no signal is present, $\mu_S > \mu_N$. Signal strength corresponds to the difference in the means $\mu_S - \mu_N$.

INSERT FIGURE 1 ABOUT HERE

Observer's perspective and the SDT model

The measurement model above describes the situation with which an observer is confronted. The world is either in state S or N (a person is approaching or not) and provides the observer with a noisy measurement x (the sensory evidence). The observer knows the value of x and wants to infer whether the state of the world is S or N, that is, whether to say "yes" or "no". The probabilities given in Eq. 1 are *measurement distributions*—the probability of getting any particular measurement x on a trial when the state of the world is, for example, S. But, for the observer, these probabilities are, in a sense, backwards. The decision-maker *knows* the value of x. What the decision-maker doesn't know is the true state of the world. Thus, from the decision-maker's perspective, $p(x \mid S)$ is the probability of getting the measurement x (that the observer knows, and therefore is no longer random) given a particular state of the world S (that the observer doesn't know). When a conditional probability is regarded in this way (the value to the left of the "I" is known and fixed, and the value after it is unknown and to be estimated or decided upon), it is referred to as a *likelihood*.

What decision should the observer make? Referring to Fig. 1B, the observer receives measurement x and thus knows the two likelihoods $p(x \mid N)$ and $p(x \mid S)$ (the values of the two curves above the measurement). A simple decision procedure is to choose the world state that is more likely:

$$\begin{aligned} &\text{Say "yes" if } p(x \mid S) > p(x \mid N) \\ &\text{Say "no" otherwise.} \end{aligned} \tag{2}$$

This is called the *maximum-likelihood* (or ML) observer. We will see below that this is a good decision rule in certain circumstances but is not always optimal. Note that the curve for $p(x \mid S)$ is always above that for $p(x \mid N)$ to the right of where the two curves cross. The ML rule is (Fig. 1C):

$$\begin{aligned} &\text{Say "yes" if } x > \frac{\mu_S + \mu_N}{2} \\ &\text{Say "no" otherwise.} \end{aligned} \tag{3}$$

The rule is that the observer should compare the evidence to a fixed *criterion* (here, the criterion is $c = (\mu_S + \mu_N)/2$) and say "yes" when the measurement exceeds the decision criterion.

The ML rule may seem a little ad hoc, since likelihood is a slippery concept; it is the probability of something you already know to be true (the measurement). It makes more sense to compare the probabilities of the two events you don't know—the possible states of the world. Thus, one might prefer to adopt the following decision rule:

$$\begin{aligned} &\text{Say “yes” if } p(S \mid x) > p(N \mid x) \\ &\text{Say “no” otherwise.} \end{aligned} \tag{4}$$

Notice that the only difference between Eqs. 2 and 4 is the order of the items in the conditional probabilities. The curves in Fig. 1A provide the values of the likelihood of each state of the world given the measurement. For the observer to determine the probabilities in Eq. 4, we need to apply Bayes' Rule to each term:

$$\begin{aligned} P(S \mid x) &= \frac{p(x \mid S)P(S)}{p(x)} \\ P(N \mid x) &= \frac{p(x \mid N)P(N)}{p(x)} . \end{aligned} \tag{5}$$

In these equations, we have new terms $P(S)$ and $P(N) = 1 - P(S)$. These are called *prior* probabilities (see Vol. 1, Chapter 26). They are the probabilities of the two possible states of the world *prior* to collecting the evidence x. For example, if you were out waiting for a friend who said they would arrive around this time, $P(S)$ would be high. In a lab experiment, there is typically an equal number of signal and no-signal trials, $P(S) = P(N) = 0.5$.

$P(S \mid x)$ and $P(N \mid x)$ are *posterior* probabilities—they are the probabilities of the two possible states of the world *after* the measurement is made (I'm using the standard convention of denoting probabilities of discrete events as P and probability densities for continuous domains as p). Thus, the decision procedure in Eq. 4 is known as the *maximum a posteriori* (or MAP) rule, as it chooses the state of the world with maximum posterior probability. The term in the denominator in Eq. 5 is a nuisance term that ensures that $P(S \mid x) + P(N \mid x) = 1$. Fortunately, we won't need to compute it. Substituting Eq. 5 into Eq. 4, the MAP rule becomes

$$\begin{aligned} &\text{Say “yes” if } \frac{p(x \mid S)}{p(x \mid N)} > \frac{p(N)}{p(S)} = \beta_{\text{opt}} \\ &\text{Say “no” otherwise.} \end{aligned} \tag{6}$$

The value on the left-hand side is called a *likelihood ratio*. It is the ratio of the values of the two curves above the measurement (Fig. 1B). The right-hand side is called the *prior odds* and provides a criterion, β_{opt}, that the likelihood ratio must exceed to say "yes". If the prior odds are equal to one (i.e., no-signal and signal trials are equally likely to occur), then Eq. 6 yields the same decision procedure as maximum likelihood (Eq. 2). Fig. 1D illustrates the value of the likelihood ratio as a function of x. As you can see, the likelihood ratio increases monotonically so that Eq. 6, a criterion on likelihood ratio, will again result in a procedure that compares the strength of the evidence to a criterion:

$$\begin{aligned} &\text{Say “yes” if } x > c_{\text{opt}} \\ &\text{Say “no” otherwise.} \end{aligned} \tag{7}$$

For the MAP procedure, the optimal criterion c_{opt} will depend on the means (μ_S and μ_N), the common standard deviation (σ), and the prior odds.

INSERT FIGURE 2 ABOUT HERE

There are two possible states of the world and two possible decision outcomes:

		Decision: "Yes"	Decision: "No"
State of the world	*S*	Hit	Miss
	N	False Alarm	Correct Reject

Once the various elements of the theory are known (the specifics of the two measurement distributions and the criterion), the theory predicts the probability of each of these four possible trial outcomes (Fig. 2A,B). This 2x2 set of possible outcomes should sound familiar. It's the same 2x2 one encounters in typical descriptions of hypothesis tests in statistics (see Vol. 1, Chapter 25). Type I error, or α, corresponds to the false-alarm rate, $P(\text{“yes”} \mid N)$—rejecting the null hypothesis when it is correct. Type II error, or β, corresponds to the miss rate, $P(\text{“no”} \mid S)$—accepting the null hypothesis when it is false. The same 2x2 appears in medical decision-making, where diagnostic tests for disease are rated by their sensitivity (the probability of detecting the disease when the patient is sick), $P(\text{“yes”} \mid S)$, and specificity (the probability of failing to detect the disease when the patient is healthy), $P(\text{“no”} \mid N)$.

Examining Eq. 6, if signal trials are more prevalent than no-signal trials, the prior odds, $P(N)/P(S)$ will be low and thus the observer will require only a small likelihood ratio to lead to a "yes" response. In other words, the criterion c will be low, so that only weak evidence is required to say "yes" (Fig. 2C). In our example, if you already expected a friend was arriving, then the slightest hint of an approaching person will lead to a conclusion that your friend is arriving. The result of a low criterion is a high hit rate (correct "yes" responses when the signal is present) and low correct-reject rate (correct "no" responses when the signal is absent). Conversely, if no-signal trials are more prevalent (e.g., you are on a road that is rarely travelled), the resulting criterion will be high, so that stronger evidence is required to say "yes" (Fig. 2D). The result is a low hit rate as well as a high correct-reject rate. The rates of these two correct responses trade off as the criterion is varied.

The ideal observer

Among other fields, sensory neuroscience has been heavily influenced by the idea of an optimal, normative or ideal observer (Geisler, 1989). Human performance can be compared to predicted optimal performance to determine human efficiency at a given task. Consider a visual signal-detection task in which the observer discriminates a small, briefly presented visual pattern vs. a uniform gray field. If you place that image on a known place on the observer's retina, then one can calculate the expected number of photons landing on each receptor for the uniform field and for the patterned stimulus.

The ideal observer calculates a weighted sum of the photon catches of the receptors. For dim but not completely dark conditions, the resulting predictions are isomorphic to standard signal detection theory.

For signal detection theory, ideal observers were developed from the very start (Green & Swets, 1966). I now develop the ideal observer for standard signal detection theory. To do this, we need to decide on a "cost function" (i.e., what is it that we are trying to optimize?). I assume the observer is aware of the value of each possible decision outcome (hit, false alarm, etc.):

		Response	
		"Yes"	"No"
Stimulus	*S*	$V(S, \text{“yes”})$	$V(S, \text{“no”})$
	N	$V(N, \text{“yes”})$	$V(N, \text{“no”})$

Here, the values of this *payoff matrix* might be in units of monetary payoff or in units of psychological utility. Typically, the values associated with correct answers (hits, correct rejects) are positive, and the other two values are negative (i.e., losses). I assume again that the observer is aware of the design of the experiment and, in particular, the prior probability that the signal is present, $P(S)$. The simplest payoff matrix is symmetric, resulting in a gain for correct answers (hits and correct rejects) and a loss for incorrect answers (false alarms and misses). But, real-world examples often have strongly asymmetric payoff matrices. For an airport baggage screener, a false alarm just leads to a more careful search of a suitcase and an annoyed and inconvenienced passenger. A miss, on the other hand, can lead to an attempt to blow up the plane! The case of a radiologist examining a mammogram is similarly asymmetric.

The ideal observer is supplied with a measurement x and maps that measurement to a response ("yes" or "no") by choosing the response that maximizes the expected gain—the average value the observer will gain if a trial with that measurement and response were repeated a large number of times. The expected gain of each response depends on the various probabilities and associated values:

$$\begin{aligned} \mathbb{E}\left[V(\text{“yes”}\,|\,x)\right] &= V(S, \text{“yes”})P(S\,|\,x) + V(N, \text{“yes”})P(N\,|\,x) \\ \mathbb{E}\left[V(\text{“no”}\,|\,x)\right] &= V(S, \text{“no”})P(S\,|\,x) + V(N, \text{“no”})P(N\,|\,x), \end{aligned} \tag{8}$$

where $\mathbb{E}[]$ denotes expected value. I am computing the expectation of the value of a given response (e.g., the expectation of $V(\text{“yes”}\,|\,x)$). This requires an expectation because the value depends on the true state of the world, and the evidence only specifies the probability of each possible state. The expected value of each response is equal to the value of that response in any given world state (e.g., "S") times the probability of that world state given the evidence (e.g., $P(S\,|\,x)$). This is summed over all possible world states ("S" and "N"). The ideal observer responds "yes" when $\mathbb{E}\left[V(\text{“yes”}\,|\,x)\right] \geq \mathbb{E}\left[V(\text{“no”}\,|\,x)\right]$. Substituting Eq. 8 into this inequality we find that the ideal observer should

$$\text{Say “yes” if } \frac{P(S \mid x)}{P(N \mid x)} \geq \frac{V(N, \text{“no”}) - V(N, \text{“yes”})}{V(S, \text{“yes”}) - V(S, \text{“no”})} \qquad (9)$$

Say “no” otherwise.

Thus, the ideal observer says “yes” when the posterior odds (the ratio on the left-hand side) exceeds a criterion derived from the payoff matrix. This criterion is the excess value of being correct (rather than incorrect) on no-signal trials (the numerator) divided by the excess value of being correct on signal trials.

Making the same substitutions as we did for the MAP decision procedure, the maximum-expected-gain decision rule becomes

$$\text{Say “yes” if } \frac{p(x \mid S)}{p(x \mid N)} \geq \frac{P(N)}{P(S)} \frac{V(N, \text{“no”}) - V(N, \text{“yes”})}{V(S, \text{“yes”}) - V(S, \text{“no”})} = \beta_{\text{opt}} \qquad (10)$$

Say “no” otherwise.

This is again a decision rule based on the likelihood ratio. On the right are the prior odds (as in Eq. 6) and a second term from the payoff matrix. When no-signal trials are prevalent (large $P(N)$), the likelihood ratio will have to be large to convince the observer to say “yes”. Similarly, when the extra value for being correct on a no-signal trial (the numerator on the right) is much bigger than the extra value for being correct on signal trials (the denominator), the likelihood ratio will have to be large to convince the observer to say “yes”.

In summary, the criterion the observer uses determines the observer’s bias for saying “yes”. There are two principal ways to affect bias: priors and payoffs. If the observer is in a situation in which, on most trials, the signal is present (e.g., you expect your friend to show up on that foggy night), and they are aware of this prior distribution ($P(S)$ is near one), then it makes sense to be easily swayed to respond “yes”—to set a low value of c_{opt}. In contrast, if signals are rare (few patients have this particular disease, very few pieces of baggage contain guns or bombs), then perhaps a high (conservative) criterion is appropriate. At the same time, decisions have consequences. Allowing a bomb onto a plane or sending a sick patient home without treatment can be disastrous, indicating a very high cost of a miss for baggage screeners or doctors, so a liberal criterion is appropriate. This is the difficult situation faced by airport baggage screeners and by radiologists—the payoff matrix implies use of a liberal criterion, but the priors suggest a conservative one. In fact, the evidence suggests that human observers are ill-equipped to select an optimal criterion when the priors are far from 50:50 (Wolfe et al., 2007).

Standardized model

The model outlined in Fig. 1 is general in the sense that there are parameters for both means (μ_S and μ_N) and the common standard deviation (σ). However, all of the derivations of the normative model above, and analysis of data below, are based only on the response rates (hit rate, false-alarm rate, etc.) and the likelihood ratio. None of these values will change with a change of variables for the decision variable involving a horizontal shift or a rescaling of the decision variable. One standard presentation of

signal detection theory, especially for yes-no tasks such as I've described, imposes a change of variables so that the decision axis is in units of *z*-score for the no-signal distribution. That is, we apply a change of variables $y = (x - \mu_N)/\sigma$. This change of variables leads to the following model (Fig. 3):

$$p(y|N) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y^2}{2}\right] \text{ and}$$
$$p(y|S) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-d')^2}{2}\right], \text{ where} \quad (11)$$
$$d' = \frac{\mu_S - \mu_N}{\sigma}.$$

Here, d' ("d prime") is a ratio between the strength of the signal and the standard deviation of the noise—the signal-to-noise ratio. This term appears in many engineering disciplines concerning signal processing, where signal strength is usually given in units of power and noise in units of variance. That is, the signal-to-noise ratio $SNR = d'^2$.

INSERT FIGURE 3 ABOUT HERE

All three decision rules (ML, MAP and maximum expected gain) result in a criterion value of the likelihood ratio (typically denoted as β). For the standardized model, the relationship between the decision variable, y, and β is particularly simple, showing that a criterion β_{opt} on likelihood ratio corresponds to a criterion c_{opt} on the decision variable:

$$\beta_{\text{opt}} = \frac{p(c_{\text{opt}}|S)}{p(c_{\text{opt}}|N)} = \frac{\exp\left[-\frac{(c_{\text{opt}}-d')^2}{2}\right]}{\exp\left[-\frac{c_{\text{opt}}^2}{2}\right]} = \exp\left[c_{\text{opt}}d' - \frac{d'^2}{2}\right] \quad (12)$$
$$\log\beta_{\text{opt}} = c_{\text{opt}}d' - \frac{d'^2}{2}$$
$$c_{\text{opt}} = \frac{d'}{2} + \frac{\log\beta_{\text{opt}}}{d'}.$$

Thus, the optimal criterion on the decision variable is determined by, and monotonically increases with, the optimal criterion on the likelihood ratio. If one requires a large likelihood ratio to say "yes", that will lead to a large, conservative criterion on the decision variable. For the maximum expected gain model (Eq. 10), this implies that

$$
\begin{aligned}
c_{\text{opt}} &= \frac{d'}{2} + \frac{\log \beta_{\text{opt}}}{d'} \\
&= \frac{d'}{2} + \frac{1}{d'}\left[\log \frac{P(N)}{P(S)} + \log \frac{V(N, \text{“no”}) - V(N, \text{“yes”})}{V(S, \text{“yes”}) - V(S, \text{“no”})}\right].
\end{aligned}
\tag{13}
$$

Thus, the effects of priors and payoffs on the optimal criterion are additive.

With the standardized representation of signal detection theory (Fig. 3), the main benefit of signal detection is made clear—the distinction between discriminability and bias. The stronger the signal, the more accurately one can perform the task. In this representation, signal strength or discriminability corresponds to the separation between the two distributions, d'. On the other hand, given a fixed amount of information (fixed d'), the bias toward responding “yes” or “no” is reflected in the position of the decision criterion c and may be defined for the normative model as $c_{\text{opt}} - c_{\text{ML}}$, where $c_{\text{ML}} = d'/2$ is the neutral or ML criterion. The optimal criterion, c_{opt}, is determined by the priors and payoffs. If the payoffs are symmetric (equal benefits for hits and correct rejects, equal penalties for false alarms and misses), then the optimal behavior is MAP. If, in addition, there are equal priors ($P(S) = P(N) = 0.5$), the optimal behavior is ML.

Data Analysis: The Experimenter’s Perspective

Parameter estimation from data

We now examine signal detection theory from the perspective of the experimenter. From this perspective, the goal is to use behavioral data to infer something about how an observer’s decisions were made. In the case of signal detection theory, an experimenter might like to infer the model parameters from data. The full model includes details of the stimulus encoding (μ_S, μ_N and σ), the prior ($P(S)$), and the payoff matrix (the four values of V). We would like to estimate these parameters because there is no guarantee that observers use accurate estimates of these parameters in formulating a decision—humans may not behave in accordance with the normative model.

However, given that the general encoding model is equivalent in all of its predictions to the standardized model, the only parameters that can be estimated are d' (i.e., $(\mu_S - \mu_N)/\sigma$) and the criterion (c). Experimentally, we know that the prior and payoff matrix can affect this criterion, but for a given, fixed set of conditions, all we can estimate are d' and c. There are two degrees of freedom in the data we collect (hit and false-alarm rate) and two degrees of freedom in the standardized model (d' and c), enabling a direct mapping from a pair of hit and false-alarm rates to estimates of d' and c.

Looking again at the standardized model in Fig. 3, we see that the value of c is the distance of the criterion to the right of the mean of the noise distribution, and d' is the sum of that distance plus the distance from the criterion to the mean of the signal distribution. In other words:

$$\begin{aligned} d' &= z[P(\text{Hit})] + z[P(\text{Correct reject})] \\ &= z[P(\text{Hit})] - z[P(\text{False alarm})] \text{ and} \\ c &= z[P(\text{Correct reject})], \end{aligned} \tag{14}$$

where $z(P) = \Phi^{-1}(P)$ and $\Phi(z) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{z} \exp(-x^2/2)dx$ is the cumulative standard normal distribution. $\Phi(z)$ is the area to the left of z under the standard bell curve, that is, the probability of drawing a random value that is z or less. Thus, $z(P)$ is the *z*-score corresponding to a particular probability—the position on the *x*-axis corresponding to a particular left-hand-tail probability of the standard normal distribution.

As a data-analysis method, the only change is to replace the theoretical probabilities in Eq. 14 with their empirical estimates. $P(\text{Hit})$ is replaced by the proportion of signal trials in which the observer's response was "yes" and $P(\text{Correct reject})$ is replaced by the proportion of no-signal trials in which the observer responded "no", etc.

As an example, suppose that there were 60 signal trials of which 47 were hits and 60 no-signal trials of which 21 were false alarms. We find that

$$P(\text{Hit}) = \frac{47}{60} = 0.783.$$

$$P(\text{False alarm}) = \frac{21}{60} = 0.35.$$

$$d' = z[0.783] - z[0.35] = 0.784 - (-0.385) = 1.169.$$

$$c = z[P(\text{Correct reject})] = z[1 - 0.35] = 0.385.$$

There is, however, one possible complication. The z function results in infinite values if supplied a probability of zero or one. That can result if the data contain no false alarms or 100% hits. There are two standard procedures for these cases. A fairly typical procedure is to take the problematic row in the results table and add 1/2 of a trial to both columns. So, if you had zero false alarms and 20 correct rejects (a correct-reject rate of 100%), you would process the data as if you had 1/2 trial worth of false alarm and 20.5 correct rejects, for a correct-reject rate of $\dfrac{20.5}{21} = 97.6\,\%$. However, Hautus (1995) suggests that a less biased procedure is to *always* add 1/2 trial to all 4 elements of the results table. That would modify the results of our example above as follows:

$$P(\text{Hit}) = \frac{47.5}{61} = 0.779.$$

$$P(\text{False alarm}) = \frac{21.5}{61} = 0.353.$$

$$d' = z[0.779] - z[0.353] = 0.768 - (-0.379) = 1.147.$$

$$c = z[P(\text{Correct reject})] = z[1 - 0.353] = 0.379.$$

The psychometric function, varying signal strength

So far, we have described a particularly simple experiment in which there are *only* two possible stimuli (signal and noise). Suppose that you are interested in human performance as a function of the "strength" of the signal. For example, suppose that you hypothesize that human perception is tuned to images of faces. You measure detection performance for faces vs. outdoor scenes as a function of stimulus contrast in the presence of a fixed noisy background image. That is, for each target you estimate detectability (d') as a function of stimulus contrast. You choose 6 contrast levels for the target, running the face trials in a separate session from the scene trials. In each session, on half of the trials, there is no target (and the correct answer is "no"), and the other half of the trials are split equally between the 6 contrast levels (and the correct response is "yes"). These trials are run in random order. If feedback is supplied after each trial, it indicates that the correct answer is "yes" and "no" equally often. In this design, there are hit rates $P(\text{"yes"} \,|\, S, \text{level} = i)$ for every signal level i, but there is a single, shared false-alarm rate from the no-signal trials.

This shared false-alarm rate does not complicate the analysis. One can use the d' and c formulas given above (Eq. 14) separately for every stimulus level. Each estimate of d' goes hand-in-hand with a corresponding estimate of c. But, the calculation of the criterion c only uses the false-alarm rate, that in turn is based only on the trials with no signal. That is, the single, shared false-alarm rate is used for the computation of d' and c for all signal levels. As a result, the criterion c is identical for all signal levels. This is important—how would the observer be able to use a different value of the criterion for each signal level, when those levels are randomly intermixed and not perfectly identifiable by the observer (who only has the noisy measurement x available)? Fortunately, this multi-intensity analysis is consistent with the use of a single criterion throughout.

The ROC curve

Signal detection theory makes a specific prediction of the consequences of changing one's criterion for performance in a signal-detection task. Fig. 4 illustrates these predictions. The graphs shown here are referred to as "Receiver Operating Characteristics" or ROC curves. They illustrate the tradeoff between correct "yes" answers (hits) and incorrect "yes" answers (false alarms) as the criterion varies. Each curve corresponds to a particular value of signal discriminability (d'). For any given value of d', a liberal (i.e., low) criterion leads to a large hit rate but also a large false-alarm rate (toward the upper-right of the plot), whereas a conservative (i.e., high) criterion reduces both hit and false-alarm rates; it shifts performance toward the lower-left portion of the plot. As d' increases, the curves push up to the upper-left corner of the plot (i.e., closer to perfect performance—100% hits, 0% false alarms). The negative diagonal on this plot corresponds to $P(\text{hit}) = 1 - P(\text{false alarm}) = P(\text{correct reject})$. Looking at Fig. 1C, equal hit and correct-reject rates corresponds to using the criterion where the curves cross, that is, the neutral (unbiased) or ML criterion.

INSERT FIGURE 4 ABOUT HERE

Signal detection theory nicely segregates two aspects of a binary, yes-no decision—discriminability (how good you are at discriminating signal from noise) and bias (do you tend to say "yes" or "no" more often). In the ROC plot, discriminability determines which curve performance will lie on and bias determines your operating point along that curve. The curves indicate theoretical, expected performance ($P(\text{“yes”} \mid S, c)$ and $P(\text{“yes”} \mid N, c)$). With a finite number of trials in a dataset, the actual hit and false-alarm rates will deviate from these values and will reflect binomial (coin-flip) variability.

Thus, an alternative approach to estimating d' is to do so via the ROC curve. The first step is to collect, for a given stimulus, a set of hit rate/false-alarm rate pairs, each corresponding to a different value of the criterion. The experimenter can induce the observer to adopt different criteria (typically in separate blocks of trials) by simply asking them to do so or by varying the priors (the proportion of signal trials) or payoffs. Alternatively, pairs of hit and false-alarm rate can be collected during a single block of trials by expanding the number of response alternatives.

For example, one can use a set of 5 possible confidence ratings as response alternatives. For our face-detection experiment, those possible responses would be: 1 = I'm sure a face is not present; 2 = I think a face is not present but with low confidence; 3 = I have no idea whether a face is present or not; etc. Then, the data can be analyzed by the *experimenter* adopting 4 different criteria: (1) treat a response of 1 as "no" and responses 2-5 as "yes"; (2) treat a response of 1 or 2 as "no" and responses 3-5 as "yes"; (3) treat responses 1-3 as "no" and responses 4-5 as "yes"; and (4) treat responses 1-4 as "no" and only response 5 as "yes". This will yield a set of four points for an ROC plot, and the experimenter can then choose the ROC curve that best fits the data (Figure 5A).

INSERT FIGURE 5 ABOUT HERE

To better understand how to think about the "best" ROC curve to fit to a set of hit rate/false-alarm rate pairs, consider another way of plotting the ROC itself (Fig. 5B). Here, hit rate is again plotted as a function of false-alarm rate. But, instead of using linear probability axes, as in Fig. 5A, I convert each probability to its corresponding *z*-score, $z(P)$. Standard signal detection theory assumes that the signal and noise distributions share a common standard deviation. Thus, if you move the criterion one SD rightward relative to the noise distribution, you have also moved that criterion one SD rightward relative to the signal distribution. This implies that the ROC "curve" on these new axes is now a straight line with slope one (Fig. 5B).

Thus, to estimate d' from a set of pairs of hit and false-alarm rates, one can plot the data using the *z*-score axes and find the best-fitting line of slope one. However, the data points in this plot have *x*-values (false-alarm rate) and *y*-values (hit rate) that are *both* dependent variables, so standard linear regression is inappropriate here. One solution is to use a maximum-likelihood method by determining the set of parameters (in our example, 4 criteria and one value of d') so that the likelihood of the data ($P(\text{confidence ratings} \mid d', c_1, \ldots, c_4)$) is maximized (Dorfman & Alf, 1969).

Deviations of human behavior from the normative model

As I mentioned, one way to collect data for multiple criteria is to run blocks of trials that vary in either priors, $P(S)$, or in payoffs ($V(S$, "yes"), etc.). To do this you will have to inform your observer of these values or allow them enough trials to experience the current priors and payoffs. For a known value of d', the normative theory indicates the optimal criterion (Eq. 13). There is, of course, no guarantee that humans perform in an optimal manner. In fact, the typical finding (termed "conservatism") is that, when priors or payoffs are made asymmetric, the criterion adopted by human observers moves in the correct direction away from the neutral ML criterion but is not moved as far as the normative theory predicts (Ackermann & Landy, 2015; Healy & Kubovy, 1978, 1981; Ulehla, 1966). One explanation of this behavior is that human behavior is, in fact, optimal, but that observed conservatism is due to a violation of the assumption of normally distributed noise and that, instead, the noise comes from a different form of distribution such as the Laplace (Maloney & Thomas, 1991). Another explanation is that humans typically use distorted values of probabilities (e.g., of the prior probability of a signal $P(S)$), leading to conservative criterion placement (Zhang & Maloney, 2012). Other violations of normative theory include data that contradict the prediction that the effects of changed priors and changed payoffs on criterion placement are additive as predicted by Eq. 13 (Locke et al., 2020) and that the criterion is fixed and stable across a block of trials (Norton et al., 2019; Norton et al., 2017).

Forced-Choice Tasks

Signal detection theory can be applied to other tasks as well, such as two-alternative forced choice (2AFC). In this task, there are two stimuli presented on each trial (e.g., a noise pattern vs. a noise pattern plus a low-contrast image of a face). The observer's task is to identify which stimulus had the signal (i.e., the face). The two stimuli could be presented in different spatial locations (e.g., left and right of visual fixation) or in different temporal intervals, sequentially. Typically, psychophysicists prefer this task, compared to the yes-no task, because this task is often described as "bias-free". That is, the participant can't be biased to say "yes", because that's not one of the response options, and there is *always* a signal presented. However, they can be biased to say "2nd interval" and, in fact, data indicate that participants often do have an interval bias (Yeshurun et al., 2008). The nomenclature for these different experiments varies, but the theory is the same—the observer has two noisy measurements (x_1 and x_2, e.g., from the first and second interval) and must decide which contained the signal. The rational decision procedure (assuming that the signal is equally likely to appear in either location or temporal interval) is to select the interval that led to the larger measurement.

On each trial, the observer has a pair of measurements (x_1, x_2)—the model for this experiment comprises a two-dimensional space of potential measurements (Fig. 6). A trial's measurement pair is now distributed as a bivariate Gaussian, and we assume the measurements in the two intervals are independent and both have standard deviation equal to one (hence the distributions are shown as a set of concentric circles in Fig. 6). We again adopt the standardized model, so that a no-signal measurement is,

on average, zero. For the typical, single-interval, yes-no task, where we only have one measurement, x_1, and the observer must say "yes" or "no", we denote the discriminability as d'_{YN}. For the forced-choice task, the mean pair of measurements for trials in which the signal is in interval one is $(d'_{\mathrm{YN}},0)$ and for interval two it is $(0,d'_{\mathrm{YN}})$.

INSERT FIGURE 6 ABOUT HERE

We can treat the forced-choice task as a signal-detection task by, for example, treating a trial in which the signal is in interval 1 as a "no-signal trial"; when the signal appears in interval 2, we treat this as a "signal" trial (with the corresponding definitions of hit, false alarm, etc.). Performance in the forced-choice task is governed by the separation between these two bivariate Gaussian distributions. We denote performance (i.e., discriminability) in the forced-choice task as d'_{FC}. From the geometry of Fig. 6, it is clear that $d'_{\mathrm{FC}} = \sqrt{2}d'_{\mathrm{YN}}$. Note that typical 2AFC behavioral data do *not* satisfy the assumptions of ideal behavior as just described. Data often indicate that people do not place a symmetric criterion between interval 1 and 2, nor do results indicate equal detectability of the stimulus in interval 1 and the stimulus in interval 2 (Yeshurun et al., 2008).

When the signal is in interval 2, the ideal observer will be correct when $x_2 > x_1$; similarly, when the signal is in interval 1, the observer will be correct when $x_2 < x_1$. Thus, the probability of being correct is the probability of the set of all pairs of measurements that satisfy either inequality (by symmetry, they are identical), so that:

$$P_{\mathrm{FC}} = P(x_2 > x_1 \,|\, s_2 = S) = \int_{-\infty}^{\infty} P(x_2 \,|\, s_2 = S) \int_{-\infty}^{x_2} p(x_1 \,|\, s_1 = N) dx_1 dx_2. \quad (15)$$

Here's a useful, important and fairly unobvious fact: For any given value of d'_{YN}, the value of P_{FC} is equal to the area under the corresponding ROC curve for the yes-no detection task (Fig. 7A)! This is clearly true at the extremes. When $d'_{\mathrm{YN}} = 0$, the observer has no information about which stimulus is which and is forced to guess, so that $P_{\mathrm{FC}} = 1/2$. The corresponding "area under the ROC" (AUROC or AUC) is the area under the main diagonal (Fig. 4)—$1/2$. Similarly, when d'_{YN} is effectively infinite, 2AFC performance becomes perfect and the area under the ROC is the entire area of the ROC plot—one.

INSERT FIGURE 7 ABOUT HERE

Between these extremes, the fact that P_{FC} is equal to the area under the corresponding ROC curve is not obvious. Consider the rectangle outlined in Fig. 7A. AUC is computed by summing the areas of such rectangles. The rectangle's height is an infinitesimal portion of the *y*-axis, $dP(\mathrm{Hit})$. The width of the rectangle is one minus the *x* value at that position on the ROC curve, $1 - P(\mathrm{False\ alarm}) = P(\mathrm{Correct\ reject})$. The full area is the sum over all such rectangles. Those rectangles can be parameterized by the criterion c that results in the point on the ROC that intersects the

rectangle. Fig. 7B shows that a differential amount of the *y*-axis in Fig. 7A corresponds to a differential amount of area under the signal distribution (S). That is, $dP(\text{Hit} \mid x = c) = p(x = c \mid S)dc$. Combining these,

$$\begin{aligned}
\text{AUROC} &= \int_0^1 P(\text{Correct reject} \mid \text{criterion } c)\, dP(\text{Hit} \mid \text{criterion } c) \\
&= \int_{-\infty}^{\infty} P(x < c \mid N)\, p(\text{measurement is } c \mid S)\, dc \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{c} P(x \mid N)\, p(\text{measurement is } c \mid S)\, dx\, dc \\
&= \int_{-\infty}^{\infty} p(\text{measurement is } c \mid S) \int_{-\infty}^{c} P(x \mid N)\, dx\, dc \\
&= P_{\text{FC}}\,.
\end{aligned} \tag{16}$$

Thus, discriminability, instantiated by the value of d', inextricably ties predictions of performance in yes-no and 2AFC tasks.

Alternative Models

The literature includes many alternatives to standard signal detection theory as a model of detection and discrimination performance. One class of such models is the set of so-called threshold models. There are several types of threshold models, and I'll illustrate one here to give an idea of how this class of models works. I assume that the observer does not have access to the noisy measurement directly, but rather, these noisy measurements result in one of two internal states, "detect" and "no-detect" (possibly through a version of signal detection theory to which the observer does not have conscious access). On *S* trials, the detect state is entered with probability P_{Hit} and otherwise the no-detect state results. On *N* trials, the detect state is entered with probability P_{FA}. Thus, if the observer merely reports the current state, saying "yes" in the detect state and "no" otherwise, the ROC will have one point at coordinates $(P_{\text{FA}}, P_{\text{Hit}})$ (Fig. 8, filled circle).

To generate a full ROC contour, one allows for randomness in the response. Suppose the observer doesn't trust their internal state. For example, when they are in the no-detect state, they occasionally decide to say "yes" despite being in the "no-detect" state. Depending on how often they make this decision, performance will lie somewhere along the upper line in Fig. 8. Similarly, if the observer instead doesn't always trust the internal "detect" state and occasionally says "no" despite being in that state, their behavior will lie along on the lower line in Fig. 8. The result of this set of behaviors (mistrust of either the detect or the no-detect state) is an ROC contour consisting of two straight lines, rather than the smooth curve (Fig. 4) resulting from standard SDT.

INSERT FIGURE 8 ABOUT HERE

A generalization of standard signal detection theory drops the assumption of equal variances for the *N* and *S* measurement distributions. For most intensive stimulus parameters (e.g., luminance, size, weight, speed, loudness, etc.) the just-noticeable difference (the difference in intensity between a base stimulus and an increment leading to a criterion discrimination performance such as $d' = 1$) is approximately proportional to the intensity of the base stimulus, which is known as Weber's Law. There are many combinations of stimulus encoding and noise that are consistent with Weber's Law (Zhou et al., 2024). One model consistent with Weber's Law drops the equal-variance assumption and instead suggests that noise standard deviation grows approximately proportional to stimulus intensity.

INSERT FIGURE 9 ABOUT HERE

Fig. 9A illustrates an unequal-variance context in which the stronger stimulus has higher variance, consistent with this model of Weber's Law, such as might result from, for example, discrimination of image contrast. One can still posit that observers perform this task by setting a single criterion c and responding "yes" when the measurement exceeds this criterion. For the example in Fig. 9A, this leads to the asymmetric ROC curve shown in Fig. 9B. In Fig. 9A, the S distribution has a standard deviation three times as large as the N distribution. Thus, if we move c one standard deviation to the right relative to the N distribution, we will have moved that criterion only one-third of a standard deviation to the right relative to the S distribution. As a result, if I plot the ROC curve with the axes scaled as *z*-scores as I did in Fig. 5B, each shift to the right by one will lead to a shift upward of 1/3 (in *z*-score for the false-alarm and hit rates corresponding to the changed criterion). Thus, I will again have an ROC that is a straight line, but the slope is no longer one but, instead, is equal to the ratio of the two standard deviations (here: $1/3$, Fig. 9C).

We pointed out above (*Observer's perspective and the SDT model*) that the optimal decision rule, taking payoffs and priors into account, is to impose a threshold on the likelihood ratio. That derivation only used the two likelihoods and never made reference to the signal detection theory assumptions of equal variance or Gaussian measurement distributions. An ideal decision-maker bases decisions *only* on likelihood ratio, in other words, the likelihood ratio is a sufficient statistic for this decision. Inspecting Eq. 10, if we are in a situation with equal priors ($P(S) = P(S) = 0.5$) and equal payoffs (the value of being correct on noise trials is the same as on signal trials), then $\beta_{\text{opt}} = 1$. That is, the criterion should be placed where the signal and noise measurement distribution curves cross. For unequal signal and noise variances, the curves cross in *two* places (Fig. 9A) and thus the optimal decision-maker doesn't place a single criterion, saying "yes" when that criterion is exceeded. Rather, the optimal strategy is to say yes when the measurement exceeds a high criterion c_h or falls below a second, low criterion c_l, because very low values of the measurement are more likely to occur on signal than on noise trials.

INSERT FIGURE 10 ABOUT HERE

Having noticed that Eq. 10 can be applied to any pairs of distributions, we can also drop the assumption that the two distributions are Gaussian. Fig. 10 shows N and S distributions that are Poisson-distributed, differing in the expected number of counts. This is a reasonable model for a decision based on the number of photons caught by a collection of rod photoreceptors in the retina or based on the number of action potentials from a single neuron. For equal priors and symmetric payoffs, the optimal decision is maximum likelihood: pick the stimulus based on the higher curve corresponding to the current measurement (here, resulting in a criterion between one and two counts).

Applications to Neural Data

The tools I have outlined may also be applied in situations in which we have empirically measured distributions and choose not to make an assumption of a particular distributional form (Gaussian, Poisson, etc.). A particularly well-known example is the application to neural responses to visual motion in cortical area MT, the "middle temporal area" (Britten et al., 1992; Newsome et al., 1989; Salzman et al., 1990; Salzman et al., 1992); the same ideas can be applied to any discrimination based on neural responses.

Newsome and colleagues recorded from single neurons in area MT of the macaque monkey while the monkey viewed a random-dot motion display in which a subset of the dots moved in either the preferred direction of the neuron or in the opposite (anti-preferred or null) direction. The other dots moved in random directions. Across trials, they varied the stimulus *coherence*—the fraction of dots that moved together. They included a zero-coherence condition (purely random motion). Each stimulus was presented many times, resulting in a histogram of the number of action potentials from the neuron summed over the stimulus duration—one histogram for each coherence level and motion direction. At the same time, the monkey was awake and performing a discrimination task on motion direction, effectively deciding whether the stimulus moved in the currently recorded neuron's preferred or null direction. The tools I've discussed allowed Newsome and colleagues to determine what information a single neuron had concerning the stimulus being displayed (the *neurometric function*) as well as about the decision the monkey was about to make (*choice probability*).

A psychometric function for a task such as this is a measurement of behavioral performance as a function of a stimulus variable. Here, that could be, for example, d' for discriminating leftward vs. rightward motion as a function of motion coherence in a forced-choice task. The notion of a neurometric function is to generalize this concept from behavior to the information contained in neural responses. Here, the noisy measurement is the spike count from the neuron. The "behavior" is generated by an idealized decision-maker that bases its choice on the neural spike counts in response to each of the stimuli to be discriminated.

Consider the two histograms in Fig. 11A. The histogram on the left represents responses "recorded" from a neuron in response to a 10% coherence stimulus moving in the null direction; the histogram on the right shows the responses at this coherence

for stimuli moving in the preferred direction. We could pick an arbitrary criterion (such as where the curves cross) and compute a hit rate (the fraction of preferred-direction responses that exceed the criterion) and false-alarm rate (the fraction of null-direction responses that exceed the criterion) and then compute d' from these two rates. However, these histograms are empirical distributions; they are a noisy representation of the true, underlying distributions that would result from an infinite number of trials.

It makes more sense to use all the information we have in these histograms to compute a measure of the ability of this neuron to discriminate these two stimuli. What Newsome and colleagues proposed is to use the same trick as described above for using confidence ratings. Place a "criterion" at 1 action potential, and compute hit and false-alarm rates based on that criterion. Repeat with the criterion equal to 2, 3, 4, … spikes. When you are done, you have produced a piecewise-linear ROC curve (Fig. 11B). Discrimination performance can be summarized using the area under the ROC. Recall that the area under the ROC is equivalent to performance in a 2AFC task. Here, that task is that I give you a random sample from the left-hand histogram and a random sample from the right-hand histogram. You decide that the sample drawn from the preferred direction's distribution is the sample with more action potentials. The area under the ROC is the predicted proportion of correct decisions in that 2AFC task. Finally, this exercise can be repeated with the pairs of neural-response histograms corresponding to each coherence level used in the experiment. The result is a neurometric function, with predicted direction-discrimination performance as a function of stimulus coherence (Fig. 11C), yielding a description of the information content in a single neuron's firing rate for this task.

INSERT FIGURE 11 ABOUT HERE

The second question that Newsome and colleagues asked was "How informative is this neuron about the behavior of the animal?" The approach to this question was quite similar. Obviously, when stimulus coherence was high, response variability was low (the animal was correct most of the time). The highest response variability was when the stimulus had no information at all (zero coherence). Fig. 11D again shows two histograms of neural responses, but this time both are (simulated) responses to completely random, zero-coherence motion stimuli. However, this time the histograms are conditioned on the monkey's response.

The left-hand histogram corresponds to when the animal decided the stimulus moved in the neuron's null direction, and the right-hand histogram shows neural responses when the monkey decided the stimulus moved in the neuron's preferred direction. In both cases, the stimuli themselves were completely random and uninformative. Given the two histograms, we can compute the areas to the right of each possible criterion, yielding hit and false-alarm rates for the task of deciding what behavioral response the monkey made given knowledge only of this neuron's response. From these rates, we can construct a ROC and compute the area under the ROC, that they called the choice probability for this neuron. Choice probability is a measure of how useful this neuron is in discriminating what behavioral choice the monkey subsequently made. This approach shows the usefulness of the ROC and, in particular, of the area

under the ROC as a nonparametric analysis method for summarizing discrimination performance.

Extensions to Metacognition

In this chapter, I have discussed signal detection theory solely with regard to perceptual decisions (Is the signal there or not? Are the dots moving to the right or left?). However, having completed a task, humans also typically have a feeling of how successfully they carried out the task. In the case of binary decisions (the “first-order” task), an experimenter can ask the observer something about their estimate of the probability that decision was correct (the “second-order” task). Reasoning about one’s own thoughts and actions is called metacognition.

There are many sources of information that one can use to inform a metacognitive judgment. For a random-dot display, there are stimulus cues that are correlated with the quality of the stimulus, such as the perceived randomness or inconsistency of dot motion directions, that could inform a judgment of confidence in the first-order dot-direction decision. There are also aspects of one’s own behavior that could inform a decision, such as having low confidence if one’s own reaction time for the first-order task was long or basing confidence on the previous rate of success in the task. In addition, one can use the elements of signal detection theory itself to form a second-order judgment.

INSERT FIGURE 12 ABOUT HERE

Consider again the signal-detection experiment in which the observer is asked to discriminate stimulus N (a noise pattern) from S (a noise pattern plus a low-contrast face) and subsequently indicate whether that response was made with low or high confidence (Fig. 12A). The theoretical setup is that of standard signal detection theory with two unit-variance measurement distributions separated by d' (first-order discriminability) and equal priors and payoffs.The observer’s response bias is represented by the first-order criterion c_1 (here, the neutral criterion is indicated). If a measurement lies quite close to this first-order criterion, the likelihood ratio and the posterior odds will be close to one, so it makes sense to have low confidence; if the measurement is far from the criterion, this justifies increased confidence. Thus, a simple model of the metacognitive judgment is that the observer adopts second-order criteria (one for each possible first-order response: $c_{2|r=N}$ and $c_{2|r=S}$) and responds “high confidence” if the measurement is farther from the first-order criterion than the corresponding second-order criterion.

Consider the case when the first-order response is “S”. Thus, the measurement lies to the right of c_1. For the second-order task (the confidence response), there are two possible stimuli that could have appeared (corresponding to first-order hits and false alarms) and two possible confidence responses. When the stimulus was, in fact, S the response “S” was correct. If the confidence response was “high”, we might say that high confidence was justified (because they were correct) and call that a second-order hit. When the stimulus was N, then high confidence was unjustified, and we can call that a

second-order false alarm. However, if confidence was low, that is a 2nd-order correct reject.

The first-order criterion c_1 is estimated using the first-order false-alarm rate. Given the estimated values of c_1 and d', and any possible value of the second-order criterion $c_{2|r=S}$, the predicted probabilities of 2nd-order hits and false alarms can be calculated. These are the probability that, for each stimulus, the measurement exceeds $c_{2|r=S}$ given that it already exceeds c_1. Varying $c_{2|r=S}$ over the entire range yields a 2nd-order ROC curve (Fig. 12B). The measured second-order hit and false-alarm rates are not likely to land on this theoretical ROC curve. Maniscalco and Lau (2012) proposed a measure of metacognitive sensitivity (how sensitive you are to your own stimulus information and the quality of your first-order judgment) by computing what they called meta-d' (for a comprehensive summary of metacognition metrics, see Rahnev, 2025). Any given pair of d' and c_1 values results in a second-order ROC curve.

Maniscalco and Lau defined meta-d' as the value of d' that, if paired with the criterion "corresponding to" c_1, yields the observed pair of second-order hit and false-alarm rates. One has, of course, to determine what you mean by the corresponding first-order criterion, since you are now treating the problem as if the two distributions are a different distance apart. The sensible solution, proposed by Maniscalco and Lau, is to use the criterion "c_1" along with the new value of d' so that the corresponding value of the likelihood ratio β for that criterion is the same as the value β derived from the first-order responses.

The literature on metacognition has many other definitions of metacognitive sensitivity. The development of the meta-d' metric was, among other things, an attempt to develop a metric that estimates metacognitive sensitivity independent of (metacognitive) bias and first-order discriminability. However, meta-d' tracks first-order sensitivity. That is, if d' is high, meta-d' is likely to be high as well. Thus, to derive a metric for the quality of metacognition itself, Maniscalco and Lau proposed that researchers report the M-ratio (meta-d'/d')—the fraction of the information in the first-order judgment that is effectively used in the second-order confidence judgment.

Conclusion

In the 70 years (as of this writing) since signal detection theory was introduced, it has become the standard model and data-analysis technique for detection and discrimination experiments in a wide variety of research areas, often far from the sensory experiments in which it was first described. It allows the researcher to separately estimate observer sensitivity (d') and response bias (c or β). Through careful estimation of the ROC curve, it also allows the experimenter to test the underlying assumptions (continuous decision variable, equal variances, Gaussian distributions). The method can be used with other distributions, including discrete distributions, such as are found in single-unit neural measurements. It is a general model of behavior and the information in neural coding, with wide applicability.

References

Ackermann, J. F., & Landy, M. S. (2015). Suboptimal decision criteria are predicted by subjectively weighted probabilities and rewards. *Atten Percept Psychophys*, *77*(2), 638-658.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience*, *12*(12), 4745-4765.

Dorfman, D. D., & Alf, J., E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating-method data. *Journal of Mathematical Psychology*, *6*, 487-496.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, *96*(2), 267-314.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Krieger.

Hautus, M. J. (1995). Corrections for extreme proportions and the biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*, 46-51.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Mem Cognit*, *6*(5), 544-553.

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 344-354.

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Atten Percept Psychophys*, *82*(6), 3158-3175.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: A User's Guide*. Cambridge.

Maloney, L. T., & Thomas, E. A. C. (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*, *35*, 443-470.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn*, *21*(1), 422-430.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*(6237), 52-54.

Norton, E. H., Acerbi, L., Ma, W. J., & Landy, M. S. (2019). Human online adaptation to changes in prior probability. *PLoS Comput Biol*, *15*(7), e1006681.

Norton, E. H., Fleming, S. M., Daw, N. D., & Landy, M. S. (2017). Suboptimal Criterion Learning in Static and Dynamic Environments. *PLoS Comput Biol*, *13*(1), e1005304.

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions IRE Professional Group on Information Theory*, *PGIT-4*, 171-212.

Rahnev, D. (2025). A comprehensive assessment of current methods for measuring metacognition. *Nature Communications*, *16:701*.

Salzman, C. D., Britten, K. H., & Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, *346*(6280), 174-177.

Salzman, C. D., Murasugi, C. M., Britten, K. H., & Newsome, W. T. (1992). Microstimulation in visual area MT: effects on direction discrimination performance. *Journal of Neuroscience*, *12*(6), 2331-2355.

Ulehla, Z. J. (1966). Optimality of perceptual decision criteria. *J Exp Psychol*, *71*(4), 564-569.

Van Meter, D., & Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Transactions IRE Professional Group on Information Theory*, *PGIT-4*, 119-145.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford University Press.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen*, *136*(4), 623-638.

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Res*, *48*(17), 1837-1851.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front Neurosci*, *6*, 1.

Zhou, J., Duong, L. R., & Simoncelli, E. P. (2024). A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proceedings of the National Academy of Sciences USA*, *121*(25), e2312293121.

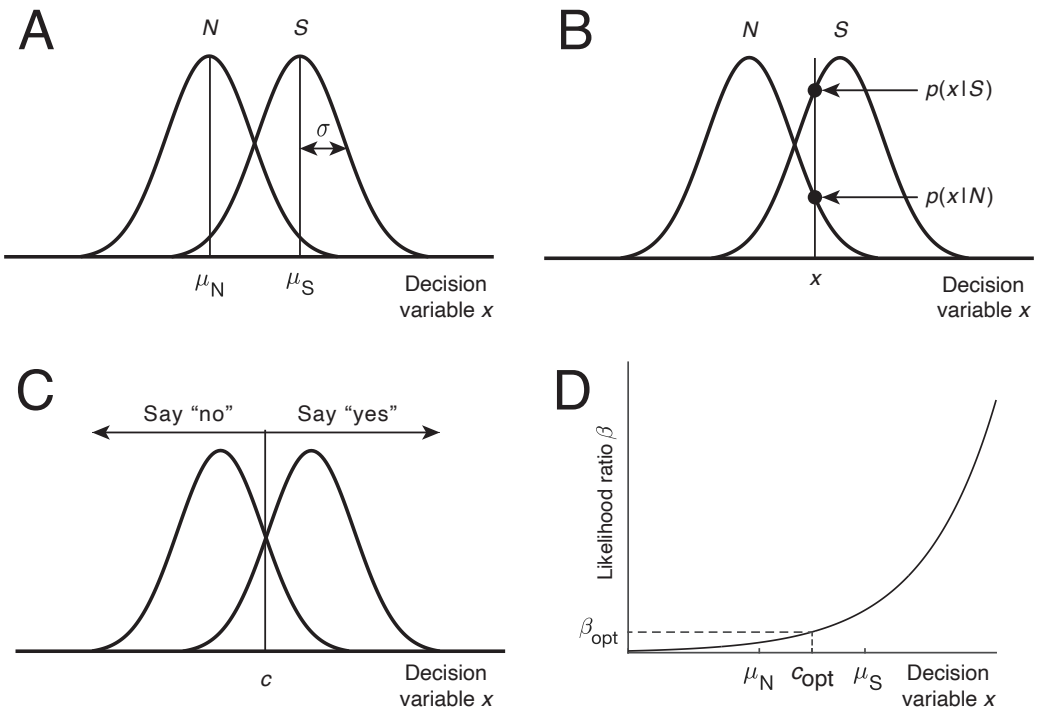

Figure 1. Standard signal detection theory. (A) The probability distribution for the evidence x on no-signal (*N*) and signal trials (*S*) with means μ_N and μ_S and common standard deviation σ. (B) The likelihood values $p(x \mid N)$ and $p(x \mid S)$ corresponding to a measured evidence value x. (C) The maximum-likelihood observer sets a criterion c where the curves cross and says "yes" when the decision variable exceeds that criterion. (D) The likelihood ratio ($p(x \mid S)/p(x \mid N)$) increases monotonically with x. Any criterion on likelihood ratio corresponds to a criterion on the decision variable x. The optimal criterion, for equal priors and payoffs, c_{opt} (where the curves cross in panel C) corresponds to a criterion on likelihood ratio, $\beta_{\text{opt}} = 1$.

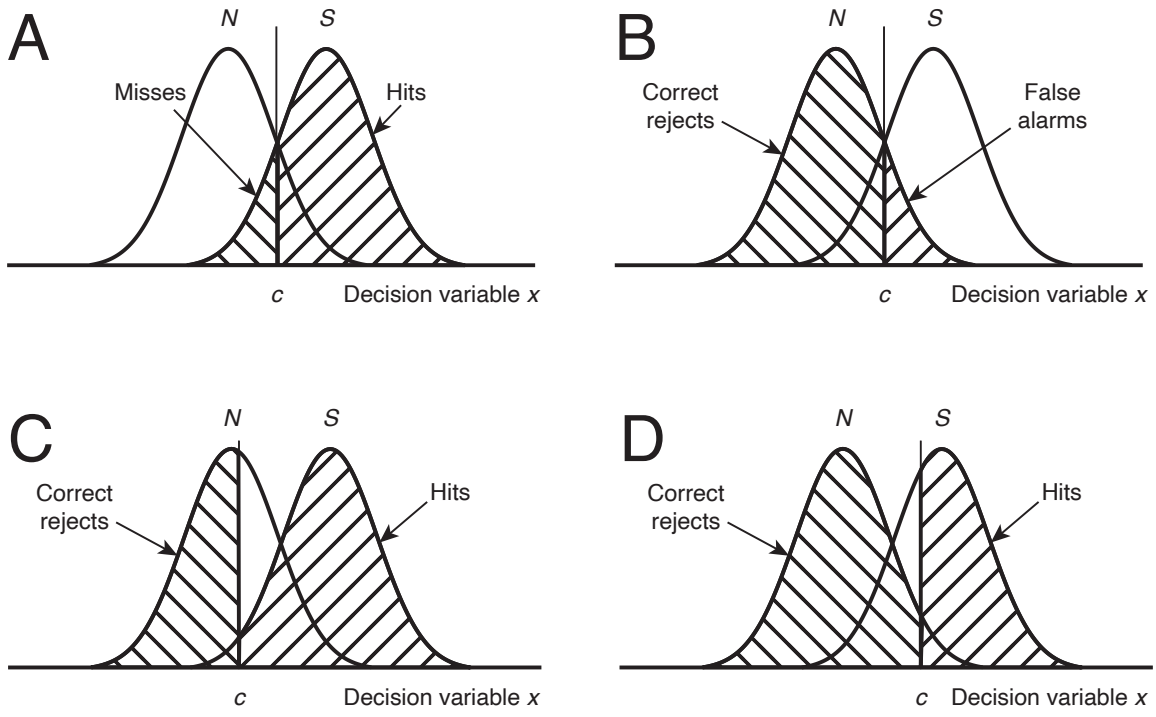

Figure 2. Criterion and decision outcomes. (A) When a signal was presented, these areas represent the probabilities of a hit and a miss. (B) When no signal was presented, these areas represent the probabilities of a false alarm and a correct reject. (C) The criterion c indexes the observer's bias to say "yes". Low values lead to a liberal bias: a high hit rate and few correct rejects. (D) Moving the criterion rightward leads to a conservative bias: a reduced hit rate, but increased correct rejects.

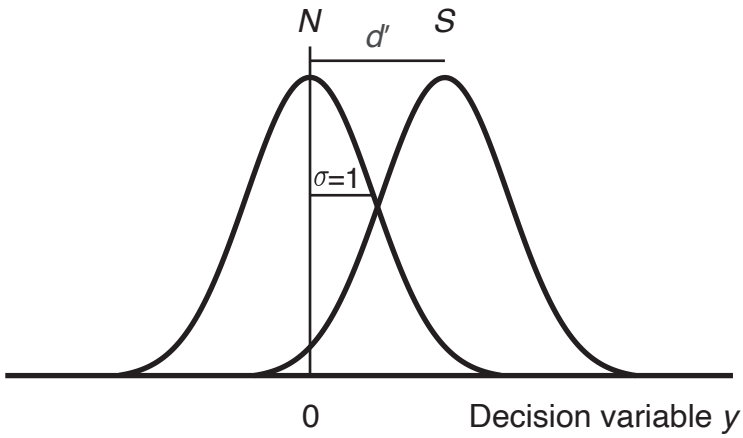

Figure 3. Standardized signal detection theory. Once standardized, the separation between the distributions (d') provides a metric for discriminability.

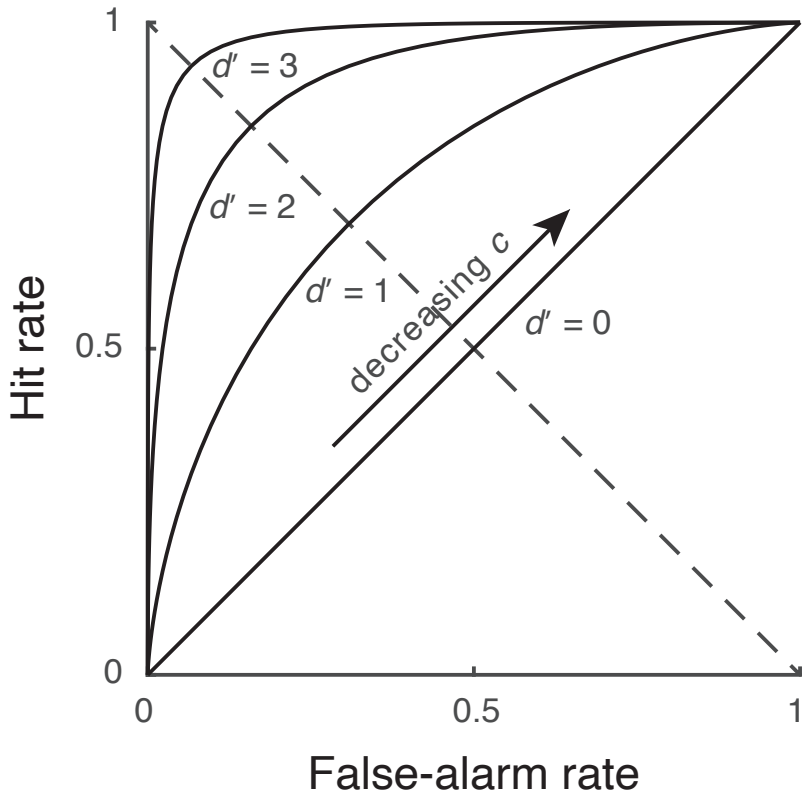

Figure 4. ROC curves for four values of d'. The dashed negative diagonal corresponds to predicted performance for a neutral criterion (where the two curves cross as in Fig. 1C), so that hit rate is identical to the correct-reject rate (i.e., one minus the false-alarm rate). For any value of d' the corresponding curve is traced from lower-left to upper-right as the criterion c decreases.

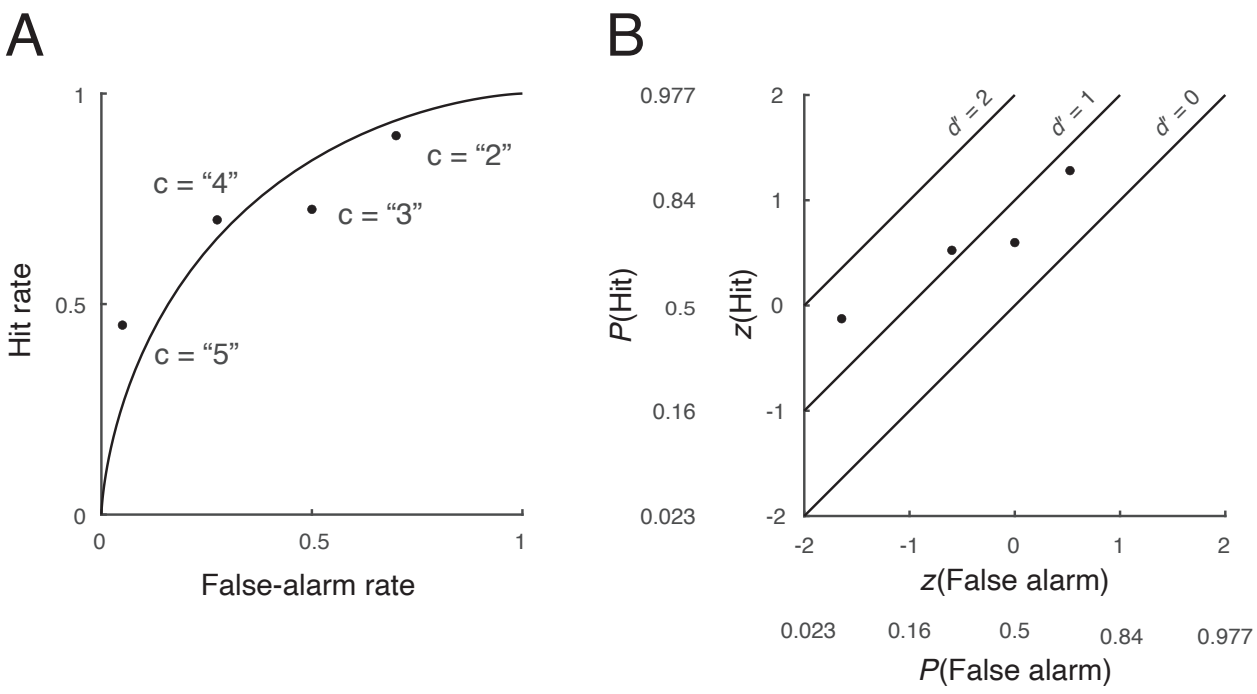

Figure 5. Estimation of d' using confidence ratings and the ROC curve. (A) From five detection confidence levels we can derive four points on the ROC curve and determine d' from the best-fitting ROC curve. (B) Plotted on probability (*z*-score) axes, the ROC curves become lines with slope 1.

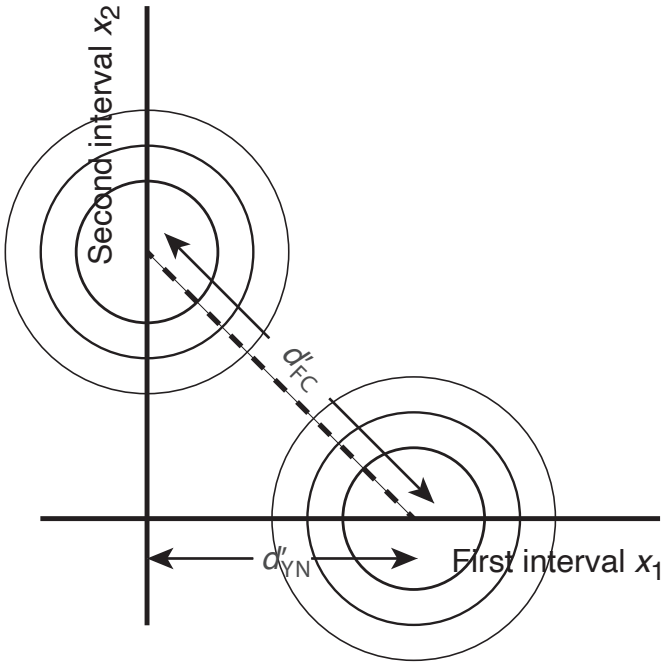

Figure 6. Two-alternative forced choice (2AFC) task. In each trial, there are two stimuli, leading to measurements x_1 and x_2. A no-signal measurement is, on average, zero. A measurement of a signal is, on average, equal to d'_{YN}—the value of d' one would have on a single-interval, yes-no task. The concentric circles represent the bivariate distribution of (x_1, x_2). The distance between the two distributions, d'_{FC}, governs performance in the 2AFC task.

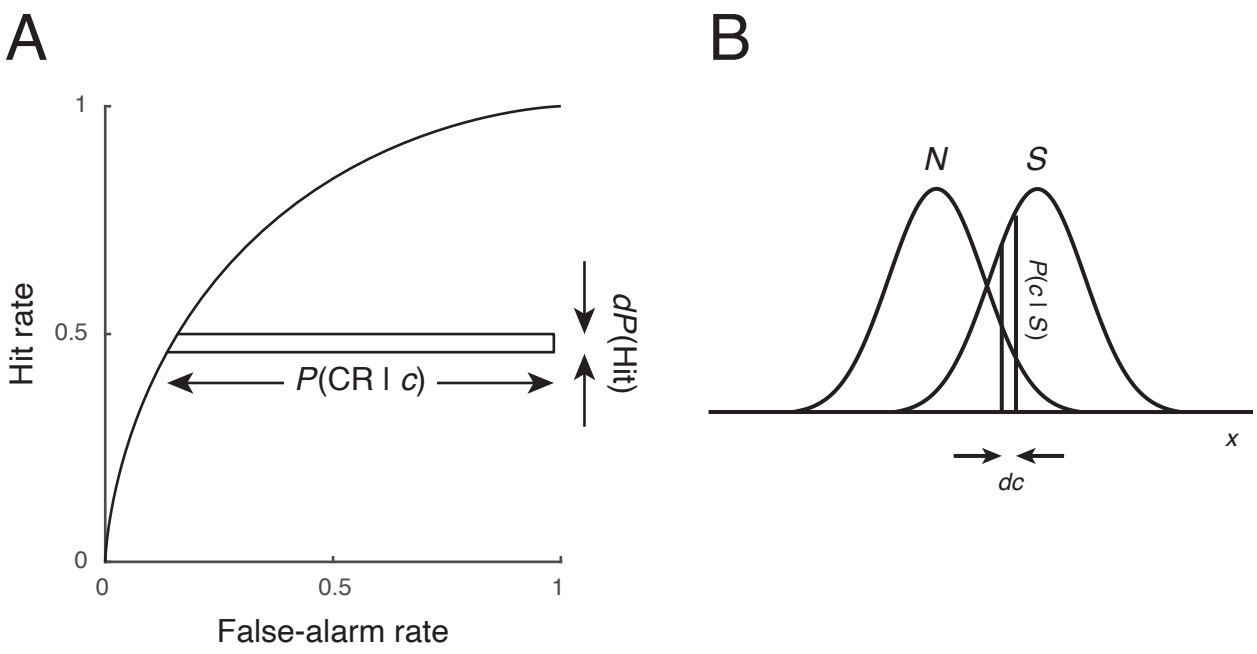

Figure 7. Demonstration that the area under the ROC equals predicted 2AFC performance. (A) The area is the sum of differential areas with width equal to the probability of a correct reject for a criterion that leads to the hit rate for that rectangle. (B) The height dp(Hit) of the rectangle in (A) is equal to the area of the rectangle shown here.

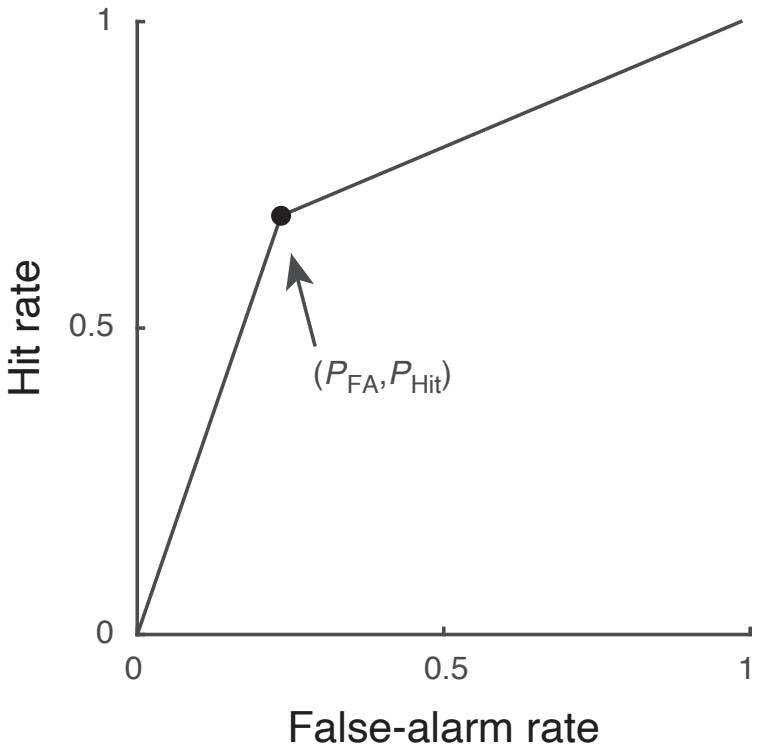

Figure 8. The ROC curve resulting from a threshold theory in which the detect state is entered on no-signal trials with probability P_{FA} and with probability P_{Hit} on signal trials.

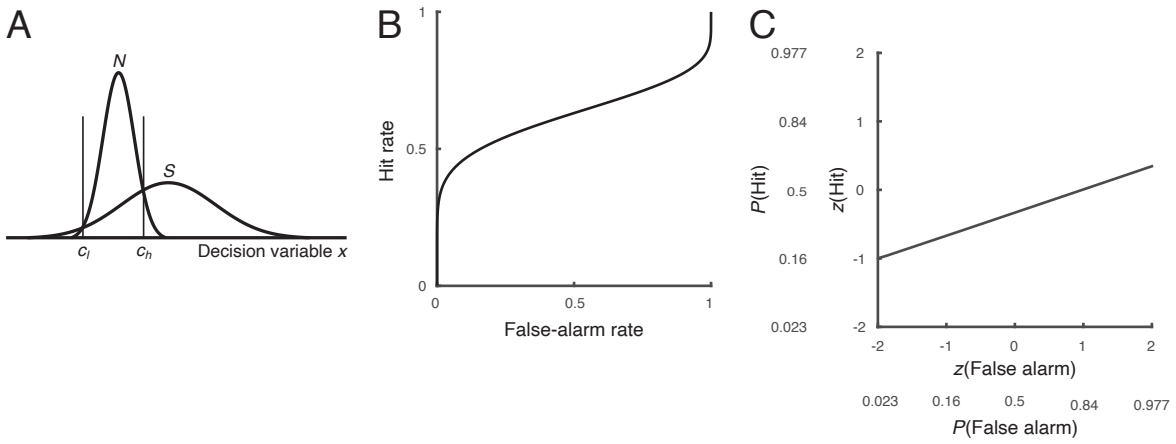

Figure 9. SDT with unequal variances. (A) The measurement distribution $p(x \mid S)$ has a standard deviation three times larger than $p(x \mid N)$. Note that S is more likely than N both to the right of c_h *and* to the left of c_l. (B) If only a single criterion is used, a non-convex ROC curve results. (C) Plotted on probability (*z*-score) axes, this ROC curve is a straight line with slope equal to the ratio of the noise and signal standard deviations (here: 1/3).

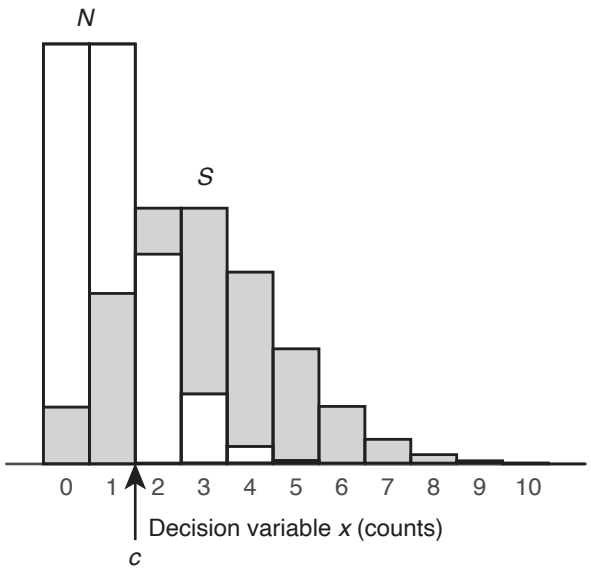

Figure 10. Signal detection theory with a discrete measurement distribution (Poisson).

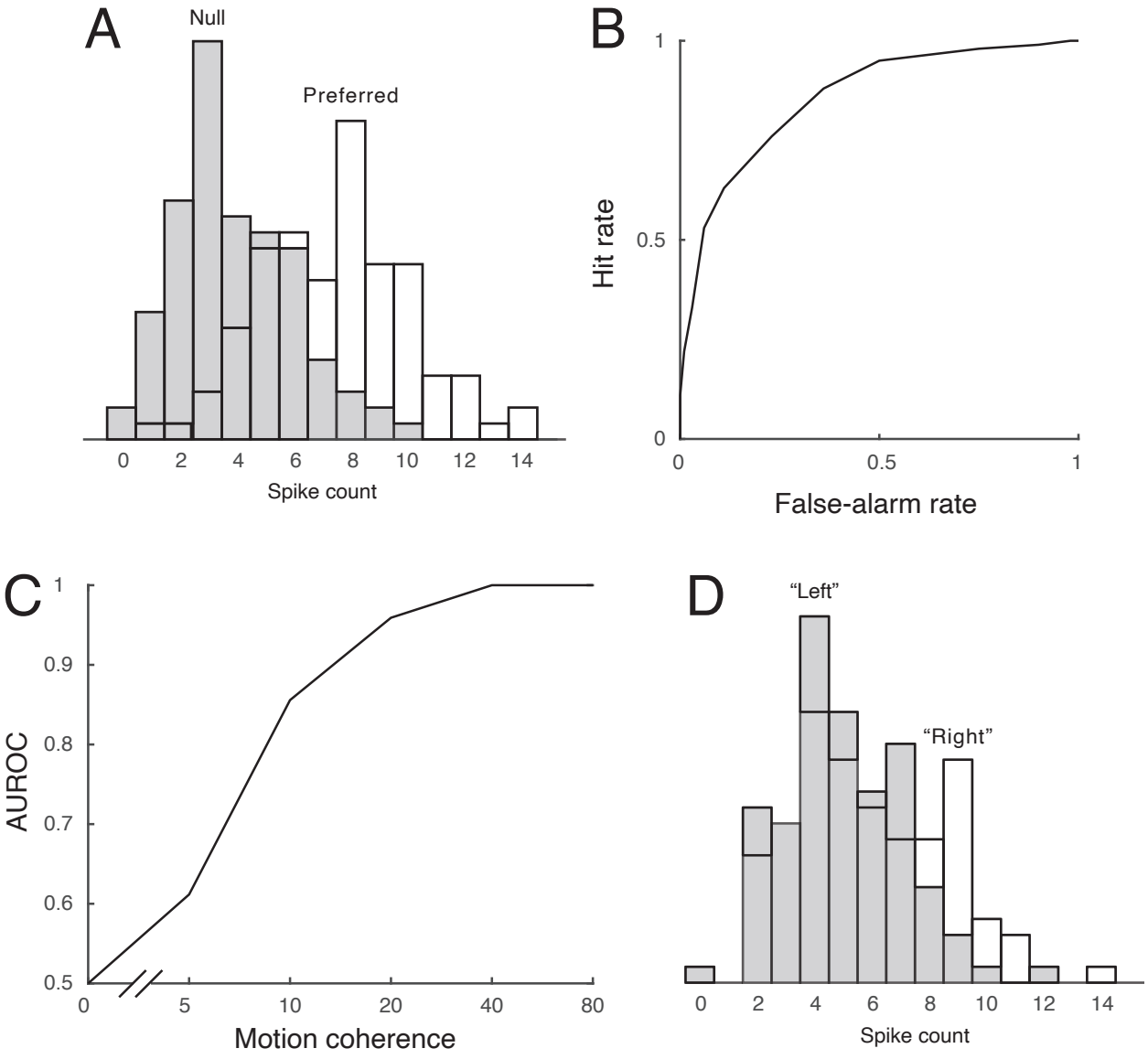

Figure 11. Signal detection theory applied to neural spiking data. (A) Simulated histograms of the number of action potentials (spikes) in response to a brief random-dot motion stimulus moving in either the neuron's preferred direction (white) or the opposite (null) direction (gray). (B) ROC curve derived from the data in (A). (C) Neurometric function—the area under the ROC (as in panel B) as a function of motion coherence. (D) Histograms of spike counts conditioned on the monkey's response for a zero-coherence motion stimulus, that may be analyzed as above to determine "choice probability".

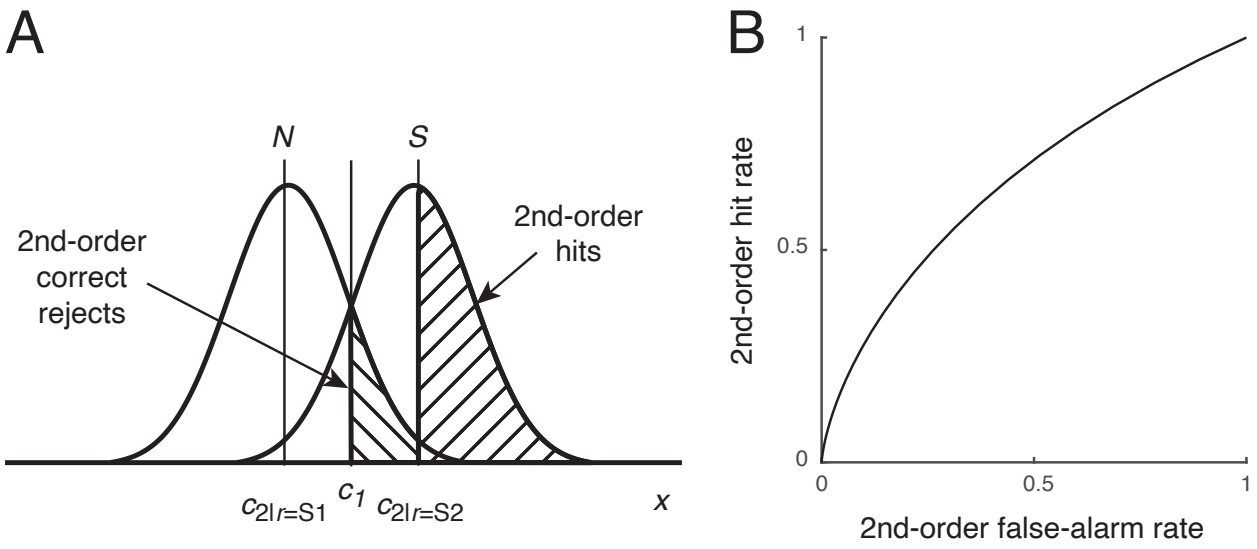

Figure 12. SDT and metacognition. The observer first discriminates between stimuli N and S and then reports whether they have low or high confidence in that judgment. (A) Criterion c_1 determines the discrimination decision. Neighboring criteria $c_{2|r=N}$ and $c_{2|r=S}$ determine the confidence response. The denoted areas correspond to high-confidence "S" reports when the stimulus was indeed S (a second-order hit); low-confidence "S" reports when it was, in fact, N (a second-order correct reject). (B) Sweeping $c_{2|r=S}$ across all possible values yields a second-order ROC, that can be compared to confidence data.