Journal of Electronic Imaging 10(2), 376–384 (April 2001).

Misrepresentations of signal detection theory and a model-free approach to human image classification

J. D. Balakrishnan
Justin A. MacDonald
Purdue University
Department of Psychological Sciences
West Lafayette, Indiana 47907
E-mail: jdb@psych.purdue.edu

Abstract. *Experimental methods and statistics derived from signal detection theory are frequently used to compare two imaging techniques, to predict human performance under different parameterizations of an imaging system, and to distinguish variables related to human visual perception from variables related to decision making. We review recent experimental results suggesting that the assumptions of signal detection theory are fundamentally unsound. Instead of shifting decision criteria under different priors, humans appear to alter the information assimilation process, representing images from categories with high prior probability more accurately (less variance) than images from categories with low prior probability. If this hypothesis is correct, detection theory measures such as d' and area under the receiver operating characteristic may be misleading or incomplete. We propose an alternative approach that can be used to quantify the effects of suboptimal decision making strategies without relying on a model of detection structure.* © 2001 SPIE and IS&T.
[DOI: 10.1117/1.1344188]

1 Introduction

A significant portion of the image quality issues encountered within the imaging sciences can be formally reduced to image classification problems of one kind or another. Some of the more conspicuous examples include medical imaging, product quality control, air traffic control, surveillance, and image database design. Relatively few of these image classification systems are automated from start to finish—usually humans are involved at some point. It is not surprising therefore that models of human image perception and classification often play a role in the development of imaging techniques. In this paper, we discuss the merits and limitations of one of the most popular of these, the theory of signal detection (or signal detection theory).[1] Using examples from radiology, we show how the detection theory models have been applied in imaging studies to measure or predict human performance. We then review some recent findings that raise some doubts about their basic validity and propose an alternative approach to measurement that does not rely on any assumptions about the structure of human detection behavior.

2 Statistical Decision Making in Medical Imaging

Consider the following prototypical issues in radiographic imaging: (i) A high resolution x-ray image can be acquired using standard film-screen techniques or using newer digital equipment, such as storage phosphor or selenium-detector based digital systems. Which imaging system conveys more relevant information about the patient's condition and what parameterizations of these systems are optimal?[2] (ii) Film-based images can be viewed on a lightbox and digital images on a cathode ray tube monitor (soft-copy viewing) or a film printer hardcopy. How does the viewing medium affect the image quality?[3] (iii) Computerized image enhancement, image reconstruction, and automated detection algorithms can provide additional information about image properties, but may also bias or interfere with the radiologist's decision making process in some way. How do these computer-aided systems compare to unassisted diagnosis?[4]

Several approaches to these measurement questions have been followed in the past, including visually comparing examples of two image formats, calculating statistical criteria (e.g., mean squared error, detective quantum efficiency, the peak signal-to-noise ratio), and recruiting radiologists for clinical evaluation. Clinical tests have the advantage of being objective, reproducible, and directly related to the image's intended function (i.e., diagnosis). Because there is no guarantee that radiologists can accurately predict their own performance with different imaging formats, the ideal clinical test is to simulate the diagnostic process so that performance can be compared under well-defined conditions. Often, diagnosis is reduced to a binary judgment, such as the presence/absence of an abnormality (possibly in each region of an image). On each trial of the study an image from one of these two different populations is randomly sampled from a collection.

Paper STA-04 received June 2, 2000; revised manuscript received Sep. 23, 2000; accepted Sep. 25, 2000.
1017-9909/2001/$15.00 © 2001 SPIE and IS&T.

Table 1 Hypothetical two-by-two contingency tables for discrimination under two different imaging conditions.

		Method A				Method B	
		Response				Response	
		AB	NR			AB	NR
Input	ab	0.6	0.4	Input	ab	0.7	0.3
	nr	0.3	0.7		nr	0.8	0.2

Signal detection theory was developed to analyze and interpret data from these two-choice classification (yes–no detection) experiments. The radiologist's judgments are converted to frequencies or relative frequencies in a two-by-two contingency table, as illustrated in Table 1. The rows of the table define the true condition, abnormal (ab) or normal (nr) and the columns define the judgment (AB or NR). The relative frequency of the AB judgment when an abnormality is present, freq[ab,AB]/(freq[ab,AB] +freq[ab,NR]), is the 'hit' or 'true positive' (tp) rate, and the relative frequency of this judgment when no abnormality exists, freq[nr,AB]/(freq[nr,AB]+freq[nr,NR]), is the 'false alarm' or 'false positive' (fp) rate. Because the proportions in each row of the table must add to 1, the other two cells (the "miss" or "false negative" and the "correct rejection" or "true negative") are redundant with the tp and fp rates and can therefore be ignored.

If imaging method A is thought to be superior to method B, the tp rate would usually be expected to be higher and the fp rate lower for this method. However, it is not uncommon to find that the tp and fp rates are both higher. Even if the increase in the tp rate is greater than the increase in the fp rate, it is not immediately obvious which method should be favored and it is also unclear what causes both the tp and fp rates to increase. In Table 1, the tp rate for method B (0.7) is only slightly larger than the tp rate for method A (0.6), but the fp rate is substantially larger (0.8 versus 0.3). Without additional information about the performance of these two diagnostic systems, it is impossible to determine which one would be superior in general practice.

Estimating parameters from one of the signal detection theory family of models makes it possible to compare two methods and to explain why the tp and fp rates may increase or decrease together. The models assume that the radiologist converts the information in the image to an "evidence value" on a bipolar continuum. Small (negative) values represent strong evidence for the normal (nr) and large values strong evidence for the abnormal (ab) condition. Somewhere between these two extremes is a cutoff between evidence states that will be mapped to AB responses and the states mapped to NR responses. Classification errors occur when the evidence falls on the wrong side of this decision boundary or criterion.

The relative frequencies of the outcomes, including the tp and fp rates, depend on the relative frequencies (the "base rates") of the nr and ab conditions in the experiment (i.e., an experimenter-controlled variable) and the probability distributions of the evidence states on nr and ab trials. When the distributions are assumed to be univariate Gaussian with equal variance, the distance between their means, or d', is a convenient index of diagnostic sensitivity or power. Less restrictive models are possible, as are other choices of the distribution functions.[1,5] The key idea behind all of these quantitative models is that the radiologist's behavior is analogous to a statistical hypothesis test: an information sample is collected from the image (the percept) and converted to a test statistic (the subjective likelihood ratio), which is mapped to one of the two diagnoses (the decision rule). Surprisingly, there is reason to challenge the model even at this very general level.

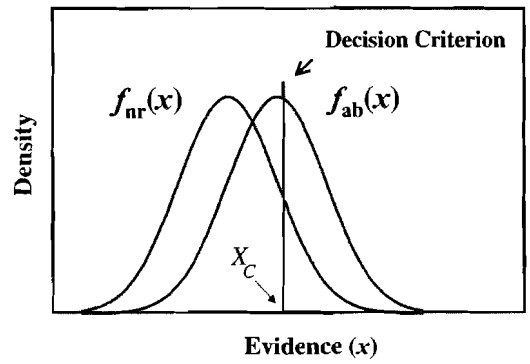


Fig. 1 The equal variance Gaussian model of binary classification. The perceptual information extracted from the stimulus is converted to an evidence value, whose distribution mean depends on the population, normal (nr) or abnormal (ab) from which the image was sampled. The decision process is a threshold (X_C) set on the evidence value.

2.1 *The Area Measure*

The equal variance Gaussian model is illustrated in Fig. 1. The criterion is displaced toward the side of the evidence scale favoring the AB response and the decision making strategy is therefore "biased" toward the NR judgment. The area under the ab distribution to the right of the criterion is the tp rate and the area under the nr distribution to the right of the criterion is the fp rate. These two response probabilities increase or decrease together as the decision criterion is shifted, presumably showing why the tp and fp rates are sometimes positively covarying in empirical studies, and why they depend on the base rates.

Calculating the tp and fp rates as the criterion is shifted across the range of evidence values and plotting one against the other produces the receiver operating characteristic (ROC) curve, which illustrates for a given subject the effect that different degrees of bias toward one of the two judgments will have on overall performance. The area under this curve is equal to the probability that a random sample from the ab distribution will be greater than a sample from the nr distribution, making it a nonparametric index of the overlap between the two distributions, and therefore an index of diagnosticity.[6]

2.2 *Validity of the Model*

The parametric index d' (defined as the distance between the means of two equal variance Gaussian distributions), the nonparametric area index, and other measures associated with signal detection theory all depend critically on the concept of an invariant encoding process with some fixed

noise level (sensitivity) and a variable detection criterion. As the decision maker's preference for a given judgment increases, for whatever reason, the set of evidence states mapped to this judgment increases, at the expense of the set mapped to the other judgment. The evidence accumulation (encoding) process, on the other hand, is assumed to be unaffected by these response preferences (sensitivity is unaffected by bias). Other kinds of decision making biases can be envisioned, and some of these could have substantial effects on the detection theory measures even when diagnostic power is constant. For example, biases may affect the amount of time that an observer spends assimilating information from an image or accessing relevant information from memory. Detection theory would generally mistake these kinds of biases for sensitivity effects.

In some decision making contexts the decision criteria are inherent and directly observable properties of the system. Automated detection systems, for example, typically convert the information from an image to a likelihood ratio test statistic that is thresholded at different values depending on a loss function or a desired fp rate. The computation of the likelihood statistic is unaffected by the choice of threshold and the decision process is a mapping of evidence values to judgments, as in detection theory.

When radiologists and other human experts make decisions about the information contained in images, there is no empirically observable evidence state and therefore no verifiable shift in a decision criterion. The popularity of signal detection theory is derived from indirect evidence for this effect, such as the positive covariance of tp and fp rates in laboratory studies when the base rates of the conditions are manipulated by the experimenter and the fact that sensitivity indices under these different biasing conditions are usually roughly constant.[7] If sensitivity and bias are not independent, it is not clear why 'pure' bias manipulations should have little or no effect on sensitivity measures.

3 Model-Independent Tests of Decision Rule Bias

Perhaps because experimental data seemed to confirm a model that was highly plausible to begin with, detection theorists appear to have overlooked a relatively straightforward, assumption-free empirical test of the criterion shift concept that follows directly from basic concepts of statistics and probability theory. To illustrate this, consider once more the two density functions, $f_{ab}(x)$ and $f_{nr}(x)$, in Fig. 1 and notice that when the detection criterion is shifted to the right of the evidence scale, there is a region of evidence values mapped to the NR response for which $f_{ab}(x) > f_{nr}(x)$. If the criterion is placed at the (single) point of intersection between the two distributions, then no such "biased response region" exists. More generally, the decision rule is unbiased if and only if

$$f_{nr}(x) \geq f_{ab}(x)$$

for all x mapped to the NR response, and

$$f_{ab}(x) \geq f_{nr}(x)$$

for all x mapped to the AB response.

Now consider once more the example in Table 1 and suppose that the proportions in the cells are the true response probabilities. For detection method B, the conditional probability of the NR judgment on nr trials, $p(R=\mathrm{NR}|\mathrm{nr})=0.2$, is less than the conditional probability of this judgment on ab trials, $p(R=\mathrm{NR}|\mathrm{ab})=0.3$. It is easy to show that if the decision rule is unbiased, then for any pair of distributions describing the evidence states,

$$p(R=\mathrm{NR}|\mathrm{nr}) > p(R=\mathrm{NR}|\mathrm{ab}),$$

and

$$p(R=\mathrm{AB}|\mathrm{ab}) > p(R=\mathrm{AB}|\mathrm{nr}),$$

Empirical results similar to those in Table 1 would therefore confirm unequivocally that the method B decision rule is biased.

To our knowledge, no such results have ever been reported in the human performance literature or in medical diagnosis, even though it is usually taken for granted that the decision rule is biased (the decision criterion is assumed to shift and a set of biased information states is therefore presumed to exist). In sensory discrimination experiments performed in our laboratory, this sufficient condition was never satisfied even when the base rate difference was relatively large (i.e., 9 to 1).[8,9] This fact by itself, however, would not be a sufficient reason to doubt the validity of detection theory. Even if the decision rule is strongly biased toward the NR response, the observable probabilities, $p(R=\mathrm{NR}|\mathrm{nr})$ and $p(R=\mathrm{NR}|\mathrm{ab})$, will still be composed of probabilities taken over both biased and unbiased evidence states. That is,

$$\begin{aligned} p(R=\mathrm{NR}|\mathrm{nr}) &= p(x<X_C|\mathrm{nr}) \\ &= p(x \in u_{\mathrm{NR}}|\mathrm{nr}) + p(x \in b_{\mathrm{NR}}|\mathrm{nr}) \end{aligned}$$

and

$$\begin{aligned} p(R=\mathrm{NR}|\mathrm{ab}) &= p(x<X_C|\mathrm{ab}) \\ &= p(x \in u_{\mathrm{NR}}|\mathrm{ab}) + p(x \in b_{\mathrm{NR}}|\mathrm{ab}), \end{aligned}$$

where X_C is the decision criterion, u_{NR} is the subset of evidence states mapped to the NR response for which $f_{nr}(x) \geq f_{ab}(x,)$ and b_{NR} is the subset of these NR response states for which $f_{nr}(x) \leq f_{ab}(x)$. If the decision rule is unbiased or biased only toward the AB judgment, then b_{NR} is an empty set and $(R=\mathrm{NR}|\mathrm{nr})$ must be greater than $p(R=\mathrm{NR}|\mathrm{ab})$. If the decision is biased toward the NR judgment, then $p(x \in b_{\mathrm{NR}}|\mathrm{nr})$ is nonzero and less than $p(x \in b_{\mathrm{NR}}|\mathrm{ab})$, but $p(x \in u_{\mathrm{nr}}|\mathrm{nr})$ will always be greater than $p(x \in u_{\mathrm{nr}}|\mathrm{ab})$. In order for the bias to be detected from these statistics, the biased response region b_{NR} would have to be quite large.

To avoid this averaging-out problem, the decision maker's judgments need to be broken down into small enough subsets of evidence states to detect the biased region near the criterion. The sufficient condition for a biased decision

rule can then be applied to each of these. That is, if the decision rule is unbiased, then for any subset of evidence states v mapped to an NR response

$$p(x \in v|\text{nr}) \geq p(x \in v|\text{ab}),$$

and for any subset w mapped to an AB response,

$$p(x \in w|\text{ab}) \geq p(x \in w|\text{nr}).$$

The only question is how to divide the subject's classification judgments into sufficiently small subsets of these (ordered) evidence states.

Fortunately, signal detection theory tells us not only what this decomposition rule should be, but also which of these subgroups should satisfy the condition for bias when a decision rule bias exists. According to the theory, the evidence axis is a bipolar continuum of subjective likelihood ratios, ranging from high confidence NR judgments to high confidence AB judgments. The criterion is set at the point of "indifference," or "least subjective certainty," in the accuracy of the judgment (or zero expected gain). In other words, the judgments given with low confidence are evidence states near the decision criterion and judgments given with high confidence are evidence states far from the criterion. In Fig. 1, the biased response region is immediately to the left of the criterion. Therefore, sufficiently low confidence NR judgments should satisfy the test for bias. All other states, including low confidence AB judgments, should fail this test.

4 Experimental Design Issues

In principle, asking subjects to give an estimate of the probability that their judgment will be correct (i.e., a subjective probability judgment on a continuous scale from 0 to 1) or an integer rating response on a scale with many levels should be enough to partition their classification judgments into sufficiently small sets of evidence states. In practice, however, subjects could simply ignore many levels of the probability scale or use large intervals to define some rating responses and small intervals to define others. This compliance issue is illustrated in Fig. 2. In addition to the critical decision criterion, other criteria are placed on the evidence scale to map internal confidence states to physically executable rating responses and the spacing between these criteria varies. Each response bin defines one of eight rating responses on a bipolar scale. Response $R=4$ is the lowest confidence NR response and response $R=5$ is the lowest confidence AB response.

When the spacing between two adjacent criteria, C_{j-1} and C_j, is large, the empirical likelihood ratio for the corresponding rating response, $p(R=j|\text{nr})/p(R=j|\text{ab})$, may not detect a biased response region contained within the interval. In Fig. 2, the spacing between the decision criterion and the criterion immediately to its left (i.e., the $R=4$, or "lowest confidence NR," response bin) is large, causing the ratio $p(R=4|\text{nr})/p(R=4|\text{ab})$ to be less than 1 despite the bias toward the NR response. When the response bin for rating response j is large, however, the relative frequency of response j is also large. Thus, subjects can be instructed to be conservative about using the lowest

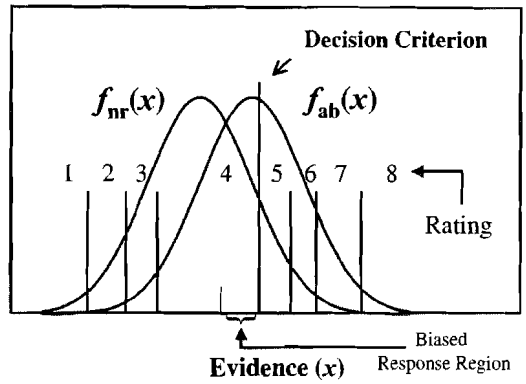


Fig. 2 Illustration of the effects of criteria placement on the test for bias in the decision rule. Internal confidence is mapped to rating responses by dividing the subjective confidence scale into response bins using additional criteria. Large spacing between the third and fourth criterion (the decision criterion) causes the proportion of "four" responses (lowest confidence NR) to be greater on nr trials, even though the ab density function is higher than the nr density function in part of this response bin. In such a case, the total proportion of four responses is also large (larger than the proportion of trials on which the evidence value falls in the biased response region).

confidence judgments (i.e., to use these responses only when they are extremely uncertain), in effect encouraging them to use small spacing of criteria adjacent to the decision criterion. If the instruction fails, the test is weak, but the weakness will be evident in the observed proportion of lowest confidence judgments (which will be large). In fact, because the criterion shift construct also predicts that the biased region, when it exists, will lie immediately to the left or right of the detection criterion, depending on the direction of the bias (see Fig. 1), the proportion of the lowest confidence NR responses is an upper bound on the proportion of biased NR responses, and the proportion of lowest confidence AB responses is an upper bound on the proportion of biased AB responses. If the test for bias fails for all rating responses and these two proportions are small, the bias must be small as well.

5 Experimental Results

Applications of these empirical tests to data from various kinds of perceptual discrimination tasks consistently lead to the same conclusion: the decision rule is either unbiased or is biased to a trivially small degree, even under strong bias manipulations (e.g., a base rate ratio of 9 to 1). Representative results from a visual shape discrimination experiment are shown in Fig. 3.[8] Subjects were asked to discriminate two L-shaped figures varying in size and to indicate how confident they were on a 14-point bipolar scale (seven levels of confidence for a given response). The test for bias failed for all confidence judgments, even when the base rates were unequal, and the proportion of lowest confidence responses (the upper bound on the proportion of biased responses) was very small (less than 1%).

5.1 *Suboptimality of the Decision Rule*

Using the same experimental methods—i.e., the rating procedure with special instructions about the relative frequency of extreme rating responses—it is also possible to

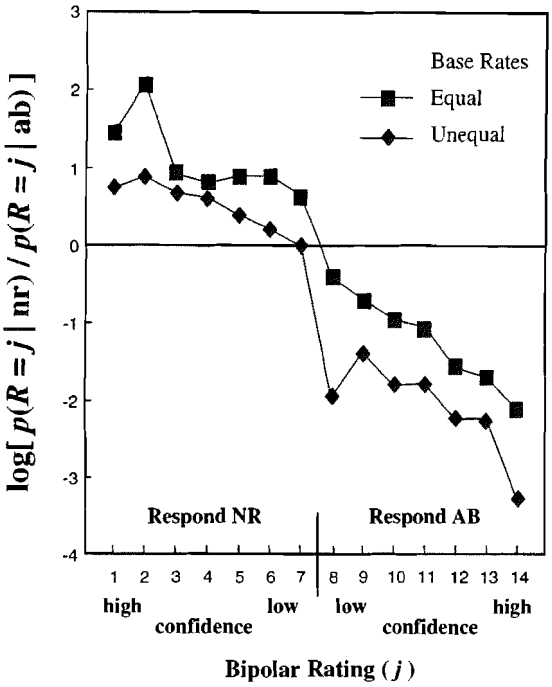


Fig. 3 The log likelihood ratio, $\log[p(R=j|\mathrm{nr})/p(R=j|\mathrm{ab})]$, for each rating response in the equal and unequal ($p_{nr}=0.9$) base rate conditions of a visual size discrimination task with a 14-point bipolar confidence rating response procedure. Rating responses 7 and 8 were the lowest confidence NR and lowest confidence AB judgments, respectively. The functions are positive when $j \leq 7$ (NR judgments) and negative when $j \geq 8$ (AB judgments), suggesting that the decision rule is unbiased.

test for suboptimality of the decision rule, where the term optimal is defined with respect to the probability of a correct diagnosis or with respect to any other objective loss function. Consider once more the detection model in Fig. 1 and suppose that the base rate of the nr condition is five times greater than the base rate of the ab condition, $p(\mathrm{nr}) = 5p(\mathrm{ab}) = 5/6$. The placement of the criterion in the example is biased toward the NR judgment, but not enough to maximize the probability of a correct judgment under these base rate conditions. The optimal location occurs at the point where the ratio of the ab density function to the nr density function is equal to the ratio of the base rates, or 5 to 1. The ratio at the decision criterion in the example is roughly 2 to 1.

This "conservative" displacement of the criterion is supposedly a strong regularity of human decision making. Presumably because they tend to underestimate or underweight base rate differences or payoff asymmetries, subjects do not shift their criteria enough to maximize their performance. The decision rule in Fig. 1 is suboptimal for accuracy because of the mapping rule immediately to the right of the detection criterion. For these evidence states, the posterior probability of the nr condition

$$p(\mathrm{nr}|x) = \frac{f_{\mathrm{nr}}(x)p(\mathrm{nr})}{f_{\mathrm{nr}}(x)p(\mathrm{nr}) + f_{\mathrm{ab}}(x)p(\mathrm{ab})}$$

is higher than the posterior probability of the ab condition

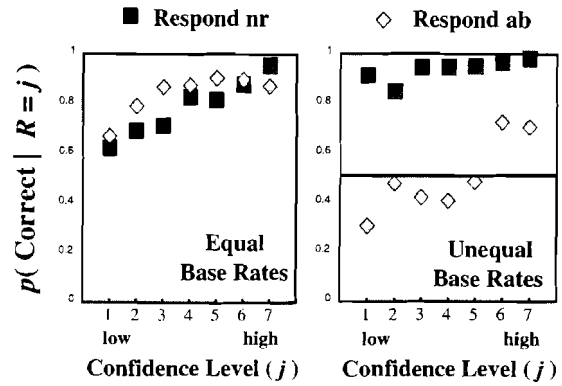


Fig. 4 Percent correct conditioned on the degree of confidence (seven levels) in the size discrimination experiment for the equal and unequal ($p_{nr}=0.9$) base rate conditions. Values less than 0.5 (left panel, low confidence AB judgments) indicate that the decision rule is suboptimal.

$$p(\mathrm{ab}|x) = \frac{f_{\mathrm{ab}}(x)p(\mathrm{ab})}{f_{\mathrm{nr}}(x)p(\mathrm{nr}) + f_{\mathrm{ab}}(x)p(\mathrm{ab})} = 1 - p(\mathrm{nr}|x).$$

Since $p(\mathrm{nr}|x) + p(\mathrm{ab}|x) = 1$, the inequality $p(\mathrm{nr}|x) < p(\mathrm{ab}|x)$ implies that $p(\mathrm{nr}|x)$ must be less than one half. A probability correct score less than one half for any confidence rating therefore immediately implies that the subject's decision rule is suboptimal for accuracy.

Assuming that human observers used unbiased decision rules in the perceptual discrimination experiment cited earlier, as the results of the bias test suggested, this test for suboptimality should be satisfied for low confidence *B* judgments when the base rates were unequal. Percent correct scores for each confidence rating response are shown in Fig. 4. Notice that in the unequal base rate condition, several of the lower confidence *B* judgments were in fact incorrect more often than they were correct, indicating that the decision rule was suboptimal. This result is important because it shows that rating data can be used to identify properties of the decision rule on one side of the decision criterion (in this case, on the right). It is difficult to maintain, therefore, that the criterion shift concept of detection theory is valid but that confidence rating data are somehow contaminated or for some other reason uninformative about decision making biases.

5.2 *The Adaptive Filter Model*

The rating paradigm was originally developed by Swets to estimate ROC curves using data from a single base rate condition.[1] Assuming that rating judgments are monotonically related to internal confidence, as in the Fig. 2 model, the cumulative distribution functions of the rating judgments on nr and ab trials are equal to the cumulative distributions of the evidence states at the upper bounds of the rating response bins (the C_j). Therefore, plotting the survivor function (one minus the cumulative distribution function) for ab trials against the survivor function for nr trials yields a rating ROC curve. Examples from the equal and unequal base rate conditions of the size discrimination experiment are shown in Fig. 5. Because the two conditions differed only in the base rates of the stimuli, signal detec-

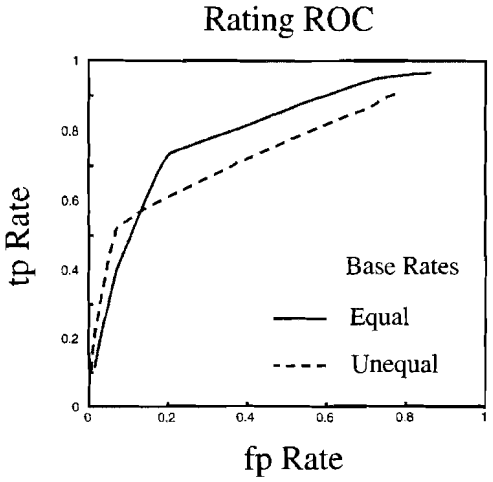


Fig. 5 Rating ROC curves for the size discrimination experiment. Signal detection theory predicts that the two curves should coincide. In the unequal base rate condition, the functions are skewed with respect to the negative diagonal, suggesting larger variance of the ab distribution.

tion theory would predict that the two curves should be identical (the criteria in Fig. 2 should shift with the base rates, but the distributions should remain the same). Instead, these curves are clearly different. When the ab base rate is low, the function is noticeably skewed with respect to the negative diagonal. The direction of the skew and its dependence on the base rates suggests the distribution model shown in Fig. 6. The density function associated with the more frequent stimulus has less variance than the function associated with the less frequent stimulus. In order to be consistent with the results of the bias and suboptimality tests, the decision criterion is placed at the point of intersection between the two distributions.

Effects of base rates on the shapes of the distributions may be taken as evidence against signal detection theory's assumption that human decision makers are "passive receptors" of external world images. Recognizing that perception requires action on the part of human observers, there are in fact many reasons to expect that the information collected from an image should depend on decision making biases. Unlike the automated detector, humans must make a number of decisions about which areas of an image to scrutinize, how to "focus" the sensors (focusing the gaze in visual image processing), and for how long. Detection theory ignores these "control" aspects of the human perception process. It is not unreasonable to suppose that expectations about what an image will contain and other decision making variables of this kind may effect these control processes in nontrivial ways. If so, the measurement of human diagnostic skill and its relationship to an image format must account for these additional variables.

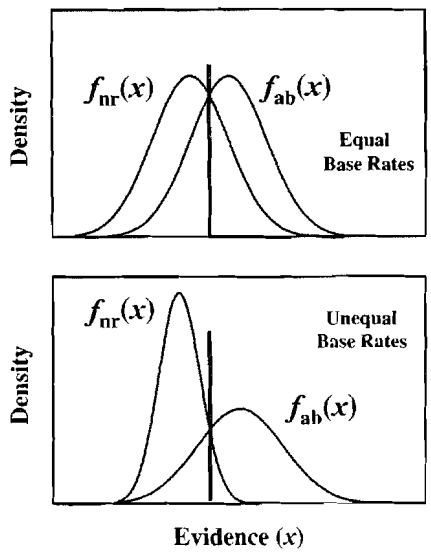


Fig. 6 A distribution model consistent with the results of the bias and optimality tests and the rating ROC curves from the size discrimination experiment.

5.3 *Applications to Image Processing*

Empirical violations of the detection theory class of models could of course be specific to visual perceptual discrimination tasks, in which physical stimulus noise is minimal and presumably internal noise is therefore the main limiting factor. In medical imaging, physical noise is also a conspicuous factor and its presence may alter the nature of the detection process. Diagnosticians are also trained over prolonged periods of time and this experience may enable them to profit from both the encoding effects exhibited in perceptual discrimination and the adjustment of their decision rules. Even if there is no particular reason to expect the detection models to work in some settings but not in others, its validity should be established or disconfirmed in each individual area, including medical diagnosis. The ideal method is to replace the "yes–no" response procedure of the detection theory paradigm with a bipolar rating scale running from high confidence NR judgments to high confidence AB judgments. An explicit cutoff between NR and AB judgments should be defined in the middle of the scale. For example, using a ten-point bipolar scale, the descending integers from 5 to 1 on the left of the response box would be the high to low confidence NR responses and the ascending integers from 1 to 5 on the right of the box would be the low to high confidence AB responses. This procedure has the advantage of forcing the subject to make both the classification judgment and the confidence rating simultaneously, avoiding the possibility that the subject has more (or less) information at the time of the rating response. If the decision rule is unbiased, sample sizes in the study will need to be large enough to show that the relative frequency of lowest confidence NR judgments on nr trials (which will be small) is larger than the relative frequency of lowest confidence NR judgments on ab trials. Detection theorists often promote the detection models by pointing out that d' and other parameters of the model can be estimated without confidence rating data and with relatively smaller sample sizes. It is important to remember, however, that if the model is invalid, none of these conveniences should be enough to justify its application.

6 A Model-Free Approach to Image Quality Assessment

Although not as popular as detection theory, other quantitative models of classification have been developed that can

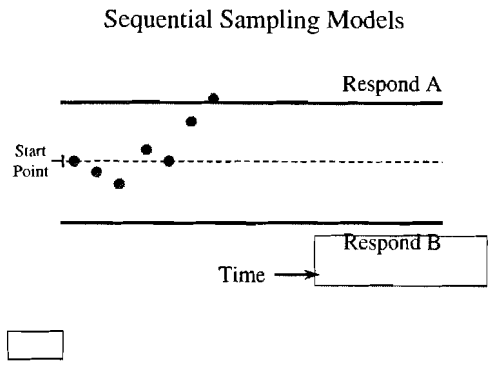


Fig. 7 The sequential sampling models of two-choice classification. The trial begins with the presentation of the image and ends when the decision maker decides that the information collected from the image is strong enough in favor of one of the two judgments to warrant a response.

allow for some of these more complex effects of decision making biases on human performance. The so-called "sequential sampling" models, for example, explicitly define the "stopping rule" used by the subject to determine when the information assimilation process should be terminated.[10,11] This class of models is illustrated in Fig. 7. The information assimilation process starts with the presentation of the image, which initiates a stochastic process that drifts toward one of two "absorption boundaries." The information accumulation process ceases when its value is larger than the positive boundary or smaller than the negative one, and the decision maker's response depends entirely on the sign of the process at the point of termination (i.e., which boundary was crossed).

Models of this kind can fit all of the data typically associated with signal detection theory, including the ROC curves, as well as other results, such as the decision times.[10] Unfortunately, they have more free parameters and parameter optimization can be difficult. Moreover, like signal detection theory, they also depend on some "technical" assumptions (e.g., the drift distributions, discrete versus continuous representation of state and time, and the form of the absorbing boundary functions). An alternative approach that does not require these modeling assumptions or model fitting is to employ the ratings technique described above and perform the optimality test illustrated in Fig. 4 on the results. Confidence rating procedures add little or no extra costs in time or effort, are already used in some areas of medical imaging research (for a recent example, see Leichter *et al.*[12]), and greatly increase the range of hypotheses that can be tested from the data.

The main reason to perform the optimality test defined above is the following: When the base rates were equal in the perceptual discrimination experiments, the function relating percent correct to confidence level in an NR response coincided with the function relating percent correct to confidence level in an AB response, and the decision rule (criterion placement) was optimal. Presumably, decision making biases were not limiting the performance of subjects when these two functions were "calibrated" in this sense (i.e., when percent correct at any given degree of confidence does not depend on which judgment, AB or NR, is given).[13] However, when the base rates were unequal, the two precent correct functions diverged and the decision rule was clearly suboptimal (even though the encoding process was apparently "adaptive," in the manner illustrated in Fig. 6). The degree of separation between these two percent correct functions can therefore be taken as a qualitative index of the degree to which the decision making processes limit discrimination performance.

When these tests are applied to results from two different imaging conditions (with identical base rates and payoffs), the percent correct functions would ideally be calibrated, and in this case the overall percent correct score would be a reasonable index of perceptual information content. If they are miscalibrated, the ideal solution would be to give the subjects feedback about this effect and train them to correct it. That is, in addition to feedback about whether the decision was correct or incorrect at the end of a trial, the experimenter may also give the subjects access to a running tally of their percent correct scores for each confidence level by response (i.e., the same information shown in Fig. 4) and ask them to calibrate these two functions. This self-correction process will obviously take time and may not always be feasible. However, in addition to a more credible final result, the method would also provide an opportunity to study the learning process and to determine why the subjects were miscalibrated initially, which may have important implications for image processing theory.

7 Implications and Some Recent Applications

Although it is possible that visual discrimination is not a good predictor of human behavior in other kinds of discrimination tasks, the fact that signal detection theory has been just as "successful" in other areas as it has been in visual discrimination may be an indication that it is invalid in many areas, for similar reasons. Taking this contrary position for granted, in this section we briefly discuss some recent applications of detection theory in imaging research and consider how the conclusions of these studies might or might not be affected.

7.1 *Ideal Observer Models and Objective (Task-Based) Image Quality Assessment*

Whatever method is used to acquire it, the information in a medical image is finite and time-invariant. Objective indices of image content, or figures of merit, can be defined by specifying in as much detail as possible the physical properties of a population of images and identifying a transformation (the response function) and decision rule that can be applied to these response functions to perform a given classification task. In this vein, Eckstein *et al.*[14] have recently shown how correlations among Gaussian distributed pixel luminance values change the function relating d' (the assumed difference in the mean pixel luminance for signal versus nonsignal pixel intensities) to the percent correct of a signal detector in an n-alternative forced choice detection task (the signal always occurs in one and only one of n locations). This corrected d' is a possible figure of merit for signal detectability.

Since an accurate distribution model would lead to accurate estimates of criterion placement, and the Gaussian model apparently failed in this respect, this particular

model cannot be an ideal choice for defining figures of merit. The fact that adding correlations that exist in an image to a statistical decision model leads to better prediction of human performance suggests nevertheless that humans are affected by these correlations as well. Barrett *et al.*[15] point out that normality is not necessarily a property to be expected of real world images and that even if normality is true of the image in some sense, it can easily be lost in the distribution of the likelihood statistic of an ideal observer. In this respect, the area under the ROC curve might be a better figure of merit for measures of image quality tied to ideal observers. An ideal observer would not exhibit, however, any effects of base rates on the evidence distributions (the distribution of the test statistic would be equivalent under some monotone transformation and therefore the ROC curves would be the same). Perhaps the simplest way to reproduce this property of human observers is to assume that image sampling time is an important factor for humans, even if this attitude is not rational from the point of view of the experimenter. As noted earlier, changes in the stopping rule of a dynamic system (the distances to the absorbing boundaries) can have arbitrarily strong effects on the detectability of a signal and therefore on the distribution of evidence (in this context, evidence is the information available at the point at which the sampling process is terminated, or in other words, the posterior likelihood of an entire sample path[16,17]). Another possibility is that some aspect of the human image transduction process makes it advantageous to alter the nature of the image transformation depending on the most likely condition of the image.

7.2 *Effects of System Versus Anatomical Noise*

In some areas of radiology, the difference between signal and noise is problem dependent: anatomical features that are irrelevant to one kind of diagnosis may be critical to the diagnosis of another kind.[18,19] This anatomical (or patient structured) noise almost surely needs to be treated differently from acquisition or system noise. Bochud *et al.*[18] have recently developed a method of quantifying the different effects of system and anatomical noise by comparing human performance to the performance of a detection model (the nonprewhitening matched-filter observer with an eye filter) when the anatomy is assumed known (and hence removed) or unknown (and hence acts as noise). Instead of the yes–no detection task, they employed the two-alternative forced choice (2AFC) design, under the assumption that percent correct in this design is equal to the area under the ROC curve in the yes–no detection task.[6] Images were simulated with different degrees of anatomical variations and three different signal profiles, tumor nodules, spherical, and cubical microcalcifications. Comparing human performance to the model's predictions with anatomy known and anatomy unknown, the authors found that the influence of anatomic noise depends strongly on the signal profile. Increasing anatomical noise from small to moderate levels had a more substantial impact on the detection of cubic microcalcifications than on nodule detection (see Fig. 4 in Bochud *et al.*)

Percent correct in the 2AFC paradigm may be a reasonable index of detectability, even if the experiment is not a realistic representation of clinical diagnosis. However, the relationship between human yes–no and human 2AFC (Green's area theorem) is lost if the criterion shift construct is invalid. For each base rate pair, there is a different pair of empirical distributions (see Fig. 5) and hence a different ROC curve, with potentially different area underneath it. Estimating the tp and fp rates under different base rate ratios defines yet another ROC curve, whose area may or may not equal the human subject's percent correct score in the 2AFC design. It seems unlikely that the area index would vary dramatically under different base rates, and there is no specific reason to doubt the conclusions of Bochud *et al.*'s study. However, until the effect of base rates on area is identified for problems involving expert diagnosis, there is some risk in assuming that the important properties of human diagnosis under different base rate conditions can be inferred accurately from behavior in the 2AFC condition.

7.3 *Computer-Aided Diagnosis*

An area in which accurate models of human decision making biases may be especially fruitful is computer-aided diagnosis (CAD). Reconstruction and detection algorithms can potentially identify image features not easily discriminated by the human visual system[20] and can also provide a "second opinion" with an explicit quantitative basis. It is important to understand, however, how information from an automated system will be combined with the radiologist's conventional approach to diagnosis.

In a recent study comparing CAD to unassisted diagnosis, Leichter *et al.*[12] used image enchancement and computerized extraction of quantitative features (spiculations) to help radiologists visualize and interpret lesions in a digitized mammograph. Participants in the study sorted images into one of five ordered categories on a bipolar scale: normal, benign, probably benign, suspicious abnormality, and highly suggestive of malignancy. The fp rates in the two conditions were equal, however the tp rates were larger and the ROC curves were clearly ordered.

From a strict detection theory point of view, a "second opinion" from a statistical algorithm should only affect the criterion set on the information that the observer has extracted from the image. The "sensitivity effect" of CAD suggests that the system either provides new information not readily accessible to the radiologist or changes the way the radiologist processes the image. It seems likely that the time to diagnose was longer in the CAD condition, which could reflect either of these two factors, assimilation of new information from the computer, or spending more time examining regions of the image due to the feedback from the CAD. CAD systems might also have an effect on the radiologists memory for specific details, or the relative weights assigned to different image properties prior to formulating an opinion. All of these effects would constitute changes in the radiologist's computation of the test statistic, leading to changes in sensitivity and in the shape of the rating ROC curve.

It is not clear what effect a manipulation of the base rates would have had on this curve. However, one interesting possibility is that receiving information from the CAD system favoring a given judgment could have the same effect that base rates appear to have in visual perceptual discrimination, changing the relative variances of the distributions of evidence effects. In other words, Fig. 6 could be a

model for experimental trials when the output of the CAD system suggests that the image is normal (and the opposite variance tradeoff might occur when the system information suggests that the image is abnormal). If so, then the decision rule is suboptimal and feedback training (recalibration) may be needed to maximize the benefits of the CAD.

8 Conclusions

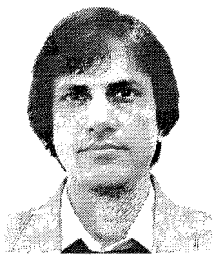
Automatic detection algorithms typically convert the intensity values of a digital image into a test statistic that is assumed or known to be monotonically related to the objective posterior likelihood of the signal. A threshold on this value maximizes accuracy for some pair of signal and noise priors, and the ROC curve generated by varying the threshold is a precise and unambiguous measure of the efficiency of the detector. If signal detection theory's assumption that human classification can be adequately modeled using a similar two-step process is not tenable, as we have argued, the significance of human ROC curves and other statistics derived from signal detection theory is more difficult to specify. Because they do not rely on a model of the detection process, interpretation of the tests that we have defined on confidence ratings is less problematic. For example, if the test for suboptimality is satisfied, then under some conditions (e.g., when the subject would normally respond to AB with low confidence), the subject should select the alternative response. Training them to recalibrate their subjective confidence states guarantees that accuracy will improve and may also eliminate other suboptimal properties of the decision making strategy. A reasonable working hypothesis is that subjects that use optimal and calibrated decision rules in the sense defined here are maximizing their performance levels. In such a case, percent correct of these subjects may be the most appropriate index of image content.

Acknowledgment

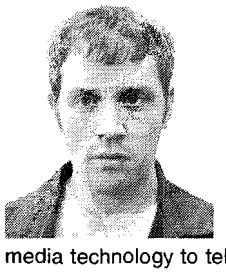
This work was supported by National Science Foundation Grant No. SBR-9709789 (J.D.B.).

References

1. D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics* (Reprint), Krieger, Huntington, NY (1974).
2. T. Aach. U. Schiebel, and G. Spekowius, "Digital image acquisition and processing in medical x-ray imaging," *J. Electron. Imaging* **8**(1), 7–22 (1999).
3. E. J. van der Jagt, S. Hofman, B. M. Kraft, and M. A. van Leeuwen, "Can we see enough? A comparative study of film-screen vs digital radiographs in small lesions in rheumatoid arthritis," *Eur. Radiol.* **10**(2), 304–307 (2000).
4. H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers," *Med. Phys.* **26**(12), 2654–2668 (1999).
5. *Multidimensional models of perception and cognition*. F. G. Ashby, Ed., Lawrence Erlbaum Associates, Hillsdale, NJ (1994).
6. D. M. Green, "General prediction relating yes-no and forced-choice results," *J. Acoust. Soc. Am.* **36**, 1024 (1964).
7. J. D. Balakrishnan and J. A. MacDonald, "Signal detection theory," *International Encyclopedia of Ergonomics and Human Factors*, E. Karwowski.
8. J. D. Balakrishnan, "Measures and interpretations of vigilance performance: Evidence against the detection criterion," *Hum. Factors* **40**, 601–623 (1998).
9. J. D. Balakrishnan, "Decision processes in discrimination: fundamental misrepresentations of signal detection theory," *J. Exp. Psychol.* **25**(5), 1189–206 (1999).
10. R. D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organization*, Oxford University Press, New York (1986).
11. J. T. Townsend and F. G. Ashby, *Stochastic modeling of elementary psychological processe*, Cambridge University Press, New York (1983).
12. I. Leichter, S. Fields, R. Nirel, P. Bamberger, B. Novak, R. Lederman, and S. Buchbinder, "Improved mammographic interpretation of masses using computer-aided diagnosis," *Eur. Radiol.* **10**, 377–383 (2000).
13. The term 'calibration' is often used in the human decision making literature to refer to the difference between subjective estimates of the probability of a correct judgment and observed probability correct. Here, it is the difference between percent correct scores for equivalent confidence levels (not probability correct judgments).
14. M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "Visual signal detection in structured backgrounds. IV. Figures of merit for model performance in multiple-alternative forced-choice detection tasks with correlated responses," *J. Opt. Soc. Am. A* **17**(2), 206–217 (2000).
15. H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions," *J. Opt. Soc. Am. A* **15**(6), 1520–1535 (1998).
16. M. Stone, "Model for choice-reaction time," *Psychometrika* **25**, 251–260 (1960).
17. A. Wald, *Sequential Analysis*, Wiley, New York (1947).
18. F. O. Bochud, J. F. Valley, F. R. Verdun, C. Hessler, and P. Schnyder, "Estimation of the noisy component of anatomical backgrounds," *Med. Phys.* **26**(7), 1365–1370 (1999).
19. F. O. Bochud, F. R. Verdun, J. F. Valley, C. Hessler, and R. Moeckli, "Importance of anatomical noise in mammography," *Proc. SPIE* **3036**, 74–80 (1997).
20. S. Baeg, S. Batman, E. R. Dougherty, V. G. Kamat, N. Kehtarnavaz, S. Kim, A. Popov, K. Sivakumar, and R. Shah, "Unsupervised morphological granulometric texture segmentation of digital mammograms," *J. Electron. Imaging* **8**(1), 65–75 (1999).



J. D. Balakrishnan received his PhD in cognitive-mathematical psychology from the University of California at Santa Barbara (UCSB) in 1991. He is a charter member of the American Psychological Society and a recent Fellow at the Hanse Institute for Advanced Study in Delmenhorst, Germany. In 1996, he received the New Investigator Award from the Society for Mathematical Psychology for his work on the organization and dynamics of human cognitive processes. He is currently associate professor in the psychological sciences department of Purdue University. His recent research focuses on the decision making aspects of human perception and categorization and applications of multimedia and virtual reality systems to air traffic and flight test control.



Justin A. MacDonald is a PhD student in the department of psychological sciences at Purdue University. He received his MS degree in quantitative-mathematical psychology from Purdue in 1999 and is currently a research assistant with funding from the National Science Foundation. His interests are in modeling and measurement of cognitive processes in auditory perception, with emphasis on speech communication, and in applications of multimedia technology to telecommunication systems.