Psychonomic Bulletin & Review
2002, 9 (4), 858-865

Decision criteria do not shift: Reply to Treisman

J. D. BALAKRISHNAN and JUSTIN A. MACDONALD
Purdue University, West Lafayette, Indiana

Recently, we argued that the detection criterion representation of decision-making biases, embedded within the theory of signal detection, is empirically testable and has, in fact, been falsified by empirical results from visual discrimination experiments. Treisman (2002) attempts to show that there is an alternative interpretation of our results that could explain them without dropping the detection criterion construct. In lieu of attempting to fit the data with a model, however, he gives two kinds of theoretical examples, both involving manipulations of the spacing of criteria on a decision axis. The first example correctly predicts that the bias estimate we developed will be zero but does so by assuming zero spacing between some criteria (some rating responses are never used). We did not observe zero spacing between any criteria and did not perform any analyses on responses that never occurred. Moreover, this example does not explain why the upper-bound bias estimates that we obtained by combining results from two criteria placements were also trivially small. His second example predicts that the bias should have been detectable with sufficiently large sample sizes. In our experiments, the sample sizes were, in fact, quite large, large enough for the results to be consistent in 18 different experimental conditions. Finally, all of Treisman's criteria placement examples also fail to explain the pronounced effects of base rates on the shapes of the rating ROC curves, and his suggestion that there are problems of logical interpretation with our proposed distribution model ignores the predictions of large classes of alternatives to detection theory, including the dynamic models of perception.

In several recent papers, we claimed to have discovered clear and compelling evidence that the general concept of a shifting decision criterion—rooted deeply within the theory of signal detection and many other formal models of behavior—is fundamentally wrong (Balakrishnan, 1998a, 1998b, 1999; Balakrishnan & MacDonald, 2001a, 2001b). Instead of changing the position of a decision criterion, biases toward one of the two judgments in a discrimination task cause a tradeoff in the variances of the encoding distributions, decreasing the variance of the bias-favored stimulus distribution and increasing the variance of the other, unfavored stimulus distribution. The tradeoff in variances causes the hit and false alarm rates to covary with signal base rate, even though the detection criterion remains fixed at the point of intersection between the two distributions. Although problematic from the detection theory point of view, this effect turns out to be perfectly consistent with the dynamic models of perception (e.g., Baranski & Petrusic, 1998; Diederich, 1997; Link & Heath, 1975; Luce, 1986; Ratcliff & Rouder, 2000; Ratcliff, Van Zandt, & McKoon, 1999; Smith, 2000; Townsend & Ashby, 1983; Van Zandt, 2000; Vickers, Smith, Burt, & Brown, 1985) and with some modified versions of detection theory developed by Green, Luce, and others, albeit for very different purposes (Green & Luce, 1974; Luce & Green, 1978; see Balakrishnan, 1999).

In his critical rejoinder, Treisman (2002) considers the results reported in Balakrishnan (1998a, 1998b, 1999) and then presents a series of theoretical arguments and counterexamples that purportedly show that the criterion shift construct is still tenable and that we overstated our case. Essentially, his point is to show that the bias estimates we reported might be explained by certain weaknesses (low power) of the tests we developed to detect criterion shifts when they occur. However, before drawing our conclusions about the detection theory models, we considered the possibilities that Treisman raises and dismissed them for reasons discussed in our previous articles.

Since there is a lot at stake for researchers in this area and both viewpoints about what models are ruled out by the data cannot be correct, it is important to examine these quantitative issues carefully to determine which conclusions are correct and which are incorrect. In this article, we will consider each of Treisman's (2002) arguments and show that the solutions he offers, although they would be adequate for some hypothetical empirical results, cannot explain the results that we actually obtained. Summarized briefly, when the signal rate is low, empirically observed likelihood ratios for low-confidence noise responses are very close to 1, indicating a zero or a trivially small bias toward the noise response, and the relative frequencies of these responses are close to zero but not equal to zero, indicating that the likelihood ratio estimate is being taken very close to the decision criterion separating signal and noise responses. This consistent finding was the basis of our conclusions about problems with the detection criterion conceptualization of bias, and none of Treisman's examples addresses it. Unfortunately, it is impossible to prove that there is no plausible alternative version of the criterion shift construct that can explain these results. In lieu of that, we have made all of the relevant datasets that we have reported previously available on the Internet (URL: http://www.psych.purdue.edu/Quantitative/dsdt.html), along with some software that can be used to perform the appropriate data analyses. Using these resources, interested readers can determine for themselves what aspects of the data can and cannot be accounted for with classical detection theory approaches and modifications.

The authors were supported by National Science Foundation Grant SBR-9709789 and by NASA Ames Grant NA57-1407 (Intelligent Systems). Correspondence concerning this article should be addressed to J. D. Balakrishnan, Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907 (e-mail: jdb@psych.purdue.edu).

Although our main point is to respond to his critique, it is also important to recognize that Treisman is the first detection theorist to have made a serious attempt to understand the methods and results we presented and to evaluate them properly before promoting the classical detection theory approach. Many of the points he makes are valid and insightful. Moreover, his discussion should make it easier for investigators familiar with the detection theory motifs to understand the sometimes confusing issues involved.

Testable Predictions of the Criterion Shift Construct

Treisman's defense of detection theory can be divided into seven different arguments. His first addresses the advantages and disadvantages of an alternative measure of sensitivity proposed in Balakrishnan (1998a). Since that issue is unrelated to the central question of the validity of the shifting-criterion construct, we consider that argument last. Also, for later reference it will be helpful to begin with a brief, alternative explanation of the methods and data that makes their relationship to detection theory more transparent.

Criterion shifts and empirical likelihood ratios. Although it may not have been conspicuous in our earlier articles, the methods we developed to test the criterion shift construct are not entirely new. They are equivalent to estimating the slope of an empirical receiver-operating characteristic (ROC) curve at different points and appealing to a well-known principle identified long ago by detection theorists—that is, that the slope of the ROC curve at a given point (i.e., at a given value of the false alarm rate) is equal to the likelihood ratio of the two distributions that generate the curve at this point (Green & Swets, 1974).

According to detection theory (or at least, one popular version of it), associated with each point of the ROC curve is an *evidence state* or percept. Somewhere on this continuum is a cutoff between signal and noise responses, which should depend on the base rates or payoffs. So, if the ROC curve could be obtained for a task in which the signal base rate is substantially less than the noise base rate, the slope of the curve at this cutoff point should be greater than 1 (the detection criterion should be biased toward the noise response).

Now suppose that in addition to the yes–no detection judgment, the subjects must also report a degree of confidence in the accuracy of this judgment. This can be done by having them make a single response on a bipolar rating scale with a cutoff in the middle separating noise and signal responses or by eliciting first a yes–no judgment and then a confidence-rating response. According to detection theory (or, again, one popular version of it), these kinds of data can be "successfully" modeled using a pair of distributions placed on a decision axis and a set of criteria that map each possible piece of evidence or percept to one and only one rating response, as illustrated in Balakrishnan (1998b, Figure 2, and 1999, Figures 2–3) and in Treisman (2002, Figures 3–6).

The main point of our results was that the slope of the ratings ROC curve at the cutoff between signal and noise responses was always 1 or close to 1, regardless of the base rates and despite the fact that the hit and false alarm rates were, as was expected, strongly affected by the base rates. Although we used a different function to perform these tests,

$$U_{\mathrm{R}}(k) = F_{\mathrm{R,n}}(k) - F_{\mathrm{R,s}}(k),$$

where s and n denote signal and noise trials, respectively, k is the rating response (R) on a bipolar scale, and F denotes the cumulative distribution function, instead of the empirical likelihood ratio, $f_{\mathrm{R,s}}(k) / f_{\mathrm{R,n}}(k)$, we also pointed out that the information in a graph of $U_{\mathrm{R}}(k)$ is equivalent to computing the slope of the ROC curve (Balakrishnan, 1998b, 1999). That is, the slope of the ROC curve is less than or greater than 1, depending on whether $U_{\mathrm{R}}(k)$ is increasing or decreasing. We point this out again because it is important to recognize that there is nothing unique about the properties of $U_{\mathrm{R}}(k)$ that determine whether our conclusions are justified or not; the issue is whether detection theory can explain the observed $U_{\mathrm{R}}(k)$ function and, hence, the observed likelihood ratio functions obtained under different base rate conditions.

Fitting these likelihood ratio data would be easy to do if the criteria on the decision axis could be placed freely, with arbitrary spacing between them. However, their placement must also correctly predict the relative frequencies of the individual rating responses. If these are small, either the spacing between the criteria must be small, or the area under the distributions between these criteria must be small (or both). Essentially, this constraint on the model allows it to be falsified empirically. The more dense these criteria points are, the closer are the slopes of the curve to single likelihood ratios defined by single evidence states, and the smaller are the relative frequencies of the rating responses that generate these points. On the basis of the relationships we observed between response frequencies and likelihood ratios, we concluded that the likelihood ratio at the detection criterion is always 1 or close to 1 (i.e., unbiased). Treisman's purpose is to show that the detection criterion might have been biased despite these results, for one of the several possible reasons enumerated below.

Treisman's Objections

1. *The test for bias assumes that the spacing of the criteria on the evidence axis is constant, which is an unreasonable expectation and is contradicted by the instructions given to subjects to use extreme responses infrequently* (Treisman, 2002, p. 851). *Violations of the equal spacing assumption "distort" the information available in our test function, $U_{\mathrm{R}}(k)$.*

The criteria chosen by the subject need not be equally spaced in order for bias in the decision rule to be identified or overlooked. What is important is the relative frequency of the lowest confidence-rating responses. If these are small, yet the likelihood ratios associated with them are close to 1, the detection theory model cannot explain the effect of the base rate manipulation by shifting the position of the detection criterion (to a point where the ratio departs substantially from the value 1). Furthermore, we do not need to "approximate the theoretical $F_{\mathrm{R,n}}(k) - F_{\mathrm{R,s}}(k)$ curve," using the observable curve $U_{\mathrm{R}}(k)$. Our claims about detection theory are based on the inability of this class of models to account for the observed likelihood ratio data combined with the observed response frequency data (unless they assume that the distributions change shape and the detection criterion remains unbiased, as we proposed).

2. *It is possible to construct a detection model in which some of the rating responses have zero frequency, in such a way that the $U_{\mathrm{R}}(k)$ function will miss the bias, even if it is estimated accurately (i.e., with large sample sizes).*

The theoretical, detection theory example given by Treisman (2002, Figures 5E and 5F) does, in fact, constitute a case in which the test for bias fails even though the decision rule is biased. As he points out, responses with zero frequency would be modeled by assuming that some of the criteria have the same value, so that there is zero area under the signal and noise distributions between them (and hence, zero probability of the corresponding response).

However, in our experiments, all of the rating responses on the 14-point bipolar scale had nonzero frequencies (Balakrishnan, 1998b), or the responses with zero frequency were ignored (i.e., when subjects reported their confidence using a 200-point sliding scale, Balakrishnan, 1999). The key issue with regard to the strength of the case against detection theory is the relative frequencies of the lowest confidence-rating responses when these frequencies are nonzero and, hence, there must be some space between the criteria in the detection model in order to fit the data.

Because the empirical $U_{\mathrm{R}}(k)$ function could, in principle, have many different possible shapes and varying degrees of estimation error, depending on the sample sizes used to estimate it, it is difficult to give a list of instructions that users can follow to interpret this function properly without properly understanding the theoretical issues involved. The "instructions" in Balakrishnan (1998b) referred to the case in which $U_{\mathrm{R}}(k)$ is increasing up to the cutoff and decreasing thereafter, and hence, there are no rating responses with nonzero frequency, as in the experiments we were reporting in that article.

For Treisman's case (2002, Figure 5E), the upper bound on the proportion of biased noise judgments would be the occurrence frequency of rating response "4," which is quite large (about .32). In our experiments, when the $U_{\mathrm{R}}(k)$ function reached its peak at the response cutoff, the upper-bound estimate was never greater than .034 (Balakrishnan, 1998b). Similarly, the upper bound on the proportion of biased signal judgments in Treisman's example would be the occurrence frequency of rating response "8," which is also large (about .23) and dramatically different from our empirical results. Although Treisman is right to point out that even by collecting large samples generated from his model example, the results would be inconclusive, this model cannot fit our results. If we had obtained data of this kind, we would not have concluded that the detection models should be rejected. In each of the 10 conditions reported in Balakrishnan (1998a, 1998b), both the estimated proportion of biased responses and the estimated upper bounds on the proportion of biased responses were trivially small. In Balakrishnan (1999), the estimated proportion of biased responses and the upper bound were identical (because the confidence response was effectively continuous valued), and all of these estimates were trivially small. For comparison with Treisman's examples, empirical estimates from Balakrishnan (1998b) are listed in Table 1.

Another way to illustrate why the spacing argument will not suffice to protect the shifting-criterion construct is to consider the predictions of these models about the rating ROC curves under different base rate conditions. All of Treisman's examples predict the same underlying ROC curve, observed at different points when the criteria shift, as is illustrated in Figure 1 (top panel). In Balakrishnan (1998b, Figure 9), the shapes of the empirical ROC curves were strongly influenced by the base rates, in a manner suggesting the variance tradeoff model that we proposed to account for the results (see Balakrishnan, 1998b, Figure 8, and 1999, Figure 9). Examples from two of the conditions in Balakrishnan (1998b) are shown in the lower panels of Figure 1. Instead of mapping out a single underlying function, the curves become noticeably skewed when the base rates are unequal. To fit ROC curves like these, detection theorists would typically

Table 1
Estimates of the Detection Theory Parameters and Results of the Distribution-Free Tests for Decision Rule Bias in the Two Experiments Reported in Balakrishnan (1998b)

Condition	β_{SDT}	Max(lr)	SDT$_p$	Ω_p	Max(p_{bias})
		Experiment 1			
1 (successive)*	3.454	0.976	.163	.000	.031
2 (simultaneous)*	2.772	0.850	.203	.000	.002
3 (mixed successive)*	3.056	1.020	.207	.004	.039
4 (mixed successive)†	1.167	0.523	.033	.000	.011
		Experiment 2			
1 (successive)*	2.803	1.298	.088	.007	.037
2 (simultaneous)*	3.012	0.764	.120	.000	.034
3 (mixed successive)*	2.800	1.535	.154	.024	.098
4 (mixed successive)†	1.457	0.615	.044	.000	.014

Note—The values of SDT$_p$ given in Balakrishnan (1998b) were obtained—that is, by fitting the detection model to the rating data, finding the predicted peak of $U_{\mathrm{R}}(k)$ from this fit, and then computing the proportion of rating responses at this predicted peak or closer to the response cutoff (e.g., the proportion of "6" and "7" responses if the predicted peak was at response "6" and the response cutoff was at "7"). β_{SDT} is the estimated likelihood ratio of the decision criterion derived from signal detection theory, Max(lr) is the maximum value of the empirically observed likelihood ratio on the "noise" side of the rating response scale, SDT$_p$ is the estimated proportion of biased responses across signal and noise trials derived from signal detection theory, Ω_p is the distribution-free estimate of the proportion of biased responses across signal and noise trials, and Max(p_{bias}) is the conservative upper bound on this distribution-free estimate (see Balakrishnan, 1998b, 1999). *Base rate ratio: 9 to 1 (optimal likelihood ratio at the cutoff, lr_7, is 9.0). †Base rate ratio: 1 to 1 (optimal likelihood ratio at the cutoff, lr_7, is 1.0).

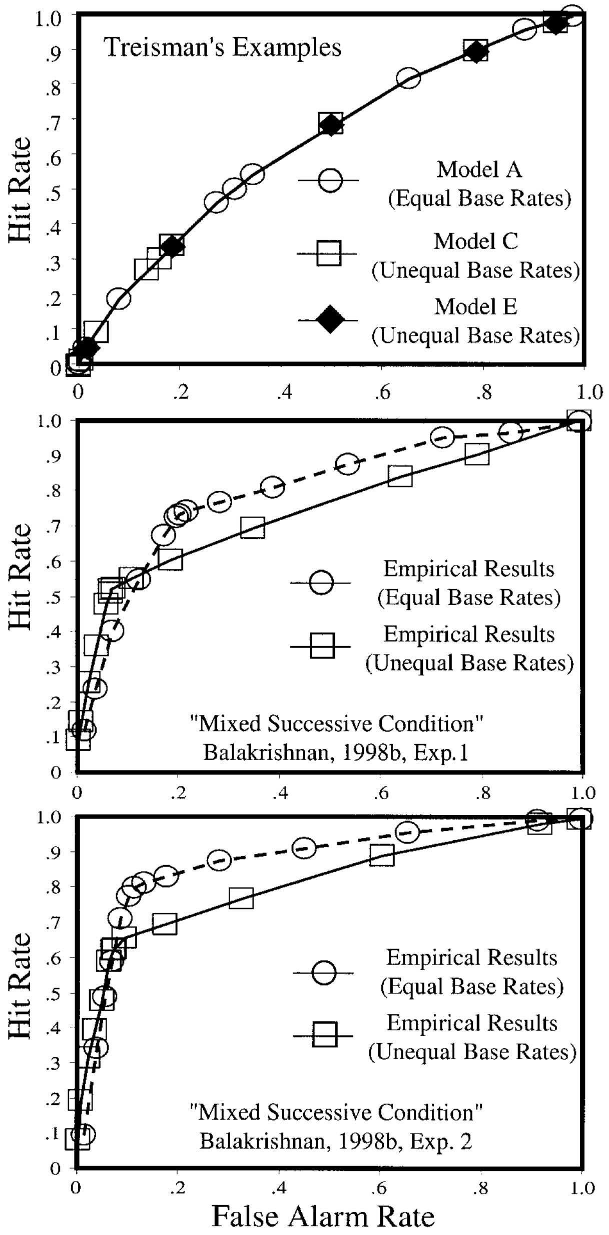

Figure 1. Rating receiver-operating characteristic (ROC) curves under equal and unequal base rate conditions predicted by Treisman's proposed models and the empirical rating ROC curves in Balakrishnan (1998b). Changing the base rates should have no effect on the shape of the curve under the detection theory assumptions. Instead, they become sharply skewed with respect to the negative diagonal when the base rates are unequal, consistent with the variance tradeoff model proposed in Balakrishnan (1998b, 1999).

propose two distribution models, an equal variance normal distribution model for the equal base rate condition (because the curve is symmetric around the negative diagonal) and an unequal variance model for the unequal base rate condition (because the curve is positively skewed with respect to the negative diagonal).

3. *The criteria are spaced in such a way that the bias would be detectable in the true $U_R(k)$ function, but the spacing is so small that this function may not be estimated with sufficient accuracy to identify its maximum value.*

The ROC curve data should be enough to show that Treisman's examples cannot fit the data adequately to save the detection models, regardless of the sizes of the spacing between the criteria and the resulting sample sizes. Treisman's example (Figure 5C) is also misleading because, in order to make the $U_R(k)$ function relatively flat at its center, he assumes a small value of d' (.5) and a relatively small degree of bias ($\beta = 1.4$). In our experiments, the d' and β values were much larger. So, if the same criteria spacing is used that Treisman used, the peak of the $U_R(k)$ function to the left of the response cutoff would be obvious. Even without considering these results, however, sample size and estimation errors can be ruled out for a more obvious reason: The results replicate. In each of 10 different conditions in the two rating scale studies (2 in Balakrishnan, 1998a, and 8 in Balakrishnan, 1998b), the relative frequency of the lowest confidence noise response was trivially small, and the likelihood ratio associated with this response was less than 1 or close to 1, indicating a zero bias or a trivially small bias, independent of the base rates.

Likelihood ratios and the relative frequencies of the four lowest confidence noise responses from the two experiments reported in Balakrishnan (1998b) are given in Table 2. Also shown are the total number of samples per stimulus (multiplying the sum of these two values by the relative frequency of a rating response yields the frequency of this response). Across all eight conditions, the largest observed value of the likelihood ratio was 1.535, despite the fact that the relative frequencies of the rating responses associated with these maximum ratio values were always small, implying small spacing of the criteria in the detection model fit. If the true likelihood ratio at the decision criterion and the true proportion of biased responses were in the ranges indicated by the fit of the detection model (i.e., 2.772–3.454 for the likelihood ratio and .088–.207 for the predicted proportion of biases judgments), it would be difficult to explain why our likelihood ratio estimates were consistently much closer to 1 and the estimated proportion of biased responses from the $U_R(k)$ functions were consistently zero or close to zero.

It should be noted that our sample sizes were quite large (in the thousands, in many cases), as compared with the range that Treisman (2002) appears to have in mind (e.g., around 200), which presumably explains why our estimates have small variance across conditions. We never claimed that our tests for bias could be reasonably carried out in the types of small studies to which many investigators are accustomed. Because they must identify the shape of two distributions in their tails (i.e., between the signal and the noise distributions), the sample sizes need to be large. This raises, of course, some issues about the effects of combining over sessions and subjects and the relationship between decision-making behavior in a small study, as compared with behavior in a large one; we will consider this issue (not raised by Treisman) in our discussion.

Table 2
Estimates of the Likelihood Ratio Function and Rating Response Probabilities from Balakrishnan (1998b)

Condition	lr_4	lr_5	lr_6	lr_7	p_4	p_5	p_6	p_7	N_n	N_s
				Experiment 1						
1 (successive)*	0.442	0.482	0.669	0.976	.166	.110	.035	.031	12,153	1,336
2 (simultaneous)*	0.547	0.510	0.850	0.414	.171	.056	.021	.002	11,175	1,226
3 (mixed successive)*	0.538	0.676	0.815	1.020	.157	.075	.035	.004	6,099	690
4 (mixed successive)†	0.443	0.404	0.405	0.523	.108	.074	.045	.011	2,956	2,952
				Experiment 2						
1 (successive)*	0.285	0.322	0.935	1.298	.123	.042	.031	.007	4,889	554
2 (simultaneous)*	0.292	0.483	0.463	0.764	.171	.132	.069	.034	11,175	1,226
3 (mixed successive)*	0.474	0.502	1.535	0.799	.145	.074	.021	.003	3,139	357
4 (mixed successive)†	0.210	0.411	0.454	0.615	.102	.074	.033	.015	1,363	1,330

Note—lr_k is the observed likelihood ratio for rating response k ("7" is the lowest confidence noise response; "4" is a moderate confidence noise response); p_k is the observed proportion of rating response k; N_n and N_s are the noise and signal trial sample sizes, respectively. *Base rate ratio: 9 to 1 (optimal likelihood ratio at the cutoff, lr_7, is 9.0). †Base rate ratio: 1 to 1 (optimal likelihood ratio at the cutoff, lr_7, is 1.0).

4. *The confidence ratings may not provide a serviceable estimate of the true likelihood ratio values that determine performance—that is, these data may be contaminated in some way. Moreover, since the tests rely on responses with very low relative frequency, the results should be highly susceptible to even a small number of "bad" trials. (Note: "contaminated" and "bad" are terms we, not Treisman, use to discuss this issue.)*

First, with regard to the general appraisal of the ratings method, there is of course no way to prove that confidence ratings are not simply "bad data" whose properties should be ignored by all investigators. However, several of these properties certainly give the impression that these data do provide some important information about the discrimination process. For example, in our studies and elsewhere, observed percent correct scores increase monotonically with increasing confidence. Second, with regard to the (relatively) small frequencies of the lowest confidence-rating responses, the shapes of the ROC curves under different base rate conditions seem to constitute clear evidence that the violations of the detection theory assumptions about criterion shifts (i.e., that the detection criterion shifts but the distributions remain the same) are not limited to properties of just the lowest confidence judgments.

More important, as we pointed out in our previous papers, it is difficult to concoct a contamination effect of any kind that causes the bias test to miss the bias in the decision rule but allows the complementary test for suboptimality to detect the suboptimality of the decision rule. As we noted earlier, the bias test relies on the empirical likelihood ratio,

$$L_R(k) = f_{R,s}(k) / f_{R,n}(k),$$

for the lowest confidence noise responses (when the noise base rate is high), whereas the suboptimality test is based (equivalently) on the empirical odds ratio,

$$O_R(k) = \frac{f_{R,s}(k)p_n}{f_{R,n}(k)p_s},$$

where p_s and p_n are the signal and noise base rates, respectively, and k represents the lowest confidence signal response (again, when $p_n > p_s$). It is not clear (at least to us) how a contamination effect could distort the results in such a way that one of these tests misses the bias, whereas the other detects the suboptimality.

5. *Showing that the decision rule is suboptimal does not refute the detection theory model. Moreover, the results of the test for suboptimality (percent correct scores around 25%) cannot be reconciled with the fact that the conditional probability of a signal when the likelihood ratio is equal to 1 is equal to the signal base rate (.1 in the experiments).*

The first point is certainly true, and that is why we never made such a claim. We did point out that if the decision rule is unbiased when the base rates are unequal, one should expect to find that it is suboptimal (although this does not *necessarily follow*; see the next section). The "conservatism" account of the decision process derived from estimates of the detection theory parameters (i.e., undershifting rather than no shifting of the criterion; e.g., Creelman & Donaldson, 1968; Green & Swets, 1974; MacMillan & Creelman, 1991; Maloney & Thomas, 1991) leads to the prediction that the decision rule is suboptimal.

The fact that the percent correct values associated with the lowest confidence signal responses were less than 50% confirmed this prediction, but the exact value of these scores does not establish anything about the tests for bias. The numerical examples given by Treisman (2002) assume that the distributions are unaffected by the base rates (an assumption that mispredicts the ROC curves) and that the likelihood ratio function is defined by two equal variance normal distributions (and small d'). The likelihood ratio function at the intersection between these two distributions is not very steep. When these distributions can change shapes under different base rates, the likelihood ratio function can be very steep at the point of intersection between them (as is suggested by the shapes of the empirical ROC curves), so that the criteria spacing might need to be extremely small before the percent correct score would come close to the value 0.1. Without knowing the distributions, it is impossible to determine whether the values we observed are too high with respect to the (qualitative) distribution model we used to illustrate the implications of the bias and suboptimality tests.

More important, this percent correct value need not approach the value 0.1 at all, even in the limit. In fact, using the dynamic models as a basis for identifying possible empirical results, the decision rule could be optimal even though the base rates are very different and the decision rule is completely unbiased. We explain why this is so in the next section.

6. *The proposed new distribution model (encoding plasticity and decision-making rigidity) does not predict the dependence of β on the base rates. Moreover, it is difficult to see how the encoding distributions could change shape unless the observer somehow knows which stimulus is presented, which makes no sense.*

In the variance tradeoff model we proposed, the amount of area under the signal distribution to the right of the (unbiased) criterion (i.e., the hit rate) relative to the amount of area under the noise distribution to the left of the (unbiased) criterion (i.e., the correct rejection rate) trades off as the variances trade off, predicting the effect of base rates on the β parameter (see Balakrishnan, 1998b, Figure 8).

Although it is not surprising that this distribution model would seem implausible to a detection theorist, who has learned to think in terms of an uncontrollable sensor whose output is fed into a decision process, it is entirely plausible from other points of view, including some already developed in the literature. In fact, Van Zandt (2000) has recently shown that the dynamic models can predict precisely the effects of the base rates on the shapes of the empirical rating ROC curves that we observed. The reasons for this prediction are somewhat technical in nature. However, to gain some idea of how this prediction arises, consider the simple random walk model in which random samples from one of two equal variance normal distributions, differing in the signs of their means, are accumulated until the sum exceeds an upper *signal threshold* or a lower *noise threshold.* These thresholds identify both the stopping rule (when to stop sampling and respond) and the decision rule (what response to emit once the sampling process is terminated).

In detection theory, the "effect" of the stimulus is a single random number, which can be converted to a likelihood ratio—that is, the relative frequency of this value on signal trials, divided by its relative frequency on noise trials. In the random walk model, the effect of the stimulus presentation is a sequence of samples, rather than a single one. However, it can still be converted to a likelihood ratio to determine what information about the stimulus is contained in the samples that have been collected. Suppose, for example, that the effect is the sequence 5, -3, 15, which exceeds a signal response threshold that was set at 10. This specific sequence will have one relative frequency (density) on signal trials and another relative frequency on noise trials. The ratio of these two is the information about the stimulus embodied in the three samples collected on this particular trial.

On each trial, a sequence of samples will be collected by the subject, and therefore, one specific value of this likelihood ratio will be generated. Moreover, this univariate ratio value will have one of two distributions, depending on which stimulus was presented. Jiang (2000) showed that whenever the accumulation process crosses the signal threshold, the relative frequency (density) of this set of random samples on signal trials is necessarily larger than its relative frequency on noise trials (note that the base rates are irrelevant when calculating these densities). This means that however subjects group these signal-crossing sample paths into signal-rating responses, the relative frequency of this rating response will be higher on signal trials than on noise trials. And of course, the same arguments hold for noise-rating responses, which will also be necessarily unbiased.

If these stopping rule thresholds are close enough to the starting point of the walk, this unbiased decision rule will be suboptimal when the base rates are unequal. However, moving them away from the starting point will eventually ensure that whenever the process crosses the signal threshold, the conditional probability that the signal was presented will be greater than .5. In crude terms, in these response time models, the subjects' stopping rule may never allow them to terminate sampling and respond when the likelihood ratio is close to 1. Thus, Treisman's claim that the probability of a correct response for the lowest confidence signal responses in our experiments should have been closer to 0.1 is valid only under the assumptions of detection theory (and in fact, only for some parameterizations of these models).

Sensitivity indices. After observing our plots of $U_{\mathrm{R}}(k)$ for data from experiments in which we varied the size differences between the stimuli to be discriminated, Jim Townsend called to our attention the fact that the "order"—or relative heights at each point—of these curves decreased consistently with increasing difficulty of the task for each subject. Following up on this empirical result and other results developed by Townsend on stochastic dominance (Townsend & Ashby, 1978, 1983), we discovered a result, initially from Feller (1968), which states that for any pair of independent, additive random variables—that is, whenever the noise and the signal distributions can be represented by two independent random variables $\mathbf{X}$ and $\mathbf{X} + \mathbf{Y}$, respectively—the area between their two cumulative distribution functions is equal to the mean of the *inserted variable*, $\mathbf{Y}$. Letting $\mathbf{X}$ and $\mathbf{X} + \mathbf{Y}$ be subjective confidence "explained" the consistency in the empirical results noted by Townsend and Ashby (1978, 1983): The area under the $U_{\mathrm{R}}(k)$ function is an estimate of the separation of the two observable confidence-rating distributions.

In addition to criticizing this approach to sensitivity testing, Treisman (2002) offers some interesting insights about this mathematical relationship. Because it is not a critical issue, given our other results and the theoretical analyses they led us to, we will only briefly address his claims in general terms. First, it is important to recognize that his informal, geometric proof does not work as well as Feller's (1968) formal, mathematical proof. For example, Treisman's arguments do not explain why the result holds for the presumably important case in which **X** and **Y** are normally distributed; yet Feller's proof does (as well as for any other continuous distribution models).

More important, perhaps, is the practical issue about the effectiveness of the measure we proposed in behavioral applications. All of the concerns raised by Treisman (2002)—which were again based on specific examples derived from signal detection theory—were already discussed in Balakrishnan (1998a). We based our arguments about the utility of this measure on its performance with empirical data—that is, the probability that it would correctly order the sensitivity levels of two experimental conditions, as compared with the probability that other measures would correctly order these conditions (the *resampling method*). On this basis, our measure was more reliable than d' and the more than 20 other proposed measures of sensitivity with which we compared it. If an application requires a more exact statement about the relationships between stimuli and performance, investigators may need to fit a parametric model to the data to obtain this kind of description. However, from our point of view, the fact that quantitative models not only *may be* incorrect, but probably should be *expected to be* incorrect ought to raise some issues about the meaningfulness of these kinds of parametric analyses.

Summary and Conclusions

Treisman's article illustrates some important principles that investigators ought to be aware of before applying the methods and tests developed in our previous papers. First, the way in which subjects map their unobservable, subjective states to observable confidence ratings (e.g., the spacing of criteria on a decision axis, if we adopt the detection theory framework) could cause the measures we developed to be uninformative. They also require relatively large sample sizes and special instructions about how often subjects should use the extreme responses on the confidence-rating scale. Treisman is incorrect, however, in suggesting that detection models can fit our data by manipulating the spacing of the criteria on the decision axis. They also fail to predict the observed effect of base rates on the shapes of the ROC curve, which suggest precisely the distribution model we proposed to account for the lack of decision rule bias—that is, tradeoffs in the variances of the distributions with increasing bias.

If Treisman had attempted to fit the data, he would have discovered these problems himself. His arguments—especially his claim that there is no sensible way to explain an effect of bias on the shapes of the distributions—overlook the many possible alternatives to detection theory developed already in the literature, including some that many investigators would consider more plausible than detection theory (i.e., models that assume that subjects sometimes trade off accuracy for speed).

Finally, if there is an alternative version of detection theory that can fit the data, it will need to explain not only the empirical ROC curves, but also the fact that we were able to detect suboptimality of the decision rule when the base rates were unequal. As far as we can tell, all of Treisman's arguments about the power of the bias test apply equally to the power of the suboptimality test, which was satisfied consistently in our experiments. As we pointed out previously (Balakrishnan, 1998b, 1999; Balakrishnan & MacDonald, 2001a), simply suggesting that the confidence-rating data are entirely irrelevant or too contaminated by irrelevant factors to identify the bias in the decision rule from low-confidence noise responses does not explain why they were relevant enough to detect the suboptimality of the decision rule from low-confidence signal responses. Many other potential alternative accounts may also be ruled out on this basis alone. For example, the fact that we combined over subjects in some experiments and over multiple sessions in others, which normally would raise many legitimate concerns about the power of the tests (see, e.g., Maddox, 1999), presumably would make it difficult to detect both bias and suboptimality, not just one of these properties. Until a detection model can be developed that accounts for both of these test results, it seems inappropriate to us for investigators to continue to promote the detection theory models—or at least, any version of these models whose fit to the data suggests that decision criteria shift.

REFERENCES

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 & 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). Misrepresentations of signal detection theory and an alternative approach to human image classification. *Journal of Electronic Imaging*, **10**, 376-384.

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 & Performance*, **24**, 929-945.

Creelman, C. D., & Donaldson, W. (1968). ROC curves for discrimination of linear extent. *Journal of Experimental Psychology*, **77**, 514-516.

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

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol. 2, 3rd ed.). New York: Wiley.

Green, D. M., & Luce, R. D. (1974). Variability of magnitude estimates: A timing theory analysis. *Perception & Psychophysics*, **15**, 291-300.

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

Jiang, B. (2000). A generalized Gaussian SPRT model for two alternative forced-choice tasks. *Dissertation Abstracts International: Section B. The Sciences & Engineering*, **60**, 5803.

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.

Luce, R. D., & Green, D. M. (1978). Two tests of a neural attention hypothesis for auditory psychophysics. *Perception & Psychophysics*, **23**, 363-371.

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

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, **61**, 354-374.

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

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

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

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

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), *Cognitive theory* (Vol. 3). Hillsdale, NJ: Erlbaum.

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

Treisman, M. (2002). Is signal detection theory fundamentally flawed? A response to Balakrishnan (1998a, 1998b, 1999). *Psychonomic Bulletin & Review*.

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

(Manuscript received January 29, 2002;
revision accepted for publication July 10, 2002.)