Is the Area Measure a Historical Anomaly?

J. D. Balakrishnan and Justin A. MacDonald, Purdue University
Hamid S. Kohen, California Institute of Technology

Abstract Green's well-known area theorem establishes an equivalence between the area under the yes-no ROC curve and the percent correct of an unbiased observer in a two-alternative forced-choice (2AFC) task with equivalent stimuli. In this article, we show that this conversion from yes-no detection data to hypothetical performance in a 2AFC task is unnecessary: The same yes-no detection data that are used to compute the area statistic can always be used to compute the percent correct of an unbiased observer in the yes-no detection task itself. We also show that the ROC curve may not be the ideal graphical device for many investigators. A more natural representation of the difficulty of a discrimination task is obtained by plotting the distribution of the posterior betting odds under equal base rates, which can be estimated from their distributions under unequal base rates. Finally, unlike the area measure and other traditional detection theory statistics, both the yes-no percent correct measure and the odds distributions generalize in an obvious and direct way to classification paradigms with more than two responses (e.g., identification).

The measures and experimental methods developed under the auspices of the theory of signal detection have had a powerful impact on research practices in many areas of psychology. One of the most influential contributions of this body of work is Green's so-called area theorem (Green, 1964; Green & Swets, 1974, p. 45-49), which establishes an equivalence between the area under the ROC curve in a yes-no detection task and the proportion of correct decisions of an unbiased observer in a comparable two-alternative forced-choice task (2AFC). In part because of this theoretical relationship, many researchers plot ROC curves and compare their areas visually or compute an estimate of this total area, A_{ROC}, to gauge the difficulty of a detection task in a rigorous way.

One of the particularly appealing features of the area measure is its "distribution-free" property: Green's result holds for virtually any univariate model of the distribution of sensory effects of the stimuli across trials. Since the equal-variance, normal distribution assumption underlying the parametric measure *d'* is often empirically untenable, A_{ROC} is sometimes regarded with less suspicion than *d'* (e.g., Macmillan & Creelman, 1991). A_{ROC} is also more convenient for some investigators because it is a percent correct score, making it easier to interpret and relate to intuitions about discrimination behaviours (Irwin, Hautus, & Butcher, 1999; Smith, 1995).

While questioning neither the theoretical nor the practical importance of Green's result, we argue in this article that the role that the area measure has played over the past several decades is a historical accident. Apparently, detection theorists overlooked a similar measure of yes-no detection performance that could have been proposed and that would have made it difficult to justify the area measure. This alternative statistic is simply the proportion of correct detection responses of an unbiased decision-maker in the yes-no detection task itself. In other words, there is no need to convert performance in yes-no detection to performance in an analogous 2AFC task – the same information that is used to compute the latter can always be used to compute the former. Moreover, there appear to be no statistical or pragmatic advantages of the conversion to 2AFC performance (at least, none that we could find). Furthermore, unlike A_{ROC}, this yes-no measure of yes-no detection generalizes immediately to classification tasks with more than two responses (*n*-choice classification). Thus, under the same basic assumptions associated with the detection theory framework, it is possible to compare bias-corrected percent-correct scores from any two classification tasks, without involving a parametric model or a parameter estimation routine.

To some extent, the "surrogate" nature of the area measure can be attributed to its origins in the geometry of the ROC curve rather than a fundamental principle of decision-making or probability theory. We will also argue that a different graphical device, directly related to the yes-no detection measure we define, may be more appropriate than the ROC curve for many research purposes. Finally, since both measures properly control for decision-making biases only under the assumptions of detection theory, we also explain how we believe they ought to be utilized and interpreted in light of recent results reported by us (Balakrishnan, 1998a,b, 1999; Balakrishnan & MacDonald, 2001a, b) and others (Van Zandt, 2000) that appear to challenge the foundational assumptions of the detection theory approach. Since the ideas involved all stem directly from well-developed areas of statistical decision-making and probability theory, and not all of these are as familiar to psychologists as the detection theory, models, we review some of the basic concepts of decision theory first before introducing the new measure.

Decision Theory and Performance Assessment

In the classical yes-no detection (discrimination) task, the subject is asked to identify which of two stimuli was presented on a given trial by reporting its prescribed label, *A* or *B*, where *A* could be, for example, the signal stimulus and *B* could be noise. Often the researcher's interest is centred on the "effect" of the stimulus (what did the subject see or hear, or what memories did it invoke?), and the accuracy of the responses across trials is a proxy for the accuracy of these perceptual or memory processes. Of course, choosing a name to report involves a decision-making process, and for a variety of reasons it is dangerous to assume that subjects simply "report what they perceive." To give just one of many possible examples, if the subject develops (rightly or wrongly) hypotheses about the probable order in which the experimenter is likely to present the stimuli, he or she may attempt to be "strategic" in choosing a response, somehow combining the information obtained from the stimulus presentation with other knowledge or beliefs about the experiment. In fact, researchers often observe what appears on the surface to be idiosyncratic behaviour on the part of some subjects, who may respond, say, *A* much more frequently than *B*, even when the two stimuli are presented equally often (i.e., their base rates are equal). Furthermore, if one of the two stimuli is in fact presented more frequently than the other, the task is easier in the sense that the subjects can avoid the rare stimulus response unless they have strong information from their senses or memory to justify it. Even if they have great difficulty distinguishing the two stimuli, their responses should be correct at least as frequently as the higher of the two base rates – hence overall accuracy is not a good measure of the accuracy of perception.

At least in rough terms, these are the basic issues that detection theorists brought to the attention of psychologists in the 1950s. Decision-making strategies can play an important role in shaping a subject's behaviour, and the analysis of discrimination data should somehow take these processes into account. The approach that the detection theorists developed, however, capitalized on only one type of decision-making process, the selection of a response at the "end" of the trial, when all of the perceptual information has been collected (or a memory search has been completed). From a broader point of view, the subjects in a discrimination task must make three fundamentally different kinds of decisions, because they are controlling a system with three connected components, a sampling device (e.g., a sensor), a knowledgebase, and a response generator. As a trial unfolds, the subject must continuously operate the sensor, directing it toward the stimulus in some manner and recording the resulting output. This "direction of attention" involves decision-making. At some point, the subject must stop collecting information from the sensor, which implies another type of decision process, and finally combine information from the sensor and the knowledgebase to select a response, *A* or *B*, to be passed on to the response generator.

The operation of the sensor involves some potentially crucial decisions about what kinds of observations to sample from the outside world during the course of the trial. For example, subjects must decide when and where to look in a visual discrimination task, how to focus their gaze, and how to organize the movements of their eyes. Some of these decisions may be "forced" on the subject in an important sense. For example, fatigue may affect the way in which the subject attends to the display, causing the quality of the information collected to deteriorate with time. Although the subject might not deliberately choose to sample the stimulus in this less vigorous fashion, from a formal perspective, there are more than one possible sample modes, and hence there is a *sampling rule* that governs the behaviour of the sensor. Of course, once the nature of this sampling process is identified (if this is possible), it is up to the investigator to interpret the psychological significance of the result (e.g., whether the sampling process is less than optimal due to involuntary factors or due to deliberate choices on the part of the subject).

In detection theory, the sampling process is assumed to be passive or uncontrolled (the sensor itself cannot be manipulated) and is therefore constant across trials and experimental conditions in an important sense. In formal terms, the sampling rule that governs the nature of the sampling process is invariant across trials and conditions. From a more general purview, subjects might make important decisions about how to sample the stimulus, and biases in these decision processes may affect their performance. Different biases in different conditions may lead to performance differences that are due to decision-making styles rather than sensory perception per se.

The second type of decision, when to terminate the sampling operation (or, more properly, when to generate a response, since the sensor might continue to operate after the response has been selected), is already recognized by many psychologists as an important issue in measurement and theory testing practices, largely due to the apparent prevalence of speed/accuracy trade-off effects. Even when instructed to be as accurate as possible, most subjects still seem to be motivated to respond as quickly as possible, sometimes even before the stimulus has disappeared from the display. Presumably this is due to the extra effort involved in spending more time on a trial, or perhaps because they feel some incentive for identifying the stimulus more quickly.

The factors that determine when the response will be generated define what decision theorists call the *stopping rule*. Signal detection theory assumes that the time at which the response is generated is independent of (cannot be predicted from) the information obtained from the sensor during the course of a given trial. To appreciate the significance of this assumption, consider the following two possible stopping rules. Prior to the start of a trial (i.e., the presentation of the stimulus), Observer A decides how long to wait before generating a response. The sensor is then activated for this predetermined interval, and its output during this period is then converted to a response. Observer B, on the other hand, decides to continue to observe the display until enough information has been collected to warrant a response, that is, when the perceived strength of the evidence at any time *t* exceeds a threshold for responding (which may be constant or may vary in time). Although the time spent sampling the stimulus for Observer A may vary from trial to trial, it will not depend in any way on the output of the sensor. For example, if the first few sample values are relatively uninformative, this does not imply that the total sampling time is likely to be large. Observer B, on the other hand, is likely to spend more time processing the stimulus on trials when the information provided by the sensor early on is weak, and less time when this initial sensor information is strong.

The dynamic or sequential sampling models assume that subjects behave in the manner of Observer B, continuously evaluating the incoming sensory (or memory) data and stopping when the total amount of accumulated evidence meets some criterion (e.g., Baranski & Petrusic, 1998; Busemeyer & Townsend, 1993; Diederich, 1997; Edwards, 1965; Link & Heath, 1975; Luce, 1986; Ratcliff, 1978; Ratcliff & Rouder, 2000; Ratcliff, Van Zandt, & McKoon, 1999; Smith, 2000; Stone, 1960; Townsend & Ashby, 1983). In these models, decision-making and perception (the sampling process) are inherently intertwined, or, in more formal terms, the stopping rule is a condition set on the information accumulation process, whereas in detection theory, the stopping rule is set on time alone.

The third (and final) type of decision process is the one that connects possible sensor output records at the point of the response to permissible responses to be executed. Whatever factors determine the amount of time taken before a response is selected, the sensor will produce some output from the start of the trial until the point of the response, and thus there must be some mapping between possible sensory records and responses. As we pointed out earlier, detection theory focuses on this terminal decision process – the *decision rule* – and its effects on the subjects' ability to correctly identify a stimulus (e.g., Green & Swets, 1974; Macmillan & Creelman, 1991). Since the sampling and stopping rules adopted by the subjects are assumed to be fixed in these models, the only way that decision processes can affect a subject's discrimination performance (in the model) is by way of the decision rule. The classical detection theory model is therefore designed to identify this decision rule so that the ability of the sensor to discriminate the two stimuli can be quantified.

Although we will not need to propose or discuss any dynamic model of perception in this article, it is important to recognize that the issue of the contribution of the decision rule to overt performance need not (or at least should not) be raised only when the sampling and stopping rules are assumed to be fixed. Regardless of how these other two rules are defined, there will be a sensory record at the point at which the response is generated and a mapping from these records to responses; hence there will always be a decision rule. For better or worse, this mapping will have some effect on performance. The area measure and the alternative we will develop should be taken as measures of the quantity of this information available at the point of the decision, taking into account any suboptimal properties of the decision rule (the mapping) and also the base rates of the stimuli. Since these measures do not take into account biases in the sampling and stopping rules, they control for all decision-making biases only under the assumption that these other decision processes can be safely ignored. In the next section, we show how it is possible to study the decision rule and determine whether it is biased or suboptimal and, consequently, what it contributes to observable performance.

Optimal Decision Rules and Performance Measures

The samples collected by the subject during the course of a trial can be thought of as the outcome of the sampling process (the output of the sensor), or equivalently, the "effect" of the stimulus presentation on the observer. Whether the observer is aware of it in any sense or not, for each possible effect $E = e$ (where $E = e$ may be a unidimensional or multidimensional vector, i.e., arbitrarily complex), there is a corresponding pair of objective posterior probabilities that define the probability that the stimulus was an A, $O_A = p(S = A \mid E = e)$, and the probability that the stimulus was a B, $O_B = p(S = B \mid E = e)$. That is, knowing that the effect on this trial was $E = e$ and combining this information with any other knowledge available (e.g., the base rates), what is the true probability that the stimulus was an A (e.g., a signal) and what is the probability that it was a B (e.g., noise)? These two values determine the betting odds in favour of the A and B responses, respectively. They are not subjective beliefs or psychological constructs – they are facts. For example, if $E = \{2, 12, -3\}$ and $O_A = 0.7$, this means that on exactly 70% of the trials in which the effect is $E = \{2, 12, -3\}$, the stimulus will turn out to be an A.[1] Furthermore, since one of the two possible responses must be correct and the other incorrect, the two posterior probabilities are "complementary," $O_A = 1 - O_B$. Thus, if O_A is greater than one-half, the stimulus is more likely to be an A than a B, and if O_A is less than one-half, the stimulus is more likely to be a B than an A.

Now suppose that a decision-maker's purpose is to maximize his or her chances of identifying the stimulus correctly. In this case, the appropriate decision rule is obvious (i.e., respond A if $O_A > O_B$ and respond B if $O_B > O_A$). The response is arbitrary when these two probabilities are equal. Notice that this prescription can be given irrespective of the base rates of the stimuli – these values have already been factored into the computation of O_A. Thus, for any set of base rates, the decision-maker maximizes the probability of a correct discrimination judgment under a given information state by choosing the response that is more likely to be correct when this information state occurs (i.e., by following the true betting odds).

Task Difficulty

Regardless of how complex the effect of the stimulus may be (i.e., the complexity of the sensor output), some value of the "test statistic," O_A, must be realized on each trial (whether the subject makes proper use of it to select responses or not is a different matter). Since O_A is always a single real number, it is a univariate random variable with one distribution on A trials and another distribution on B trials. In other words, regardless of whether any of the assumptions of detection theory about the sampling and stopping rules of the discrimination process are valid, two univariate distributions must always exist in a well-defined discrimination task (unless the laws of probability do not apply to the subjects for any description of the effect of the stimulus presentation). Many psychologists would probably associate the statement that two univariate distributions describe the effect of the stimulus presentation on the observer with the detection theory models – in fact, this assumption will always be found in all stochastic models of any classification process, including the dynamic models cited earlier. Thus, no matter what mechanisms drive the classification process, the variable O_A will have some univariate distribution across trials, and the amount of information contained in the effects across trials will be embodied in this O_A distribution.

Once the sampling and stopping rules are set (conforming with the detection theory assumptions or not), the set of possible information states (effects) at the point of the decision are also set, as well as the relative frequencies with which they will occur. Since each effect has some corresponding value O_A, the objective difficulty of the task is determined by these relative frequencies. The task is difficult if the posterior probability, O_A, tends to be close to one-half (and hence O_B also tends to be close to one-half). In this case, the true betting odds are close to one-half and so the best the decision maker can do is to be correct slightly more than 50% of the time. In contrast, an easy task is one in which O_A tends to be close to 1 or 0 (and hence O_B is also close to 0 or 1). In such a case, the observer's overall probability correct score will be close to 1 (assuming that he or she follows the optimal decision rule).

Likelihood Ratios, Base Rates, and the Betting Odds Distribution

Because the test statistic O_A takes into account the base rates of the stimuli, its distribution across trials in a given condition will depend not only on the sensitivity of the perceptual system to the distinguishing features of the stimuli, but also on the base rates of the stimuli. For example, if the two stimuli are virtually identical, but the A stimulus is almost always presented

[1] Technically, this statement only makes sense if there are a finite or countably infinite number of possible effects. See, for example, Cox and Hinckley, 1990, for a treatment of continuous measures.

(i.e., the *A* stimulus base rate is close to 1), then O_A will almost always be close to 1 (favouring the *A* response) and the task is easy in the sense that the optimal decision-maker can be correct on almost all of the trials, even though the two stimuli might be difficult to discriminate. This is one reason why detection theorists usually focus their discussions of the statistical decision-making concepts underlying the detection models on the likelihood ratio,

$$l_{E=e} = \frac{f(E=e \mid S=A)}{f(E=e \mid S=B)},$$

rather than O_A or on the stimulus odds ratio,

$$O_{E=e} = \frac{O_A}{O_B},$$

which would also be affected by the base rates. The likelihood ratio is the relative frequency (or density) of an effect on *A* trials divided by the relative frequency (or density) of this effect on *B* trials. Although it is a function, it is important to remember that it is also a random variable that takes on a specific value on each trial of the experiment. That is, since the effect, $E = e$, will vary from trial to trial, the likelihood ratio corresponding to the effect will also vary from trial to trial. Furthermore, on *A* trials, the likelihood ratio computed from the varying effects will have relative frequencies determined by the *A* stimulus distribution, whereas on *B* trials the effects will have relative frequencies determined by the *B* stimulus distribution. In other words, the likelihood ratio corresponding to the effect of the stimulus will have one of two distributions, depending on which stimulus was presented. We will denote these two variables as l_A and l_B, for the *A* and *B* trials, respectively.

If the sampling and stopping rules are both unaffected by changes in the base rates (an assumption that we have argued is grossly violated empirically – see below), the distribution of the likelihood ratio on *A* trials (the distribution of the random variable l_A), and the distribution of the likelihood ratio on *B* trials (the distribution of the random variable l_B), will also be invariant across different base-rate conditions and under different biases in the subjects' decision rules.

Since the likelihood ratio function *l* is asymmetric, in the sense that it takes on values between 0 and 1 when the denominator, $f(E = e \mid S = A)$, is greater than the numerator, $f(E = e \mid S = B)$, but values between 1 and infinity when $f(E = e \mid S = A)$ is less than $f(E = e \mid S = B)$, the distributions of the log likelihood ratio, log (*l*), on *A* and *B* trials are usually preferred to *l*. A relatively "complete" representation of the difficulty of a detection task, controlling for the base rates and any biases in the decision rule, can therefore be obtained by plotting the distributions of the log likelihood ratio, log (*l*), on *A* and on *B* trials.

In principle, we could stop here and recommend that these two distributions be estimated from data and plotted whenever a researcher wishes to study discrimination performance in a manner that takes the decision rule (but not the sampling and stopping rules) into account. However, this representation requires plotting two distributions, whereas the area measure is derived from a single curve. Moreover, as measurement scales go, the probability of a correct response (the betting odds) is probably more easily interpreted for most researchers than a log likelihood ratio. Since O_A is affected by the base rates, we cannot use its distribution as a basis for comparing performances across conditions that may involve different base rates. Fortunately, however, there is a simple solution to this problem. Specifically, given any value O_A obtained from an experiment with any base rates, it is possible to compute the corresponding (or, "corrected") value that O_A *would have been* on this trial if the base rates had been equal. Using this conversion formula as a starting point, the best possible performance that should be expected from the subject if the base rates were equal can be computed. Note that this is also what the area measure is intended to offer – the best possible performance – but for a different paradigm (2AFC).

To see how this conversion is carried out (and what it means), suppose that the base rates in an experiment are 3 to 1 in favour of the *A* stimulus (i.e., the *A* stimulus is presented on 75% of the trials and the *B* stimulus on the remaining 25% of the trials). On a given trial, suppose that the subject stops attending to the stimulus after collecting two samples, resulting in the effect $E = \{3, 5\}$, where 3 is the value of the first sample and 5 is the value of the second sample obtained on this trial. Suppose further that the corresponding value of O_A for this particular effect is 0.6. In other words, given the perceptual information that the subject has collected on this trial (3,5), there is a 60% chance that the stimulus is an *A*. Now, applying Bayes' rule, we know that on this particular trial,

$$O_A = 0.6$$

$$= \frac{f(E=e \mid S=A) \cdot p_A}{f(E=e \mid S=A) \cdot p_A + f(E=e \mid S=B) \cdot p_B}$$

$$= \frac{f(E=e \mid S=A) \cdot 0.75}{f(E=e \mid S=A) \cdot 0.75 + f(E=e \mid S=B) \cdot 0.25},$$

where e is the effect {3, 5} and p_A and p_B are the base rates of the A and B stimuli, respectively.

Therefore, if we wish to determine what O_A would have been on this trial if the base rates had been equal, we simply replace the values .25 and .75 in the equation above with the value .5 and find the resulting odds value. Fortunately, this can be done without knowing the values of $f(E = e \mid S = A)$ or $f(E = e \mid S = B)$. The conversion from O_A to the "corrected" odds value, O_A^*, is simply:

$$O_A^* = \frac{O_A \cdot p_B}{O_A \cdot p_B + (1 - O_A) \cdot p_A}. \tag{1}$$

In our example, O_A was .6, p_A was .75 and p_B was .25, and so the conversion can be performed using

$$O_A^* = \frac{O_A \cdot 0.25}{O_A \cdot 0.25 + (1 - O_A) \cdot 0.75} = 0.33.$$

The value of O_A^* can also be computed from the relative frequencies of the effect on A and B trials, using the formula

$$O_A^* = \frac{f(E = e \mid S = A)}{f(E = e \mid S = A) + f(E = e \mid S = B)},$$

which is derived from Equation 1 by substituting the Bayesian expression for O_A into the equation and simplifying the result.[2]

The values of O_A^* indicate how difficult the individual trials would be if the base rates were equal on each trial. However, because the A stimulus is presented more frequently in the experiment, recording these O_A^* values across trials and examining their distribution across trials of the experiment would not provide a suitable performance index, since this distribution would still be affected by the base rates. What is needed is the distribution that these O_A^* values would have across trials if the A and B stimuli were presented equally often. One more step is needed to calculate this "equal base-rate" distribution from unequal base rate data. This step is to find the relative frequencies of the O_A^* values on A trials, $f(O_A^* \mid S = A, p_A \neq p_B)$, and their relative frequencies on B trials, $f(O_A^* \mid S = B, p_A \neq p_B)$, in the *unequal* base-rate experiment, and then take the average of these two to determine their frequency under *equal* base rates,

$$f(O_A^* \mid p_A = p_B) = \tfrac{1}{2}\, f(O_A^* \mid S = A, p_A \neq p_B) + \tfrac{1}{2} f(O_A^* \mid S = B, p_A \neq p_B). \tag{2}$$

The single, univariate function, $f(O_A^* \mid p_A = p_B)$, is the distribution that the betting odds values would have had across trials if the base rates had been equal and all other factors (i.e., the sampling and stopping rules) were held constant – in other words, it is a suitable basis for quantifying the amount of information contained in the effects of the stimuli on the observer.

Although this may seem like a complex series of transformations to introduce our new measurement curve, compared to the relatively simple definition of an ROC curve, it may be worth recalling at this point that Green's theorem, which is needed to justify the area measure, is by no means trivial. Perhaps more importantly, the measurement formulas we will eventually end up with are in fact simpler than the formulas needed to calculate the area measure, are derived in a direct and intuitive way from the odds distribution, $f(O_A^* \mid p_A = p_B)$, and involve intermediate steps that provide model-free tests for bias and suboptimality of the subject's decision rule.

Some Illustrative Examples: Classical Psychophysics

To see how the odds distribution controls for decision rule biases and base rates, but not for changes in stopping or sampling rules, we will work through two simple examples. First, suppose that the presentation of a stimulus invokes one of two possible sensory states, "detect" and "no detect", as in the classical psychophysical models. The A stimulus is presented on 80% of the trials and the B stimulus is presented on the remaining 20% the trials. For whatever reason, the subject decides to respond A when in the detect state and B when in the no-detect state. In such a case, the two possible effects, E, are: "detect-respond-A" and "no-detect-respond-B." Suppose that on A trials, the detect state is generated 60% of the time and the no-detect state is generated 40% of the time. On B trials, the detect state is generated 35% of the time and the no-detect state is generated 65% of the time.

Since the A stimulus is presented on 80% of the trials and the detect state will occur on 60% of these, 80% x 60% = 48% of the trials will be correct A responses. Of the remaining 52%, 20% x 65% = 13% will be correct B responses. Overall percent correct for this subject would therefore be 48% + 13% = 61%. Notice that since the subject could be correct 80% of the time by simply responding A on every trial, regardless of the detection state, this subject's decision rule is suboptimal (for accuracy). This is not surprising, given that the decision rule is unbiased in the detection theory sense.

[2] We thank Helena Kadlec for pointing out this relationship.

That is, the relative frequency of the detect state on *A* trials (60%) is greater than the relative frequency of this state on *B* trials (35%), and the subject always responds *A* in the detect state. Similarly, the relative frequency of the no-detect state on *B* trials (65%) is greater than the relative frequency of this state on *A* trials (40%), and the subject always responds *B* in the no-detect state. When the higher of the two relative frequency distributions always determines the response that will be emitted, the decision rule is unbiased. The unbiased decision rule is usually – but not always – suboptimal when the base rates are unequal. The bias is often needed to take account of the prior probabilities of the stimuli (and also for payoff asymmetries, etc).

Now let us see what the odds distribution computations would tell us about this subject. In this experiment, there are only two different observable responses, *A* and *B* (confidence, response time, etc., were not recorded). However, these responses uniquely identify each possible sensory state (detect -> respond *A* and no detect -> respond *B*). In such a case, there will be no difference between the true odds distribution defined by a complete description of the effects of the stimuli and the odds distribution estimated from observable behaviour. Of course, this is unusual. We will return to this issue of the differences between the true underlying odds distribution and the experimenter's estimate of it later.

The possible values of O_A in this example are

$$p(S = A \mid E = detect) =$$

$$\frac{p(\,E = detect \mid S = A)\cdot p_A}{p(\,E = detect \mid S = A)\cdot p_A + p(\,E = detect \mid S = B)\cdot p_B}$$

$$= \frac{.60\cdot.80}{.60\cdot.80+.35\cdot.20} = .873,$$

and

$$p(S = A \mid E = no\,detect) =$$

$$\frac{p(\,E = no\,detect \mid S = A)\cdot p_A}{p(\,E = no\,detect \mid S = A)\cdot p_A + p(\,E = no\,detect \mid S = B)\cdot p_B}$$

$$= \frac{.40\cdot.80}{.40\cdot.80+.65\cdot.20} = .711.$$

It should be clear from this that the subject should respond *A* in both the detect state and the no-detect state; hence the suboptimality of the subject's actual decision rule and an overall correct response rate of less than 80%. Notice also that because the subject's response is uniquely determined by the sensory state, it follows that $p(S = A \mid E = detect) = p(S = A \mid R = A)$ and $p(S = A \mid E = no\ \ detect) = p(S = A \mid R = B)$. This is why the true odds distribution and the estimated odds distribution obtained from observable behaviour will be identical in this example, as we noted earlier.

The next step is to compute, from the $p(S = A \mid R = A)$ and $p(S = A \mid R = B)$ values obtained from the unequal base-rate experiment, the corrected odds values using Equation 1. These are

$$O_{A,\,detect} = p(S = A \mid R = A) = .873 \Rightarrow$$

$$O^*_{A,\,detect} = \frac{.873\cdot.2}{.873\cdot2+.127\cdot.8} = .632,$$

and

$$O_{A,\,no\,detect} = p(S = A \mid R = B) = .71 \Rightarrow$$

$$O^*_{A,\,no\,detect} = \frac{.71\cdot.2}{.71\cdot2+.29\cdot.8} = .381.$$

The two values on the abscissa of the odds distribution are therefore .381 and .632. To complete the odds distribution, we need to find the relative frequencies with which these two corrected odds values would occur across trials when the base rates are equal. These are obtained using Equation 2,

$$f(O^*_{A,\,detect} \mid p_A = p_B) =$$

$$\frac{1}{2}\big(f(R = A \mid S = A) + f(R \quad A \mid S = B)\big) =$$

$$\frac{1}{2}\big(.6+.35\big) = .475,$$

and

$$f(O^*_{A,\,no\,detect} \mid p_A = p_B) =$$

$$\frac{1}{2}\big(f(R = B \mid S = A) + f(R = B \mid S = B)\big) =$$

$$\frac{1}{2}\big(.4+.65\big) = .525.$$

The heights of the odds distribution are therefore .525 at the abscissa value .381, and .475 at the abscissa value .632. In other words, on 47.5% of the (equal base rate) trials, the subject will be experiencing a subjective state that causes the *A* stimulus to have a .632

probability of being present, and on the remaining 52.5% of the trials, the subject will be experiencing a state that causes the *A* stimulus to have a .381 probability of being present (and hence, a .619 probability that the *B* stimulus is present). In other words, the betting odds are never very good, and hence sensitivity is fairly low.

Interpretation issues

In this relatively unrealistic example, we assumed that the information collected from the stimulus on each trial could take on only one of two possible values, detect or no detect, and further that there was a deterministic, one-to-one mapping between these states and observable responses, *A* or *B* (the subject always responded *A* when the sensor output was "detect" and *B* when the output was "no detect"). The two possible effects were therefore detect-respond-*A* and no-detect-Respond *B*. In a real situation, there will be (infinitely) many possible effects, and each possible observable behaviour will define a many-to-one grouping of these effects. The odds distribution calculated from the observable behaviours may therefore differ from any odds distribution that would be computed from a more detailed description of the events occurring between the presentation of the stimulus and the execution of an *A* or *B* response. However, the difference between the true distribution and the estimated one will not be arbitrary: The latter will be an "averaged" or "smoothed" version of the former, with some fortunate consequences.

To illustrate this important principle, suppose that instead of always responding *A* in the detect state and *B* in the no-detect state, the subject adopts a nondeterministic decision rule, responding *A* on 80% of the detect state trials and *B* on the remaining 20% of these trials, and similarly responding *B* on 80% of the no-detect trials and *A* on the remaining 20% of these trials. In this case, there are now four possible effects, detect-respond-*A*, detect-respond-*B*, no-detect-respond-*A*, and no-detect-respond-*B*. Yet, the experimenter still has only the two observable responses to work with, generating the two statistics, $p(S = A \mid R = A)$ and $p(S = A \mid R = B)$. Each of these two observable probabilities will be an average of the two conditional probabilities associated with the individual effects they represent. That is, $p(S = A \mid R = A)$ will be a (weighted) average of the unobservable values $p(S = A \mid R = A, \textit{detect})$ and $p(S = A \mid R = A, \textit{no detect})$, and $p(S = A \mid R = B)$ will be a (weighted) average of the values $p(S = A \mid R = B, \textit{detect})$ and $p(S = A \mid R = B, \textit{no detect})$.

The weighting coefficients in this average will be determined by the relative frequency of the effect within the pair. So, for example, since the effect "$R = A$, *detect*" will occur more frequently on $R = A$ trials, the value $p(S = A \mid R = A, \textit{detect})$ will be weighted more than the value $p(S = A \mid R = A, \textit{no detect})$ in the average leading to the observed value $p(S = A \mid R = A)$. The observable odds distribution in this example would therefore be an averaged version of the true odds distribution, with some important consequences. The more effects the experimenter is able to observe, the more accurate should be the resulting odds distribution (provided that the experimenter records properties of behaviour that are related to the discrimination process, such as response time and response confidence – we return to this issue later).

It is also important to remember that the information contained in the odds distribution estimated from experimental data in a given experimental condition may also depend on the stopping and sampling rules adopted by the subject in this condition, as we pointed out earlier. Suppose, for example, that the *A* stimulus is a brief flash of light on the right side of a display and the *B* stimulus is a flash of light on the left side. If the *A* stimulus is presented more often, the subject may decide to direct his or her attention to the right side of the display, where the stimulus occurs more frequently. This sampling rule bias would presumably change the nature of the sensory information elicited from the stimulus. Controlling for bias in the decision rule and the effects of the base rates on performance by computing the odds distribution would not control for this encoding effect (even if there were no averaging). The frequent presentation of the *A* stimulus might also cause the subject to spend less time examining the display before responding, since his or her accuracy level might still be reasonably high due to the base rate difference. This stopping rule bias would most likely decrease the estimated sensitivity level even after decision rules and base rates are controlled for. The impact of stopping rule biases can be studied (in a model-dependent manner) using some well-developed response time models (e.g., Ratcliff, 1978; Ratcliff & Rouder, 2000; Ratcliff, Van Zandt, & McKoon, 1999; Smith, 2000). However, none of the measures or approaches we know of can control for encoding biases. In each empirical situation, it is up to the investigator to make an assessment about the potential effects that these other kinds of biases might have on performance before drawing inferences about sensitivity.

Some Illustrative Examples: The Ratings Paradigm

The main point of the previous example was to show how observable responses are converted to probabilities and how the relative frequencies of these probabilities under equal base rate conditions are computed. These are the steps that would be followed when the researcher wishes to estimate the odds distribution from discrimination data in which only the discrimination judgment is recorded (i.e., in the cases in which area of the ROC curve would be estimated from a single point on the curve). Although many performance measures are derived from these two judgments alone (i.e., from the hit and false alarm rates), we would argue that, due to the averaging problem just discussed, it is always preferable to observe and record more information about the discrimination judgment. Response confidence is a logical choice from a statistical decision-making point of view, and is also the most popular additional record in detection theory analyses. However, it is also possible to use response times to subdivide the *A* and *B* responses into smaller subsets, since response time and confidence are strongly correlated. In this way, a measure of the odds distribution with much less averaging can be obtained without any modification of the classical yes-no detection paradigm.

In this section, we work through the steps involved in computing the odds distribution when the *A* and *B* responses can be partitioned into at least two different categories per discrimination judgment. Since our decision rule in the previous example was unbiased, this time we will also discuss the issue of how decision rule biases are controlled for using the odds analysis. The response categories could be confidence ratings (low vs. high confidence *A* responses, low vs. high confidence *B* responses) that the subject may be asked to assign to the discrimination judgment, or the speed of the response (fast vs. slow *A* responses, fast vs. slow *B* responses), or any other aspect of discrimination performance that is expected to subdivide the possible effects of the stimuli into subgroups with different objective odds.

Since it is the most common design, we will assume that the responses are four different confidence ratings, two for each discrimination response, and are labeled A_1, A_2, B_1, and B_2. We will also assume that the base rates of the experiment are 60 to 40 in favour of the *A* stimulus. From the data collected in this hypothetical experiment, the first step is to compute the four conditional stimulus probabilities associated with the four possible responses, that is, $p(S = A \mid R = A_1)$, $p(S = A \mid R = A_2)$, $p(S = A \mid R = B_1)$, and $p(S = A \mid R = B_2)$. The first of these is the proportion of correct judgments in the set of trials on which the subject responded A_1, the second is the proportion of correct judgments in the set of trials on which the subject responded A_2, the third is the proportion of incorrect judgments in the set of trials on which the subject responded B_1, and the last value is the proportion of incorrect judgments in the set of trials on which the subject responded B_2. These are the four values of O_A corresponding to the four possible responses in the experiment. Suppose that they turn out to be the following:

$$O_{A,R=A1} = .65,\ O_{A,R=A2} = .75,\ O_{A,R=B1} = .5 \text{ and } O_{A,R=B2} = .43.$$

The second step is to determine what these objective stimulus probabilities would be if the base rates were equal, using Equation 1. In this example, the result is:

$$O_A = .43 \rightarrow O_A^* = .33,$$
$$O_A = .5 \rightarrow O_A^* = .4,$$
$$O_A = .65 \rightarrow O_A^* = .55,$$

and

$$O_A = .75 \quad \rightarrow \quad O_A^* = .67.$$

These are the four values on the abscissa of the odds distribution. The last step is to determine how often these betting odds under equal base rates would occur when the base rates are equal. To do this, we must first calculate from the data the relative frequencies of each response on *A* trials and the relative frequencies of each response on *B* trials in the unequal base-rate experiment. Suppose that on *A* trials, these turn out to be the following:

$$f(R = A_1 \mid S = A) = f(O_A = .75 \mid S = A) = f(O_A^* = .67 \mid S = A) = .2,$$

$$f(R = A_2 \mid S = A) = f(O_A = .65 \mid S = A) = f(O_A^* = .55 \mid S = A) = .5,$$

$$f(R = B_1 \mid S = A) = f(O_A = .50 \mid S = A) = f(O_A^* = .40 \mid S = A) = .2,$$

and

$$f(R = B_2 \mid S = A) = f(O_A = .43 \mid S = A) = f(O_A^* = .33 \mid S = A) = .1.$$

In other words, the subject made the A_1 response on 20% of the *A* trials during the experiment, the A_2 response on 50% of the *A* trials, etc. (note that the sum of these percentages must add to 100%). These response frequencies also determine the frequencies of their corresponding betting odds (O_A) on *A* trials during the unequal base-rate experiment, and the corresponding betting odds that would have been calculated during the experiment if the base rates were assumed to be equal in this experiment, O_A^*, even though they were not equal. For example, since the A_1 response occurred on 20% of the *A* trials, the subject experienced the objective betting odds value of .75 on 20% of the trials during the actual experiment, and on each of these trials, the betting odds would have been calculated to be .67 if the base rates had been equal.

Suppose that on *B* trials, the relative frequencies are found to be:

$$f(R = A_1 \mid S = B) = f(O_A = .75 \mid S = B) = f(O^*_A = .67 \mid S = B) = .1,$$

$$f(R = A_2 \mid S = B) = f(O_A = .65 \mid S = B) = f(O^*_A = .55 \mid S = B) = .4,$$

$$f(R = B_1 \mid S = B) = f(O_A = .50 \mid S = B) = f(O^*_A = .40 \mid S = B) = .3,$$

and

$$f(R = B_2 \mid S = B) = f(O_A = .43 \mid S = B) = f(O^*_A = .33 \mid S = B) = .2.$$

We now know the betting odds values that would be experienced in the equal base-rate condition (the abscissa values of the odds distribution), and their relative frequencies on *A* and on *B* trials when the base rates are unequal. To determine how often these betting odds values would occur across both *A* and *B* trials when the base rates are equal (i.e., the heights of the odds distribution), we apply Equation 2,

$$f(O^*_A = .67 \mid p_A = p_B) = \tfrac{1}{2}\,(.1 + .2) = .15,$$
$$f(O^*_A = .55 \mid p_A = p_B) = \tfrac{1}{2}\,(.4 + .5) = .45,$$
$$f(O^*_A = .40 \mid p_A = p_B) = \tfrac{1}{2}\,(.2 + .3) = .25,$$

and

$$f(O^*_A = .33 \mid p_A = p_B) = \tfrac{1}{2}\,(.1 + .2) = .15.$$

The resulting odds distribution is shown in graphical form in Figure 1. The values on the abscissa (.33, .40, .55, .67) are the objective betting odds that would be experienced and the values on the ordinate are the relative occurrence frequencies of these odds under equal base rates. Interpreting the graph, on 60% of the trials, the objective betting odds will favour the *A* response (15% + 45%) whereas on the remaining 40% of the trials they will favour the *B* response. In this sense, the sensory system itself (not the decision rule) is biased toward the *A* stimulus in this particular example.[3] Since this sensory bias cannot be due to the decision rule, it must be due to a property of the sensor or how it is controlled by the subject (i.e., a bias in the stopping or sampling rule).

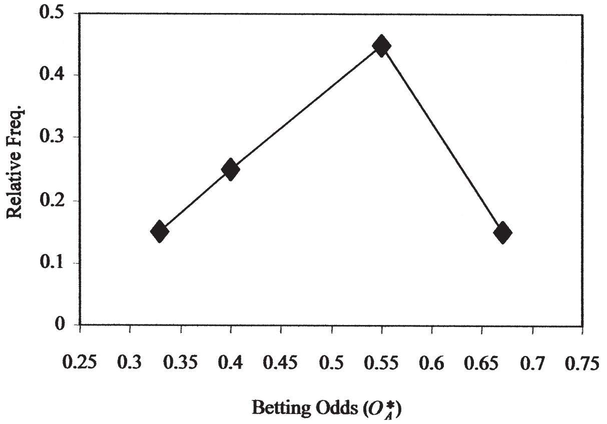

Figure 1. An odds distribution when there are only four possible values of the objective betting odds value in an experiment with unequal base rates (p_A = .6; p_B = .4). The curve describes how the amount of information at the point of the decision will vary from trial to trial in this hypothetical experiment, after controlling for the base rates. On about 60% of the trials, the sensory information would favour the *A* response (O^*_A > 0.5). Since the decision rule is controlled for, this effect must be due to inherent properties of the sensor or to biases in the sampling and/or stopping rules adopted by the subject.

Effects of decision rule bias and suboptimality

To determine whether the decision rule in this example is biased or unbiased in the detection theory sense, the relative frequencies of each possible *A* response on *A* trials must be compared to the relative frequency of the same response on *B* trials. If the latter is greater than the former for any one of these *A* responses, the decision rule is biased. Similarly, if the relative frequency of any *B* response is greater on *A* trials than on *B* trials, the decision rule is biased. The test for suboptimality is slightly easier: If any of the values of O_A derived from an *A* response in the experiment are less than one-half, or any of the O_A values derived from a *B* response are greater than one-half, then the decision rule is suboptimal.

In the example above, the decision rule appears to be optimal – each value of O_A is one-half or greater for the two *A* responses and less than or equal to one-half for the two *B* responses. The rule is also unbiased – each of the two *A* responses (at confidence level 1 and at confidence level 2) had higher relative frequency on *A* trials, and each of the two *B* responses had higher relative frequency on *B* trials. When the tests for bias or suboptimality fail for every response, however, it is possible that this result is merely due to the inevitable

[3] The fact that 60% was also the base rate of the *A* stimulus in the hypothetical experiment is merely a coincidence.

averaging over effects. That is, by asking the subject to report, say, six different levels of confidence instead of four, the tests for bias and/or suboptimality might turn out to be satisfied for some of these responses even though they failed when there were only four possible confidence responses. In same sense that knowing the heights of three points on a function might be sufficient to show that it is *not* a linear function but could never be sufficient to show that it *is* a linear function, these tests can be used to establish that the decision rule is biased or that it is suboptimal, but not to establish that it is unbiased or optimal.

Now let us consider the effects that a bias and suboptimality in the decision rule would have on the odds distribution. Suppose that there was a recording error in this experiment: Whenever the subject responded A_1, the computer always recorded the B_1 response, and whenever the subject responded B_1, the computer always recorded the A_1 response. This would cause the (apparent) decision rule to be both biased and suboptimal. It would also have a substantial effect on the hit and false alarm rate, and hence on the value of *d'*, if this parametric statistic was computed from the data. However, it has absolutely no effect on the estimated odds distribution. Looking closely at the steps involved, the results of the odds calculations depend on how strongly a given observable response predicts the *A* stimulus, $p(S = A \mid R = r)$, but it does not matter whether this $R = r$ is an *A* response or a *B* response (the four responses in the example could have been labeled, say, C_1, C_2, C_3, and C_4, without consequence) or whether the subject believes that the *A* response is more or less likely when selecting this response, or how the subject combines knowledge about the experiment with sensory information obtained on a given trial in order to select a response.

Diagnosticity

The limitations of an odds distribution obtained from empirical data do not ensue from any model assumptions that may be violated, but instead from the potential effects of averaging on the estimated odds distribution. The actual sequence of events occurring between the presentation of a stimulus and the observation of a response will almost surely be infinitely complex for any living organism. Many aspects of this sequence will be observable but unrelated to the discrimination process and many other aspects will be related but unobservable. Fortunately, it is possible to gauge the diagnostic strength of a given observable measure in a relatively simple manner. If the conditional probabilities, $p(S = A \mid R = r)$, generated by the measure (e.g., if *r* is response time, the *A* responses would be divided into subgroups by speed and the *B* responses would also be divided into subgroups by speed) vary widely across a substantial range (e.g., .01 to .99 would be extremely large), then the measure is very diagnostic. If these values are similar or identical, then they will provide little or no more information about sensitivity than that available already from the two discrimination judgments, that is, from $p(S = A \mid R = A)$ and $p(S = A \mid R = B)$.

The Odds Distribution and Task Difficulty

Because the odds distribution as we defined it above distinguishes between the two "sides" of the betting odds continuum (the optimal decision is *A* when $O^*_A > 1/2$ and *B* when $O^*_A < 1/2$), we will refer to it as the *articulated odds distribution* (*AOD*). It is also possible to ignore which response was prescribed by O^*_A and focus on the probability of a correct judgment. This *unarticulated* odds distribution (*UOD*) would be a plot of the probability distribution of the maximum of the two values, O^*_A and $1 - O^*_A$, across trials, since the maximum of these two is the probability that the optimal response, *A* or *B*, will turn out to be correct. The steps involved in computing the *UOD* would be similar to the steps in the example just given – we simply replace O_A with the maximum of O_A and $1 - O_A$ for each response, then use the same formula (Equation 1) to convert to the maximum value of O^*_A and $1 - O^*_A$.[4]

The difficulty of the task (with respect to accuracy) is relatively transparent in the *UOD* curves – if the area under the curve is massed toward the right edge, the task is easy, since the objective probability correct on a given trial is almost always close to 1. In other words, the unbiased and hence optimal decision-maker in the equal base-rate condition is usually making highly confident responses that turn out to be correct. If it is massed toward the left side of the scale, the task is hard, since the probability correct is always close to its minimum, 0.5. The decision-maker in this case is usually making very low confidence responses that often turn out to be incorrect.

Examples of the odds distribution under some familiar models of the distribution of sensory effects (e.g., normal) on *A* and *B* trials are shown in Figure 2. The *AOD* curves in the figure were obtained as follows. For each possible effect, $E = e$, the conditional probability that this effect is a sample from the *A* distribution, $O^*_A = p(S = A \mid E = e)$, is computed using Bayes' rule with

[4] Since two different observed values of O_A might now result in the same maximum value (e.g., $O_A = .2$ and $O_A = .8$), the relative frequencies of these two events need to be added together to obtain the relative frequency of this shared maximum value.

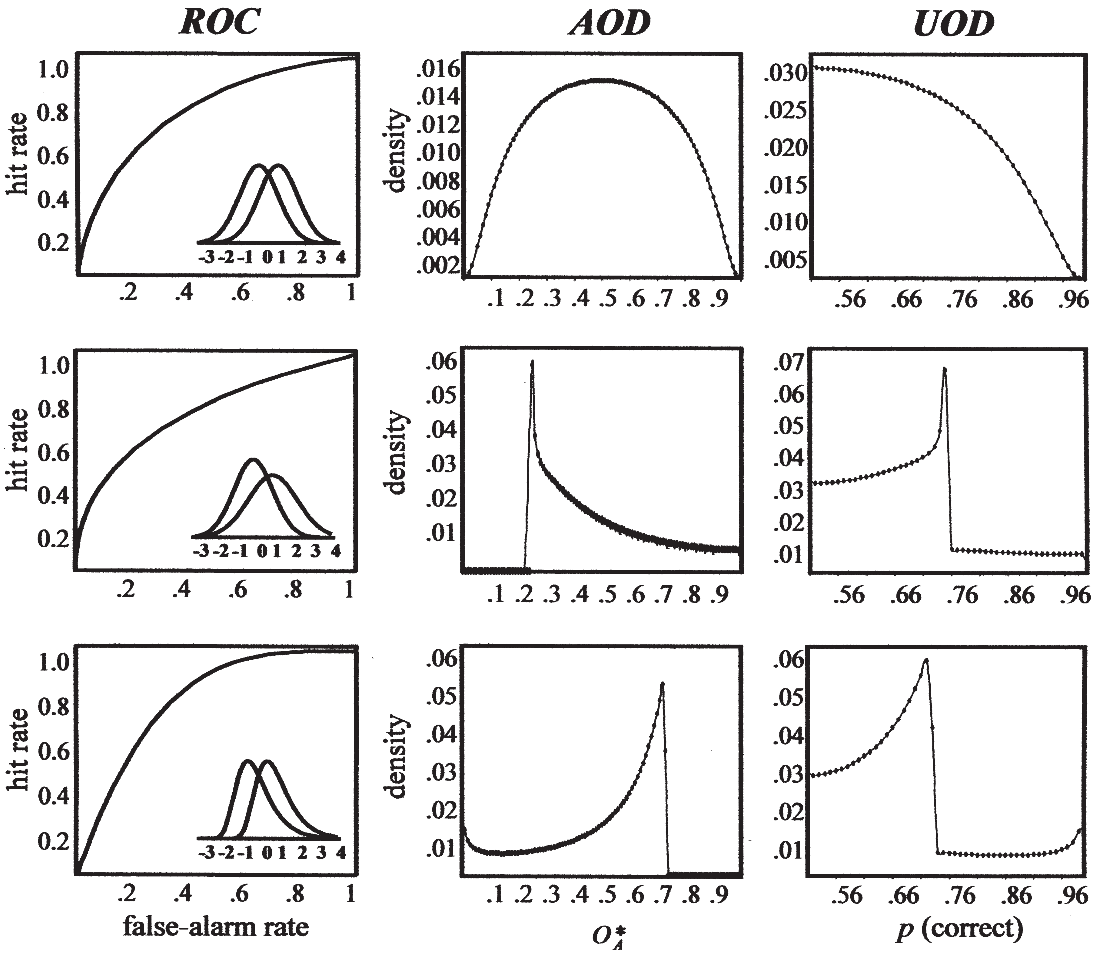

Figure 2. Comparison of *AOD* and *UOD* functions to their corresponding ROC curves. First Row: Normal Equal Variance model; Second Row: Normal Unequal Variance model; Third Row: Double Exponential model.

equal base rates. The relative frequency of this conditional probability O_A^* across trials in these theoretical examples is the relative frequency across trials of the effect *e* that generates this value of O_A^*. The heights of the *AOD* curve for each possible value of O_A^* are therefore obtained by averaging the two distributions (densities) at the value *e* (see Equation 2). For the distribution models chosen for these examples, each value of *e* gives rise to a unique value of O_A^* – this would usually, but not always, be true. The *UOD* curves are obtained in a similar way, except that in this case there are two different effects that give rise to the same probability correct value (see Footnote 2).

In order to interpret the shapes of these odd distributions, it is important to remember first that the point of the analysis is to show how often different objective betting odds occur in an equal base-rate discrimination condition. For the Normal Equal Variance model of the sensory effects (upper row of Figure 2), the *AOD* curve is symmetric and reaches its maximum at the point of maximum uncertainty (0.5). In other words, the most frequent trial in this hypothetical experiment is one in which the subject will have little or no information about the identity of the stimulus. If the spacing between the two normal distributions was increased, the curve would invert, becoming a U-shaped function instead of an inverted U-shaped function. That is, on most of the trials, the subject would have either strong information that the stimulus was an *A*, or strong information that the stimulus was a *B*. Since the *AOD* was symmetric, the *UOD* curve for this example provides essentially the same information: The most frequent events are low-certainty trials and the least frequent events are high-certainty trials.

The potential benefits of the extra information contained in the *AOD* curve becomes evident in the other two examples in the figure, representing the Normal Unequal Variance model (middle row) and the Double Exponential models (bottom row; see, e.g., Luce, 1986, Appendix B for the distribution formula). In the nor-

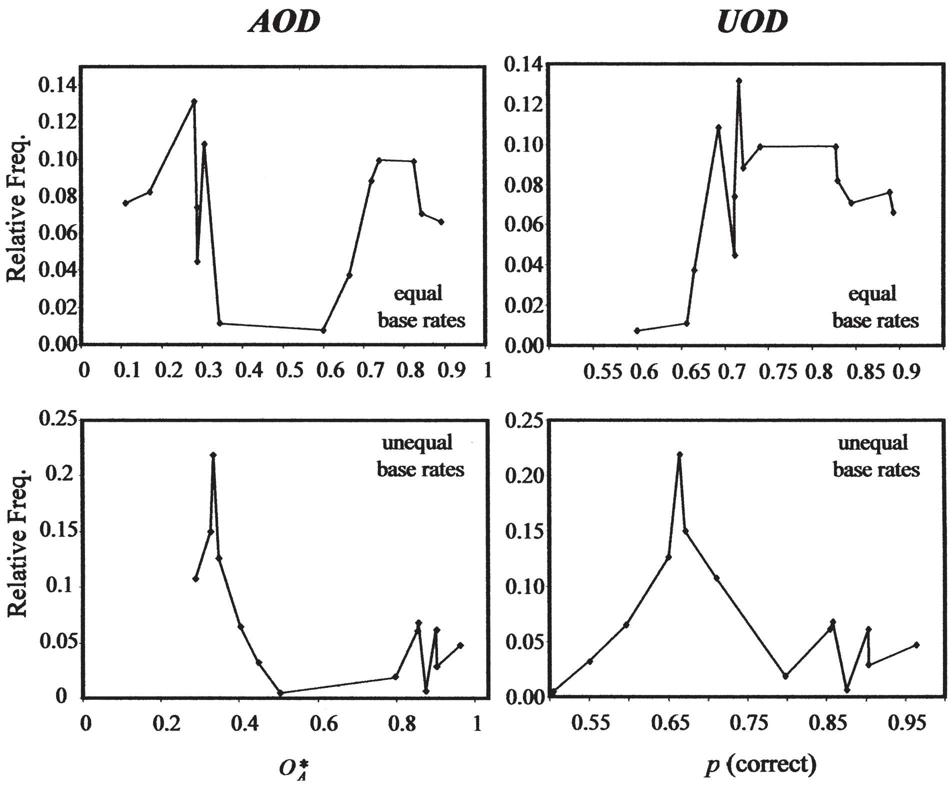

Figure 3. Empirical *AOD* and *UOD* functions from a visual size discrimination experiment reported in Balakrishnan (1998b). The results are qualitatively consistent with the Normal Unequal Variance model and the reflected Double Exponential model (change the sign of the argument).

mal model, the subject's most frequent experience will be high certainty that the stimulus was a *B*, since values on the abscissa less than one-half represent betting odds in favour of the *B* response in the *AOD*. In the Double Exponential model, the subject will most often experience high certainty that the stimulus was an *A*. Since the base rates are equal, these effects must be due to properties of the sensor, and/or to biases in the sampling or stopping rules adopted by the subject.

Hopefully it is clear from these examples that both the *AOD* and the *UOD* curves can be used to quantify the difficulty of a task in concrete terms, without learning some pattern-matching rules (e.g., large area under the ROC implies that the task is easy). The odds distributions also make it possible to show how two tasks that produce roughly equivalent overall percent-correct scores may still differ in important ways: Two odds distributions could have the same mean but very different shapes. In this respect, ROC curves are more difficult to interpret: Apart from the total area underneath them, differences in the shapes of two ROC curves do not provide any transparent information about the difficulty of the task.

The Shape of the Odds Distribution and Its Implications

In addition to providing a more direct representation of task difficulty, the odds distribution may also be useful for investigators who wish to study the effects that biases of other kinds (i.e., in the sampling or stopping rule) may have on discrimination performance. To illustrate, consider once more the middle and lower row of panels in Figure 2, representing the Normal Unequal Variance model (middle row) and the Double Exponential models (bottom row). Both of these models generate an ROC curve that is skewed with respect to the negative diagonal (i.e., the line connecting the upper left corner to the lower right corner of the graph), positive skew for the normal model, and negative skew for the Double Exponential model. The direction of the skew is unimportant, since it can be reversed in either model, that is, by switching the variances in the normal model or by reflecting the distributions (changing the sign of the argument) in the Double Exponential model.

The theoretical significance of this qualitative property of the curves is not the difference in variances of the distributions (the two variances are equal in the Double Exponential model), but instead the asymmetry in their associated odds distributions. In an important sense (but *not* in the decision-rule sense of detection theory), the asymmetric curves in the Figure may represent a bias toward the *A* (middle panels) or the *B* stimulus (lower panels): At the point of the decision, the decision-maker more often has information favouring one of the two judgments over the other, despite the fact that the base rates are controlled for. Although difficult to explain in the context of detection theory, it turns out that dynamic models can account for precisely this effect in a very natural way (i.e., by introducing biases in the stopping rule; see Van Zandt, 2000).

Empirical AOD Curves

When the base rates are unequal in visual size discrimination tasks with rating responses, the sufficient condition for suboptimality of the decision rule defined earlier turns out to be satisfied (Balakrishnan, 1998b, 1999), but the bias test always fails. Although the failure of the bias test does not rule out the possibility that the decision rule was biased, the fact that the same measure (confidence) detected the suboptimality of the decision rule makes it difficult to reconcile these data with the detection theory assumption that human subjects vary their decision criteria when the base rates are manipulated in an experiment. We suggested therefore that the decision rule was unbiased, and that other kinds of bias (in the stopping or sampling rule) cause the hit and false-alarm rates to covary under changes in the base rates.

This sensory bias should be expected to show up in the odds distributions. Some examples of these curves taken from the data reported in Balakrishnan (1998b) are shown in Figure 3. In these size discrimination experiments, the confidence ratings were given on a bipolar scale, that is, small rating values were high confidence *A* ("small") responses; large values were high confidence *B* ("large"). When the base rates were equal, the *AOD* curves were roughly symmetric. However, when the *A* stimulus was presented more often than the *B* stimulus (unequal base rates), the relative frequencies of the states constituting high evidence in favour of the "large" response (*B*) are higher than the frequencies for the "small" response (*A*), even though the odds analysis takes the base rates into account. In other words, these data exhibit a sensory bias effect, consistent with our conclusion that signal detection theory's concept of a shifting decision criterion representation of decision making biases is fundamentally violated empirically.

Performance Measures

Given the quantities they represent – that is, the distributions of the betting odds across trials when the base rates are equal – it should not be surprising to find that the overall percent correct score across *A* and *B* trials (under equal base rates) is easily derived from the odds distributions. In fact, this value is simply the mean of the *UOD* function. This "direct" measure of yes-no detection performance is our alternative to the area measure. We will refer to it as the "odds statistic" and denote it by γ_O. Since it is equal to the mean of the *UOD* function, it can be obtained by multiplying the values on the abscissa of this function by their corresponding heights and adding these terms. However, an even simpler formula, derived in the Appendix, is

$$\gamma_O = \frac{1}{2}\sum_k \max\big(f(R = k|S = A), \quad f(R = k|S = B)\big), \tag{3}$$

where the summation runs over all possible rating responses (i.e., on both the *A* and *B* sides of a bipolar scale).

In the classical yes-no detection experiment, there are only two responses, which makes it formally equivalent to a confidence-rating experiment with one level of confidence per response. The estimate of γ_O in this case is therefore

$$\gamma_O = \frac{1}{2}[\max\big(f(R = A|S = A), f(R = A|S = B)\big) + \max\big(f(R = B|S = A), f(R = B|S = B)\big)].$$

Notice that the four conditional probabilities in this expression are simply the four entries of the 2 x 2 contingency table (i.e., the hit, false-alarm, miss, and correct rejection rates). If the hit rate is greater than the false-alarm rate and the correct rejection rate is greater than the miss rate, as is almost always the case, this formula is simply the average of the hit and correct rejection rates. This is also the exact same formula used to estimate area under the ROC curve from a single point (i.e., the "trapezoidal area" obtained by connecting the points of the curve and computing the area underneath the line segments; see Smith, 1995). Thus, when derived from a single point on an ROC curve, the area measure is also a direct estimate of the subject's ability to perform the yes-no detection task.

Of course, since there will almost surely be many more than two possible effects of the stimuli (i.e., the observable *A* and *B* responses will represent more than two possible sensory effects), the hit and false-alarm rates are weighted averages over a large set of objective probability values, and hence an estimate derived from these two values alone would be relatively crude (which is one reason why detection theorists often recommend using parametric indices rather than the area measure when these are the only data available; see, e.g., Macmillan & Creelman, 1991). Although it is possible to run multiple base-rate or payoff conditions to estimate γ_O from the conditional probabilities associated with smaller groups of information states, as is sometimes done to estimate an ROC curve, we will not illustrate this analysis because from our point of view it is inappropriate. It is important to recognize that this approach introduces the assumption that the distributions of the sensory effects are invariant under different biasing conditions (i.e., that there is a single ROC curve to be estimated and not a different one for each base-rate or payoff condition). Empirical data unequivocally reject this assumption, at least for visual size discrimination: Changing the base rates dramatically changes the shapes of the underlying distributions, changing therefore the shape of the rating ROC curve (Balakrishnan, 1998b, 1999; Van Zandt, 2000).

A better solution is to observe other aspects of the subject's discrimination behaviour that might make it possible to identify smaller sets of the conditional probabilities associated with the different sensory effects. Although the confidence rating paradigm is the most popular approach along these lines (e.g., Baranski & Petrusic, 1998; Egan, Clarke, & Carterette, 1956; Green & Swets, 1974; Swets, Tanner, & Birdsall, 1961), another equally reasonable approach is to record response time, as we noted earlier, and use these data to break up the *A* and *B* judgments into arbitrarily small subsets. It is well known that accuracy depends strongly on reported confidence, and further that response time (RT) is correlated (negatively) with confidence (Baranski & Petrusic, 1998; Emmerich, Gray, James, Watson, and Tanis, 1972; Katz, 1970; Petrusic & Baranski, 1997; Shaw, McClure, & Wilkens, 2001; Vickers, Smith, Burt, & Brown, 1985; see Link, 1992 for a review of earlier work). Response time should therefore substitute quite well for a confidence rating response, eliminating the need for a confidence rating judgment. Without changing the traditional detection paradigm in any way, therefore, it is possible to obtain a performance assessment based on distributional information. At least for purposes of ordering experimental conditions with respect to sensitivity differences, performance estimates derived from distributional analyses appear to be superior to estimates based on hit and false-alarm rates alone (Balakrishnan, 1998a).

Statistical Properties of the Area and Odds Measures

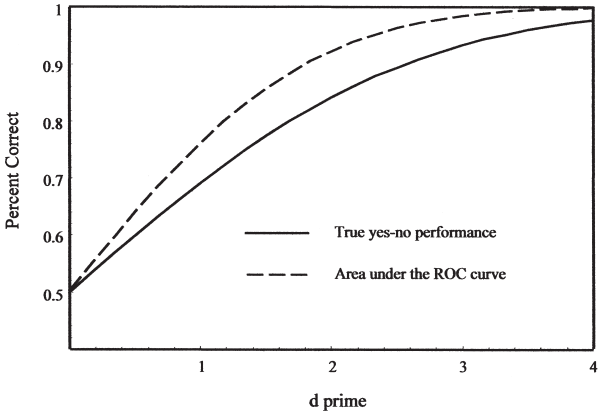

Figure 4. Comparison of the area under the ROC curve to performance in the yes-no detection task under the normal equal variance model. When estimated accurately, the area measure overestimates the true proportion of correct discrimination judgments.

Calculated from the ratings ROC curve, the estimated area under the ROC curve (trapezoidal area) can be written as

$$A_{ROC} = \sum_{i=1}^{n}\left(\frac{f_A(i)\cdot f_B(i)}{2} + F_A(i-1)\cdot f_B(i)\right)$$

where *i* is the rating response on an *n*-point bipolar scale, *f* denotes the relative frequency, and *F* is the cumulative frequency distribution (and hence $F_A(i\text{-}1) = 0$ when $i = 1$). The first component in the sum,

$$\frac{f_A(i)\cdot f_B(i)}{2},$$

is the area of a triangle, and the second, $F_A(i - 1)\cdot f_B(i)$, is the area of a rectangle (to see why, connect the points of the empirical ROC curve and consider the area to right of the line segment defined by each adjacent pair of points).

Since both estimators, γ_O and A_{ROC}, involve sums of the same components, they will have similar statistical properties. However, because the area measure involves products of these components and more terms, the variance in the γ_O estimator should be expected to be smaller. In fact, in simulations of each of the three distribution models shown in Figure 2, the variance of the odds estimator was slightly smaller for all values of the spacing between means of the distributions, distribution variances, and criteria placement that we examined. There is no obvious reason, therefore, to prefer the area measure in empirical applications, even if the purpose of the analysis is merely to order conditions by their sensitivity.

Of course, the main advantage of the yes-no detection measure γ_O over the area measure is the fact that γ_O is a measure of percent correct in the yes-no detection task rather than the 2AFC task. To see the difference, Figure 4 compares the true yes-no performance to the corresponding area under the true ROC curve for the Normal Equal Variance model. Notice that the area under the ROC curve consistently overestimates the performance of the subject in the yes-no detection task, with the greatest disparity occurring at moderate values of sensitivity. Ironically, modifications of the trapezoidal area estimate to take account of the convexity of a typical continuous ROC curve (e.g., Grier, 1971; Nelson, 1984; Smith, 1995) necessarily increase the estimator, presumably causing it to be further away from the correct yes-no detection score.

n – Choice Classification Tasks

Compared to the phenomenal success of the detection theory approach to discrimination and detection behaviour, applications of the theory to other classification designs have been relatively modest. For a variety of reasons, there are no widely accepted formulas available to compute either parametric or nonparametric performance statistics for identification data (but see Scurfield, 1998; Smith, Warren, Dutton, & Smith, 1996, for attempts along these lines). Instead, a parameter search routine is typically used to attempt to find the best-fitting parameters of a model under one of several possible sets of model assumptions (e.g., Ashby & Lee, 1991; Balakrishnan, 1997; Kadlec, 1995; Kadlec & Hicks, 1998). This model-fitting problem is not at all trivial, especially when the perceptual space is multidimensional. Even if it is reasonable to believe that the model is sufficiently accurate, in many cases it is difficult to be convinced that these fitting routines will consistently identify the correct model parameters.

Extension of the odds statistic, on the other hand, is straightforward and unambiguous. When there are more than two stimuli, the optimal decision-maker still follows the same fundamental principle, selecting the response that maximizes the probability of a correct response. To do this, the decision-maker must compute the conditional probability of each stimulus category *i* given the effect of the stimulus, $p(S = i \mid E = e)$, then choose the response accordingly (i.e., the response that maximizes the chances of a correct classification response). As in the two-choice case, the difficulty of the task is a function of the (univariate) distribution of the maximum of these *n* conditional probabilities across trials. If this maximum value is almost always close to 1 across trials, the task is easy; if it is almost always close to its smallest possible value, $1/n$, the task is hard.

The most natural graphical representation for the *n*-choice paradigms would be a plot of the distribution of this maximum of $p(S = i \mid E = e)$ across trials when the base rates are equal (i.e., $p_i = 1/n$). This would be the *n*-choice extension of the unarticulated odds distribution for experiments with more than two stimuli. The overall probability of a correct response under an optimal decision rule is once again the weighted average of the maximum of the observable conditional probabilities, which simplifies to

$$\gamma_{O,n} \quad = \quad \frac{1}{n}\sum_{k} \max_{i=1,\ldots,n} [p(R = k \mid S = i)]\,, \qquad (4)$$

where *k* runs over all the possible responses and *n* is the number of classification judgments. For example, in a confidence-rating task in which the subject first reports which of four stimuli was presented and then gives a rating on a 5-point scale, there are 4 x 5 = 20 possible responses.

In principle, this classification index can be compared across conditions with arbitrary sets of arbitrarily complex stimuli. However, if there are different numbers of stimuli in the two conditions to be compared, it is important to recognize that $\gamma_{O,n}$ may be higher in one condition merely because there are fewer stimuli to be discriminated. As before, the subjects' classification responses define the crudest possible partition of the information states induced by the stimuli, and other properties such as confidence and RT may provide the most effective means of subdividing them further.

Conclusions

Despite the many alternative approaches developed over the past half century, detection theory continues to be the preferred point of departure for most formal analyses of human discrimination behaviour. There are probably many good reasons for this remarkable record, including some related to the fundamental concepts first introduced to psychologists by the detection theorists. However, it is also important to recognize that detection theory is not a finished product – there is still substantial room for development, even with respect to its most elementary applications. The area measure is one example of this potential. By applying "first principles" of probability and decision theory, it is possible to derive a measure of yes-no detection that is comparable in every respect to the area measure except for the need to convert to a different paradigm, 2AFC. This new measure controls for suboptimality of the decision rule (whereas the area measure does not), appears to have a smaller standard error, and is easier to compute than the area measure.

Like the area measure, this yes-no statistic should be understood as a measure of the amount of information available at the point at which a decision is reached (and executed) rather than a measure of sensitivity that controls for decision-making biases – because there are other kinds of biases that this measure does not control for. Its accuracy in this properly circumscribed role will depend on how it is estimated empirically. In particular, its value should approach the true amount of information available as smaller groupings of the observer's information states are identified by the experimenter – that is, as more and more aspects of the subjects' discrimination judgment are recorded. Depending on the circumstances, the discrimination or identification judgment by itself is probably too crude a partition of these states to justify this statistic in most cases (although it may be difficult to justify any of the other measures in these cases as well). Obtaining more information by adding a confidence-rating procedure is easy enough to do, but even this extra effort may be unnecessary, since response times are always distributed and usually will be strongly correlated with the amount of information available at the point of the decision. In fact, since RT is a continuous measure, the investigator is free to choose the RT intervals to subdivide the subjects' *A* and *B* responses in such a way that the sample sizes are, say, constant or greater than some minimum for each grouping interval. Computing the probability of a correct *A* (or *B*) judgment for each of these RT subsets should produce an odds distribution that works at least as well as the distribution obtained from rating data.

Of course, if there is sufficient data available, the experimenter may also subdivide the subjects' responses by both confidence and RT, obtaining an even finer partition, or using any other measure that is correlated with accuracy. If this correlation is poor, the measure provides little or no more information than that contained in the discrimination judgments already. Having an accurate measure of the information available at the point of the decision, it may then be possible to examine the role that other kinds of biases may play, such as speed/accuracy trade-off (the stopping rule) and attention biases (the sampling rule). Simply assuming that these biases do not exist, as is often done, probably does a disservice to the field, despite the rigour and sophistication of the detection theory statistics.

This work was supported by NASA grant NRA2-37143 (Intelligent Systems) and is dedicated to the memory of Dr. Hamid Kohen, a free-thinking scientist and a great friend. We are grateful to Helena Kadlec and Ehtibar Dzhafarov for many helpful comments on an earlier draft.

Address correspondence to J. D. Balakrishnan, Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907 (E-mail: jdb@psych.purdue.edu).

References

Ashby, F. G., & Lee, W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General*, *120*, 150-172.

Balakrishnan, J. D. (1997). Form and objective of the decision rule in absolute identification. *Perception & Psychophysics*, *59*, 1049-1058.

Balakrishnan, J. D. (1998a). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors*, *40*, 601-623.

Balakrishnan, J. D. (1998b). Some more sensitive measures of sensitivity and response bias. *Psychological Methods*, *3*, 68-90.

Balakrishnan, J. D. (1999). Decision processes in discrimination: Fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1-18.

Balakrishnan, J. D., & MacDonald, J. A. (2001a). Alternatives to signal detection theory. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors: Vol. 1* (pp. 546-550). London: Taylor & Francis.

Balakrishnan, J. D., & MacDonald, J. A. (2001b). Observability of suboptimal and biased decision rules in classification paradigms. *Purdue Mathematical Psychology Technical Report*. (Submitted to the Journal of Mathematical Psychology.)

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 929-945.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic cognition approach to decision making. *Psychological Review*, *100*, 432-459.

Cox, D. R., & Hinkley, D.V. (1990). *Theoretical statistics*. London: Chapman and Hall.

Diederich, A. (1997). Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, *41*, 260-274.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, *2*, 312-329.

Egan, J. P., Clarke, F. R., & Carterette, E. C. (1956). On the transmission and confirmation of messages in noise. *Journal of the Acoustical Society of America*, *28*, 536-550.

Emmerich, D. S., Gray, J. L., Watson, C. S., & Tanis, D. C. (1972). Response latency, confidence, and rocs in auditory signal detection. *Perception & Psychophysics*, *11*, 65-72.

Green, D. M. (1964). General prediction relating yes-no and forced-choice results. *Journal of the Acoustical Society of America, A*, *36*, 1024.

Green, D. M., & Swets, J.A. (1974). *Signal detection theory and psychophysics* (Reprint). Huntington, NY: Krieger. (Earlier edition published 1966, New York: Wiley).

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, *75*, 424-429.

Irwin, R. J., Hautus, M. J., & Butcher, J. C. (1999). An area theorem for the *same-different* experiment. *Perception & Psychophysics*, *61*, 766-769.

Kadlec, H. (1995). Multidimensional signal detection analyses (MSDA) for testing separability and independence: A Pascal program. *Behavior Research Methods, Instruments, & Computers*. *27,* 442-458.

Kadlec, H., & Hicks, C. L. (1998). Invariance of perceptual spaces and perceptual separability of stimulus dimensions. *Journal of Experimental Psychology: Human Perception & Performance*, *24,* 80-104.

Katz, L. (1970). A comparison of Type II operating characteristics derived from confidence ratings and from latencies. *Perception & Psychophysics*, *8*, 65-68.

Link, S. W. (1992). *The wave theory of difference and similarity*. Hillsdale, NJ: LEA.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*, 77-105.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

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

Nelson, T. O. (1984). A comparison of current measures of accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109-133.

Petrusic, W. M., & Baranski, J. V. (1997). Context, feedback, and the calibration and resolution of confidence in per ceptual judgments. *American Journal of Psychology*, *110*, 543-572.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*. *26*, 127-140.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261-300.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks. *Journal of Mathematical Psychology*, *42*, 5-31.

Shaw, J. S., McClure, K. A., Wilkens, C. E. (2001). Recognition instructions and recognition practice can alter the confidence-response time relationship. *Journal of Applied Psychology*, *86*, 93-103.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology, 44,* 408-463.

Smith, W. D. (1995). Clarification of sensitivity measure *A*. *Journal of Mathematical Psychology*, *39*, 82-89.

Smith, W. D., Dutton, R. C., & Smith, N. T. (1996). A measure of association for assessing prediction accuracy that is a generalization of non-parametric ROC area. *Statistics in Medicine, 15,* 1199-1215.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25,* 251-260.

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

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge, UK: Cambridge University Press.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582-600.

Vickers, D., Smith, P., Burt, J., & Brown, M. (1985). Experimental paradigms emphasizing state or process limitations: II. Effects on confidence. *Acta Psychologica*, *59*, 163-193.

Appendix

The mean of the *UOD* curve is the maximum percent-correct score that the subject could achieve if the base rates were equal and the two distributions of sensory effects (*A* and *B* trials) induced by the actual experimental conditions (with unequal base rates) would stay the same if the base rates were equal. The simpler formula (Equation 3) for this mean value is derived as follows.

First, regardless of the base rates in the actual experiment, each observable response *R* (e.g., response "*A*, 3" in the ratings paradigm) will have some relative frequency on *A* trials, $f(R \mid S = A)$ and another relative frequency on *B* trials, $f(R \mid S = B)$. These two values can be used to determine what the possible bettings odds values will be when the base rates are equal (i.e.,

$$O_A^* = p(S = A \mid R) = \frac{f(R \mid S = A)}{f(R \mid S = A) + f(R \mid S = B)},$$

and also the relative frequency with which these betting odds values will occur,

$$f(O_A^* \mid p_A = p_B) = \tfrac{1}{2}\,[f(R \mid S = A) + f(R \mid S = B)].$$

The overall probability of a correct response when the base rates are equal will be the average of the different betting odds values (the maximum of O_A^* and 1 - O_A^*), weighted by the relative frequency with which they will occur. For example, if the maximum of O_A^* and 1 - O_A^* will be 0.7 on half the trials and 0.8 on the other half of the trials, overall percent correct will be 0.75.

Therefore, for each possible response, $R = r$, we need to find the maximum of the corresponding value of O_A^* and 1 - O_A^*, or in other words, the maximum of the two values,

$$\frac{f(R=r \mid S = A)}{f(R=r \mid S = A)+f(R=r \mid S = B)},$$

and

$$\frac{f(R=r \mid S = B)}{f(R=r \mid S = A)+f(R=r \mid S = B)}.$$

We then multiply this maximum value by the relative frequency with which it will occur (i.e., the relative frequency with which the value O_A^* will occur),

$$\frac{1}{2}[f(R=r \mid S=A) + f(R=r \mid S=B)] \cdot \max\left(\frac{f(R=r \mid S = B)}{f(R=r \mid S = A)+f(R=r \mid S = B)}, \frac{f(R=r \mid S = B)}{f(R=r \mid S = A)+f(R=r \mid S = B)}\right).$$

Notice, however, that regardless of which of the two betting odds values in the right term is the maximum, the denominator will cancel with the term on the left. The expression therefore simplifies to

$$\frac{1}{2}\max\big(f(R=r \mid S = A), f(R=r \mid S = B)\big).$$

Adding up these values for each possible response, $R = k$, yields the overall probability correct across trials,

$$\gamma_O = \frac{1}{2}\sum_k \max\big(f(R = k \mid S = A), f(R = k \mid S = B)\big).$$

Sommaire

Le théorème d'aire bien connu de Green établit une équivalence entre l'aire sous la courbe ROC oui-non et le pourcentage d'exactitude d'un observateur impartial dans une tâche à choix forcé à deux options (2AFC) avec un stimulus équivalent. En raison de cette relation simple, cette soi-disant mesure d'aire est devenue l'un des plus importants indices de performance de discrimination. Dans le présent article, nous montrons que cette conversion des données de détection oui-non à la performance hypothétique dans une tâche 2AFC n'est pas nécessaire. Les mêmes données de détection oui-non qui sont utilisées pour calculer les statistiques d'aire peuvent toujours être utilisées pour calculer le pourcentage d'exactitude d'un observateur impartial dans la tâche de détection oui-non proprement dite. Cette mesure de « probabilités (odds) » et la mesure de l'aire devraient toutes deux être prises en tant qu'indices de la quantité d'information disponible au point où le sujet donne sa réponse à chaque essai. À ce titre, ces mesures contrôlent les biais et les sous-optimisations d'un genre particulier, p. ex. dans une règle de décision. D'autres sortes de biais comme ceux supposant la façon qu'un sujet réagit au stimulus (biais de codage) et le temps qu'il prend à réagir (biais de la règle d'arrêt) pourraient avoir une incidence importante sur la mesure de l'aire et la mesure que nous définissons ici.

Nous montrons aussi que la courbe ROC pourrait ne pas être le moyen graphique idéal pour les chercheurs empiriques dont le but est souvent d'étudier les effets des stimuli sur l'observateur plutôt que d'apparier les taux d'avertissement de réponses bonnes et fausses. Dans toute tâche de discrimination, un essai peut être vu comme une gageure avec différentes probabilités de gageures postérieures selon l'information que le sujet a glané du stimulus. Lorsque les valeurs de probabilité sont uniformément fortes en faveur de l'un ou l'autre stimulus, la tâche est facile; lorsqu'elles sont uniformément faibles, la tâche est difficile. À partir de la distribution de ces probabilités de gageures observées en fonction de n'importe quelle condition de taux de base dans une expérience, il est possible de calculer les valeurs de probabilité et leur distribution en vertu des conditions de taux de base égales. Un tracé des données distribuées « corrigées » peut présenter une description plus naturelle et plus informative de la capacité de l'observateur de faire une discrimination. En dernier lieu, contrairement à l'aire mesurée et d'autres données statistiques de théorie de détection traditionnelles, tant la mesure du pourcentage d'exactitude des réponses oui-non et la distribution des probabilités généralisent de façon évidente et directe les paradigmes de classification avec plus d'une réponse (c.-à-d. l'identification).