Journal of Experimental Psychology:
Human Perception and Performance
1996, Vol. 22, No. 3, 615–633

0096-1523/96/$3.00

Testing Models of Decision Making Using Confidence Ratings in Classification

J. D. Balakrishnan
Purdue University

Roger Ratcliff
Northwestern University

Classification implies decision making (or response selection) of some kind. Studying the decision process using a traditional signal detection theory analysis is difficult for two reasons: (a) The model makes a strong assumption about the encoding process (normal noise), and (b) the two most popular decision models, optimal and distance-from-criterion models, can mimic each other's predictions about performance level. In this article, the authors show that by analyzing certain distributional properties of confidence ratings, a researcher can determine whether the decision process is optimal, without knowing the form of the encoding distributions. Empirical results are reported for three types of experiments: recognition memory, perceptual discrimination, and perceptual categorization. In each case, the data strongly favored the distance-from-criterion model over the optimal model.

To predict behavior, a theory of perception must include a decision process (e.g., a mapping or a rule) that ties internal perceptual effects of the stimuli to observable responses. Empirically studying the decisional elements of perception is difficult because virtually all of the special properties of a data set can, in principle, be attributed to either encoding or decision-making aspects of performance. In most quantitative models, the solution to this problem is to adopt a simple decision rule (e.g., the criteria setting in signal detection theory) and allow this rule to introduce additional free parameters into the model. The estimates of these parameters are used to isolate the contribution of the decision-making process, and the overall fit of the model to the data is the measure of its validity. The drawback of this approach, of course, is that the decision model must be tested in conjunction with the encoding model.

For experimental domains involving classification of some kind (e.g., discrimination, identification, or recognition), there are two main classes of decision models: distance-from-criterion models and optimal models. Distance-from-criterion models assume that an observer divides the different perceptual states induced by the stimuli into nonoverlapping subsets and associates each of these subsets with a different response. Thus, the response is based directly on the perceptual effect of the stimulus. Optimal models assume that the observer is able to transform perceptual information into a likelihood statistic, which tells which response is most likely to be correct for a given perceptual state. Using this statistical information, the observer can meet any performance objective (accuracy level or payoffs) that does not exceed the capacity of his or her encoding process. In contrast, the distance-from-criterion rule may or may not allow the observer to maximize performance level, depending on whether or not the response regions the observer chooses happen to match those of the optimal decision rule.

There is some empirical evidence that the low-level types of decision making involved in many perception experiments may be at least highly sophisticated, if not entirely optimal. For example, some of the early ideal observer models of signal detection theory, that is, an optimal decision process attached to a physical model of sensory transduction, can provide very good fits to standard psychophysical data, if certain aspects of the stimuli (signal phase or frequency) are assumed to be unavailable to the decision maker (e.g., Geisler, 1989; Green & Swets, 1966). More recently, Ashby and Maddox (1992) showed that participants can learn extremely complex nonverbal decision rules in a perceptual categorization task, if the advantage of these rules (i.e., the potential gain in performance level) is sufficiently large. Finally, in a series of recent articles, Glanzer and colleagues pointed out several strong regularities of recognition-memory performance that are predicted in a natural way by the optimal decision rule but not by distance-from-criterion models (Glanzer & Adams, 1985, 1990; Glanzer, Adams, & Iverson, 1993).

Apart from these results, most of the arguments are theoretical. To many researchers, an optimal decision rule seems implausible because it assumes that a participant has

J. D. Balakrishnan, Department of Psychological Sciences, Purdue University; Roger Ratcliff, Department of Psychology, Northwestern University.

This research was supported in part by Grants MH44640 and MHK00871 from the National Institute of Mental Health.

We thank F. Gregory Ashby, Mark Chappell, and Trisha Van Zandt for their comments on an earlier version of this article.

Correspondence concerning this article should be addressed to J. D. Balakrishnan, Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907. Electronic mail may be sent via Internet to jdb@psych.purdue.edu.

perfect statistical knowledge about the effects of the stimuli on the perceptual system (e.g., the exact shapes and locations of the perceptual distributions). If these distributions were completely unknown to the participant, then a very large number of trials would be needed to estimate them with any precision (the tails of the distributions would be especially difficult to learn because the participant would have very few examples of these during the experiment). The counterargument is that accurate classification is crucial for the survival of any species and, therefore, would have every opportunity to reach the high level of development assumed by the optimal model. To explain how optimality can be achieved when the stimuli are unfamiliar to the participants, one can point to the learning that occurs early on in virtually all laboratory classification tasks: Refining of the decision model is at least as plausible an account of these results as improvement in encoding or perceptual aspects of the behavior would be.

In this article, we describe some empirical tests that can be used to study the decision-making process and determine whether it is optimal. The tests take advantage of certain relationships that exist between the nature of the decision-making process and the ability of the participant to predict whether a given response will turn out to be correct (i.e., response confidence). Applied to three different kinds of classification experiments—recognition, discrimination, and categorization—the data from all three suggest the same thing: Participants use a distance-from-criterion rule in perceptual classification and recognition judgments, rather than an optimal decision rule. All of the tests are based on the predictions of the optimal classification model about the distributions of the feeling of confidence across experimental conditions (rather than, say, the average confidence level). Because the statistical theory will be unfamiliar to many, the problem of modeling confidence judgments is first introduced in a brief review of signal detection theory.

Statistical Decision Rules and Classical Signal-Detection Theory

In the classical theory of signal detection (e.g., Green & Swets, 1966), the presentation of a stimulus is assumed to induce an information state in the perceptual system, which is uniquely identified by a single number. Typically, the size of this *strength value* represents the participant's measure of the stimulus on a pertinent physical dimension (i.e., the judgment dimension). This measurement activity is noisy, with significant consequences for the participant. Instead of determining which stimulus objects cause which strength effects, the participant now must decide which objects are more likely to cause which strength effects. Therefore, likelihood is a fundamental concept of the theory.

In a yes–no detection or discrimination task, the model assumes that the participant creates a fixed, exhaustive map from strength values to responses. A crucial, but often neglected, issue is how the participant chooses this decision rule (i.e., how the participant decides which strength values should be assigned to which responses). The decision rule cannot be chosen arbitrarily, because this would cause the performance level (e.g., percent correct) to be arbitrary (i.e., performance level would be independent of the stimuli). The simple fact that physical properties of the stimuli (e.g., their physical similarity) strongly affect performance therefore implies that participants know (or guess wisely) some important facts about the relationship between strength effects and their most probable source. Signal detection theory is mute about what this knowledge might be.

Optimal Classification Models

A natural hypothesis to consider, of course, is that the participant always knows which of the two stimuli is more likely to have caused the incident strength effect on a given trial. Letting A and B be the two candidate stimuli and S be the strength effect of the presented stimulus, this hypothesis is equivalent to assuming that the participant computes the *stimulus likelihood ratio,*

$$L = \frac{P(\mathrm{B} \mid S = t)}{P(\mathrm{A} \mid S = t)} = \frac{f_{\mathrm{B}}(t)\mathrm{P}_{\mathrm{B}}}{f_{\mathrm{A}}(t)\mathrm{P}_{\mathrm{A}}},$$

where t is the observed value of S on the given trial; $f_{\mathrm{A}}(t)$ and $f_{\mathrm{B}}(t)$ denote the distributions (probability density functions) of S for Stimuli A and B, respectively; and P_{A} and P_{B} are the a priori probabilities (relative frequencies) of these stimuli in the experiment. The participant responds A whenever L is less than 1 and B whenever L is greater than 1. (The event $L = 1$ indicates indifference, and the response can be assigned randomly.)

By using this rule, the decision process always follows the betting odds—the response is emitted that has the best chance of being correct. Following the betting odds maximizes the expected percentage of correct responses, earning this decision rule the title "optimal classifier model." If the objective is to maximize expected payoffs rather than accuracy, and the payoff matrix is not symmetric, then the appropriate cutoff between A and B responses would not equal 1, but the decision rule would otherwise remain the same. To make optimal use of the information given to them by their senses, then, the participants must be able to compute L from S on each trial, which requires perfect knowledge about the probability distributions (which change with the stimuli) and the relative frequencies of the stimuli (P_{A} and P_{B}) during the experiment. Note that because L is computed from the observed value of S on each trial, it is also a random variable. Separating the trials by the two stimulus conditions, the four psychological variables in this model are S_{A}, S_{B}, L_{A}, and L_{B}.

Heuristic Decision Models

If the participants do not have this exact knowledge about the distributions of the strength values, then for the reasons explained above, they must have at least partial knowledge of them. One example of partial knowledge of the sort required would be that participants know that the distributions are unimodal or bell-shaped and can locate (or esti-

mate reasonably well) their mean values. In this case, the participants could adopt the heuristic strategy of placing a threshold on the strength values somewhere between the means of the two distributions. That is, the participants choose a value T, and if the observed value of S is less than T, then Stimulus A is chosen, and if the observed value of S is greater than T, then Stimulus B is chosen.

The value of this strength threshold would be the familiar criterion value of signal detection theory. If L is an increasing function of the value of S, then this threshold model will be mathematically equivalent to the optimal classifier model for some value of the ratio P_B/P_A. For this reason, letting this ratio be a free parameter (i.e., allowing for response bias or an asymmetric payoff matrix) seems to make the models empirically indistinguishable. In fact, we show that the two models can be empirically distinguished, without knowing whether or not L increases with S.

Modeling Confidence in a Classification Response

The heuristic and optimal models are special cases of the following more general model: (a) The participant computes a perceived stimulus likelihood ratio (i.e., the participant decides which stimulus seems more likely to be correct and to what degree), and (b) the response that this perceived likelihood ratio favors is frequently (or always) the same response that the true or objective likelihood ratio given by L favors. Thus, the psychological construct of most importance is not the strength effect but rather this perceived stimulus likelihood ratio. Because it must be based in some way on the strength effect, this psychological value will vary from trial to trial, even if the stimulus does not (i.e., it is also a random variable).

To illustrate, suppose that the perceptual effect S is equal to some value t on a given trial. Using some unspecified function or rule, the participant computes the probability that Stimulus A will cause the event $S = t$ and the probability that Stimulus B will cause this event. The perceived stimulus likelihood ratio becomes the ratio of these two values. Because the value of S changes from trial to trial, so does the perceived likelihood ratio value. Stated formally, the idea is that the participant computes a ratio of perceived conditional stimulus probabilities,

$$E = \frac{\hat{P}(B \mid S = t)}{\hat{P}(A \mid S = t)},$$

and sets a criterion on this value. That is, when E is less than some value T, then the response is A, otherwise the response is B. If T is not equal to 1, then this would indicate that the participant sometimes intends to choose a response that he or she believes is less likely to be correct (because by definition, all values of E less than 1 indicate higher confidence that the stimulus is an A, and all values of E greater than 1 indicate higher confidence that the stimulus is a B). Separating the trials by stimulus conditions, the perceived likelihood ratio variables are E_A and E_B.

The fundamental assumption involved in all of the empirical tests to be described here is that this perceived likelihood ratio, E, can be studied empirically—in effect, it is an alias for the participant's feeling of confidence in his or her response choice. When the participant rates his or her confidence level on an integer scale ranging from 1 (*most confident A*) to n (*most confident B*), we assume that the possible values of E are divided by the participant into n contiguous regions and labeled with increasing integer values from 1 to n. The rating response depends on which of these response bins the actual value of E falls into. (The partition could change from trial to trial without affecting the empirical tests that we propose. What is important is that this partition is independent of the value of E on a given trial, or not dependent in such a way that larger rating values do not imply larger E values.)

Letting R represent a bipolar rating response on a given trial (i.e., small values of R represent high-confidence A responses and large values of B represent high-confidence B responses), the statistical measure of most interest will be the cumulative frequency distributions of R under the two different stimulus conditions. That is, the experimenter can estimate $P(R_A \leq K)$ and $P(R_B \leq K)$ for each K. These estimates are useful because of the mapping relationship assumed between R and E. Specifically,

$$P(R \leq k) = P(E \leq C_k)$$

and hence

$$P(R_A \leq k) < P(R_B \leq k) \rightarrow P(E_A \leq C_k) < P(E_B \leq C_k),$$

$$P(R_A \leq k) = P(R_B \leq k) \rightarrow P(E_A \leq C_k = P(E_B \leq C)_k,$$

and

$$P(R_A \leq k) > P(R_B \leq k) \rightarrow P(E_A \leq C_k) > P(E_B \leq C_k).$$

Most of the tests of the decision models described below are based on their predictions about these distributional inequalities, or dominance patterns.

Of course, the same assumption about the mapping between confidence and confidence ratings is required in signal detection theory analyses of confidence ratings. To make quantitative estimates of sensitivity and bias possible, however, signal detection theory adds the assumptions that (a) the mapping from confidence to confidence ratings is constant across trials (otherwise the strength distributions, and hence perceptual sensitivity, cannot be estimated), (b) the strength distributions are normal, and (c) the E values are monotone transformations of S. The reasons for the third assumption are discussed in the next section.

Dependence of Perceived Likelihood Ratio on the Decision Rule

To study the decision-making process more directly, that is, without making the extra assumptions of signal detection

theory, the predictions of the decision models about confidence must be determined. For the optimal classifier, the relationship is simple: Confidence and objective stimulus likelihood ratio are equivalent. Several important tests easily follow from this basic identity. For the heuristic model, predictions about confidence come from some arguments about how the participant chooses the criterion (T) that divides the strength effects into the two types of responses. In a signal detection theory analysis of confidence ratings, the assumption is that to make the confidence judgments, the participant in effect sets more than one criterion on strength. The more general idea is illustrated as follows: First, note that as the relative frequency of, say, Stimulus A increases, the proportion of A responses given by a participant on B trials will increase (e.g., Green & Swets, 1966). Because the only decision parameter in the heuristic model is the criterion, this empirical result is explained by assuming that the relative frequencies of the stimuli cause the participant to move the criterion in one direction or the other, increasing the range of values for which the more frequent stimulus is seen as more likely. If the participant moves the criterion more and more to the right as, say, Stimulus A is more and more likely to be presented and more and more to the left as Stimulus B is more and more likely to be presented, then this implies that E increases with the signed distance between S and the criterion. The model is

$$E = g(S - T),$$

where $g(.)$ is an increasing function of its argument.

Thus, the heuristic model is really a special case of a distance-from-criterion model of confidence. As noted above, the traditional signal detection theory analysis of ratings depends on the assumption that $g(.)$ is an increasing function. If it is not, then the receiver operating characteristic (ROC) curve does not measure the perceptual effects distributions but rather another pair of distributions that depend in a fairly complex way on both perceptual and decisional processes.

Testable Predictions of the Optimal Decision Model

Objective Certainty Test

One fundamental prediction of the optimal classifier is implicit in the definition of E, that is, that $E = L$ for all values of S. If this representation of confidence is correct, then the subjective feeling of confidence will be perfectly correlated with the objective probability that the response will be correct. Among other things, this means that the proportion of correct responses should always increase with the reported level of confidence in the emitted response.

A simple test of this prediction is to plot the proportion of times that Stimulus A was presented when the rating response, R, was equal to k, for each k. That is, the value $P(\mathrm{A} \mid R = k)$ is plotted on the ordinate, against k values on the abscissa. If the participant is using the optimal decision rule, then this function should be nondecreasing (it is strictly increasing if there are more values of E than of R). If this empirical function is not monotone, then the optimal model can be immediately rejected.

The distance-from-criterion model may or may not predict this objective certainty property, depending on whether the objective likelihood ratio, L, is an increasing function of S. For example, the unequal variance, normal model of signal detection theory predicts that this function will not be monotone increasing but instead will be U-shaped. (However, unless the difference in variances is large relative to the difference in the means, the decreasing portion may be too far into the tails of the distributions to be empirically detectable.)

Stochastic Dominance Tests

If small values of E indicate high confidence that the stimulus was an A, then it seems reasonable to expect E to be small more often when Stimulus A is presented than when Stimulus B is presented. If this were not true, in fact, then the participant could improve his or her performance by responding B when E was very small, even though small E values represent high confidence that the stimulus was an A. Improving performance is not possible if $E = L$, and thus the optimal model must make some strong predictions about the distributions of E_A and E_B.

The general result is that the optimal model always predicts

$$F_{E_A}(t) \geq F_{E_B}(t)$$

for each value of t, where F denotes the cumulative distribution function. This leads to the following empirically testable prediction:

> *Distribution dominance property:* If the participant is using the optimal decision rule, then
>
> $$P(R_A \leq k) \geq P(R_B \leq k)$$
>
> for all rating responses k. (The proof is given in the Appendix.)

Thus, the optimal classifier predicts that the cumulative frequency distribution of the rating responses when Stimulus A is presented will always be greater than or equal to the cumulative frequency distribution of the rating responses when Stimulus B is presented. If this prediction is violated, then the optimal decision rule can be rejected. As before, the distance-from-criterion model may or may not make this prediction about the data, depending on what properties are assumed about the strength distributions (see the Appendix).

Stronger Forms of Stochastic Dominance

The ordering of two cumulative distribution functions is one of several possible forms of *stochastic dominance,* that is, a way of saying that one random variable "tends to be smaller" than another (e.g., Townsend, 1990; Townsend & Ashby, 1983). Some of these forms of dominance are stron-

ger than others. For example, an ordering of the cumulative distributions is a stronger form of dominance than an ordering of the means, because it implies the mean ordering, whereas an ordering of the means does not imply an ordering of the cumulative distribution functions.

The optimal classifier also predicts two additional forms of dominance that are even stronger than an ordering of the cumulative distributions. One of these is based on the so-called *hazard rate function,*

$$\frac{f(t)}{1 - F(t)}.$$

Theoretically, the cumulative distribution can be recovered exactly from the hazard rate function, and vice versa; there is no new information nor any loss of information in this redefinition of the distribution. However, empirically estimating hazard rate functions has some special advantages when a researcher wishes to identify the correct distributional form for an empirical measure (Luce, 1986).

The optimal model predicts that the hazard rate functions will also be ordered, that is:

$$\frac{f_{E_A}(t)}{1 - F_{E_A}(t)} \geq \frac{f_{E_B}(t)}{1 - F_{E_B}(t)}.$$

The other dominance prediction is between the *convexity functions,* that is:

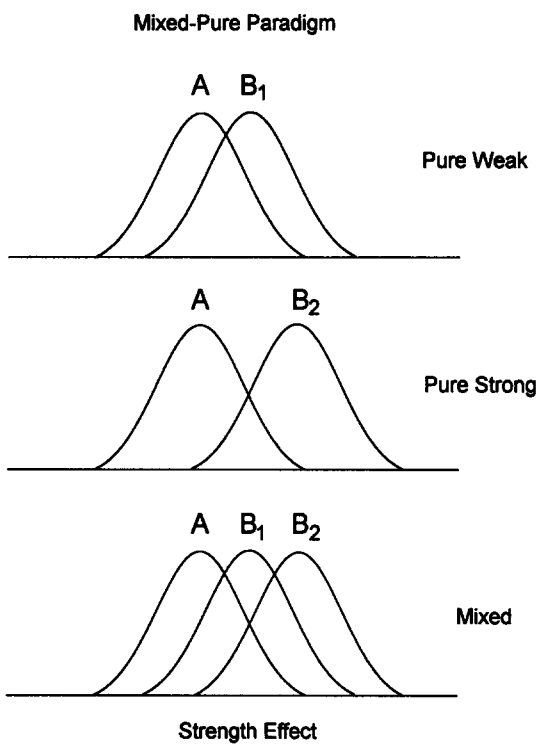

Figure 1. Strength effect distributions illustrating the three conditions of the mixed–pure paradigm. The means of the distributions represent the physical sizes of the stimuli.

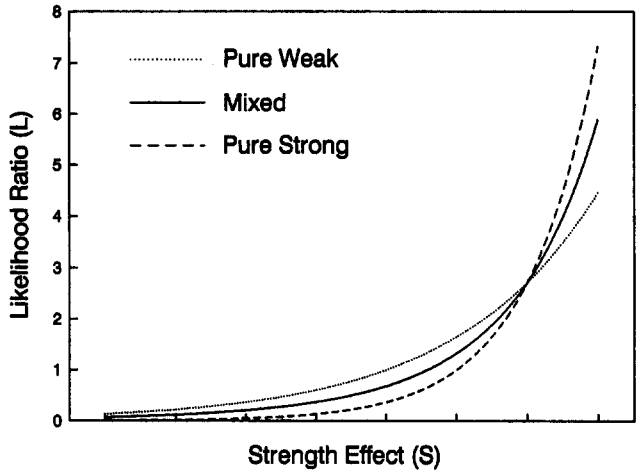

Figure 2. Likelihood ratio functions corresponding to the three conditions of the mixed–pure paradigm represented in Figure 1. Values less than one indicate greater likelihood that the strength effect is caused by the A stimulus, and values greater than one indicate greater likelihood that the strength effect is caused by the B stimulus.

$$\frac{f_{E_B}(t)}{F_{E_B}(t)} \geq \frac{f_{E_A}(t)}{F_{E_A}(t)}.$$

These functions also have some special theoretical significance in distributional analyses of empirical models (Balakrishnan, 1994; Dzhafarov & Rouder, in press).

To test whether the hazard rate functions are ordered, the experimenter can plot the ratio of *survivor functions* (1 minus the cumulative distribution function),

$$U(k) = \frac{1 - H_B(k)}{1 - H_A(k)},$$

where $H_A(k)$ and $H_B(k)$ are the cumulative frequency distributions of the rating responses for the A and B stimulus conditions, respectively (we used H instead of F to emphasize the fact that the rating responses are observable). If the hazard rate dominance property is satisfied, then this empirical function must be nondecreasing for all k. Similarly, to test for order in the convexity function, the experimenter plots the value

$$V(k) = \frac{H_B(k)}{H_A(k)},$$

against k.

If these stronger forms of dominance are satisfied, then the empirical functions will be increasing. (Proofs of these results are given in the Appendix.) As before, the distance-from-criterion model may or may not make these dominance predictions, depending on the properties of the strength distribution functions.

All three of these dominance properties, as well as the objective certainty prediction of the optimal model, are strongly supported by the empirical data from recognition-

memory, discrimination, and categorization experiments reported below. In this respect, the optimal model predicts a large and theoretically important set of empirical relationships in two choice classification tasks, when the different effects of the stimuli on the feeling of confidence are compared (i.e., in AB comparisons). In the next section, however, additional distributional tests are developed that will allow us to rule out the optimal decision model.

Mixed–Pure Paradigm

To show that participants do not use the optimal decision rule, the empirical tests described next add a third stimulus to the classification task. In this mixed–pure paradigm (Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990), the two permissible responses are still A and B, but the B stimulus is now varied between conditions. Sometimes it is a relatively weak stimulus and other times a relatively strong stimulus. For example, if the stimuli are lines differing in length, then Stimulus A is the shortest line, the weak Stimulus B (hereinafter B_1) is longer than Stimulus A, and the strong Stimulus B (hereinafter B_2) is longer than B_1.

The participant's task is to respond A when Stimulus A is presented and B when either B_1 or B_2 is presented. Sometimes B_1 and A are the only two stimuli in a given block of trials (the pure weak condition), sometimes B_2 and A are the stimuli for the block (the pure strong condition), and sometimes B_1 and B_2 occur in equal proportions (i.e., 25% of trials), while Stimulus A still occurs on half the trials in a block (the mixed condition).

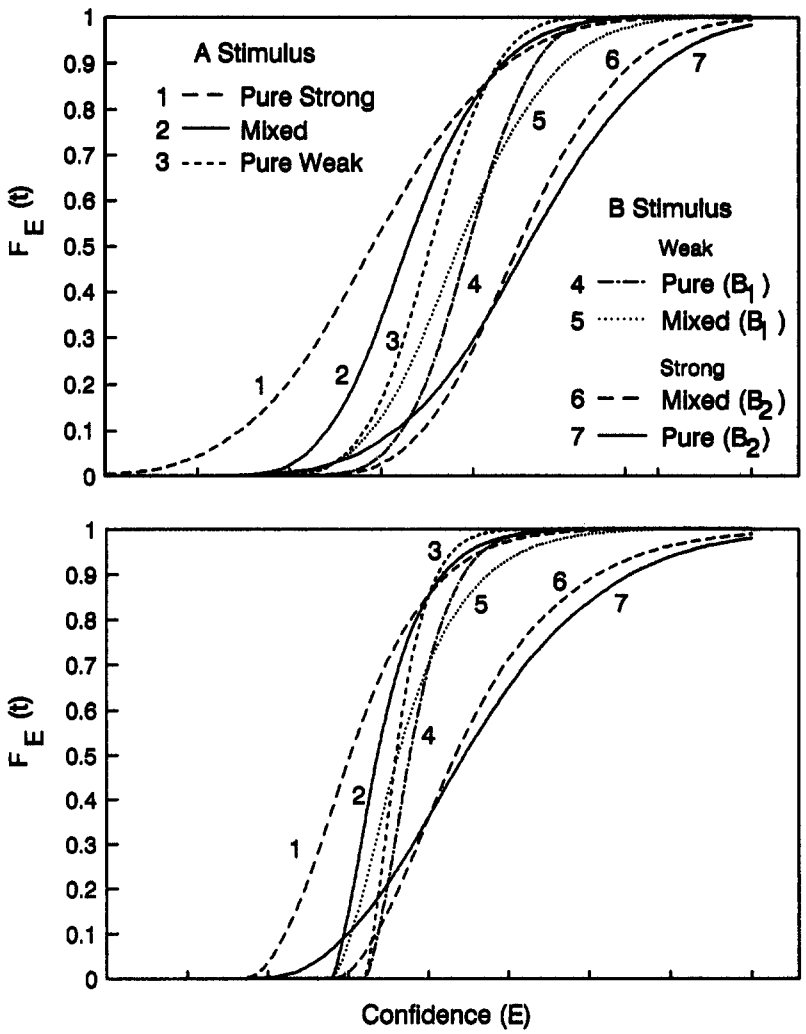

Figure 3. Distribution crossover predictions of the likelihood ratio model. The functions in the upper panel represent confidence distributions resulting from normal, equal variance strength effects and an optimal decision rule, and the functions in the lower panel represent normal distributions of strength and variance increasing with the mean.

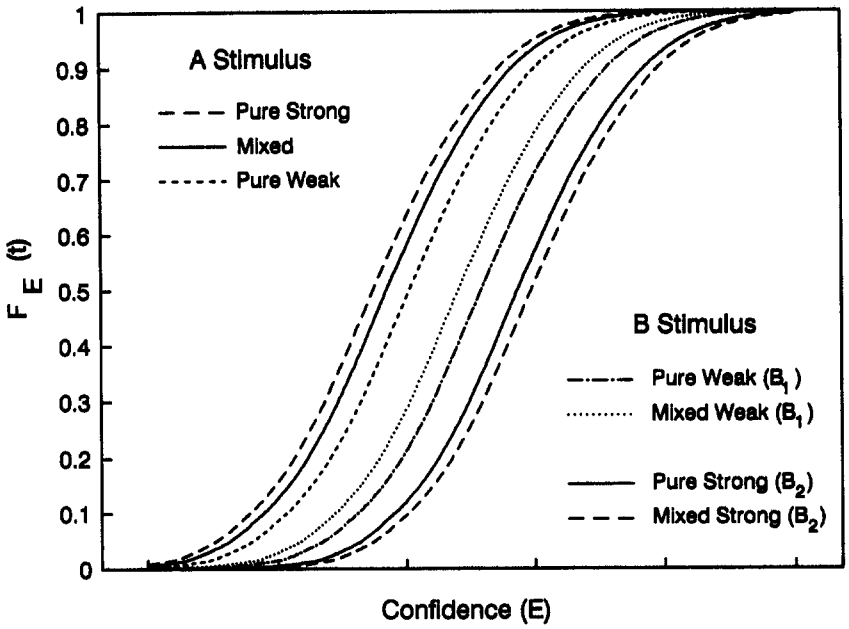

Figure 4. Confidence distributions predicted by the distance-from-criterion model when the same strength distributions as in the upper panel of Figure 3 are assumed, and the criterion separating A from B responses is placed where the distributions intersect (i.e., optimally).

Predictions of the Optimal Classifier

The three conditions of the mixed–pure experiment are illustrated graphically in Figure 1, using normal distributions to represent the strength effects of the stimuli. Note that the optimal classifier computes E differently in all three conditions, because Stimulus B is defined differently in each of them. In the pure weak (PW) condition,

$$E_{PW} = \frac{f_B(t)P_{B_1}}{f_A(t)P_A},$$

whereas in the pure strong (PS) condition,

$$E_{PS} = \frac{f_{B_2}(t)P_{B_2}}{f_A(t)P_A}.$$

The same kind of formula defines E for the mixed (M) condition, the only difference being that the strength distribution of Stimulus B is a (weighted) average of the distributions of B_1 and B_2. That is,

$$E_M = \frac{f_{B_1}(t)P_{B_1} + f_{B_2}(t)P_{B_2}}{f_A(t)P_A}.$$

The empirical tests to be described next take advantage of the fact that the strength effect distribution depends only on the stimulus that is in fact presented, whereas the decision rule of the optimal classifier depends on the two stimuli that may or may not be presented on a given trial. Suppose, for example, that the cumulative frequency distributions of the Rating Response R when Stimulus A was presented in the pure weak condition is compared with the cumulative frequency distributions of R when Stimulus A is presented in the pure strong condition. Because the transformation from S to R is different in the two conditions, the predicted cumulative frequency distributions of R are also different,

Table 1
A Summary of the Distributional Predictions of the Optimal and Distance-From-Criterion Models of the Decision Process

	AB comparisons				AA comparisons	BB comparisons	
Model	$P(A \mid R = k)$ increasing	$H_A \geq H_B$	$(1 - H_B)/(1 - H_A)$ increasing	H_B/H_A increasing	$H_{A(PS)} \geq H_{A(M)} \geq H_{A(PW)}$	$H_{B_2(PS)} \geq H_{B_2(M)}$	$H_{B_1(M)} \geq H_{B_1(PW)}$
Optimal classifier	Yes	Yes	Yes	Yes	No	No	No
Distance-from-criterion	Yes or no	Yes or no	Yes or no	Yes or no	Yes	Yes	Yes

Note. In the "yes or no" cases, the prediction of the distance-from-criterion model depends on the distribution model for the strength effects of the stimuli.

Table 2
Hit and False-Alarm Rates for Recognition-Memory Data

Condition	Hit	False alarm
Pure weak	.58	.39
Pure moderate	.70	.20
Pure strong	.79	.16
Mixed: Moderate + weak		
Weak (B_1)	.59	
Moderate (B_2)	.72	
A		.20
Mixed: Strong + weak		
Weak (B_1)	.56	
Strong (B_2)	.81	
A		.18

Note. The data were combined across participants. Hit = responded "yes" to an old item; false alarm = responded "yes" to a new item.

even though both of them represent responses to the same physical stimulus. In fact, the optimal classifier predicts that instead of being identical or ordered, these two cumulative frequency distributions will cross over at some point. The general result is that stochastic dominance should be violated in all AA or BB comparisons within the mixed–pure paradigm.

To see why these violations of dominance are predicted by the model, consider the likelihood ratio functions in Figure 2, which correspond to the strength distributions in Figure 1 for the three conditions of the experiment (pure weak, pure strong, and mixed). The abscissa represents the strength effect (S), and the ordinate represents the value of confidence that it produces (E). Because the height of these functions is equal to E, the model predicts that the same value of S does not produce the same level of confidence in the different conditions. A very small value of S in the pure weak condition, for example, should cause the participants to have less confidence than the same value of S when it occurs in the mixed condition (because the pure weak likelihood ratio function is closer to 1, or complete uncertainty, than the mixed likelihood ratio function).

It turns out that the arrangement and crossover of the likelihood ratio functions are reproduced (with the order pattern reversed) in the predictions of the optimal classifier model about the cumulative distribution functions of E, leading to the following empirical test.

Context sensitivity test: Assume that the likelihood ratio functions for the pure B_1 and pure B_2 conditions are monotone and that the B_1 and B_2 strength distributions intersect at Point w. The optimal classifier model predicts that

$$H_{A(PW)}(k) \leq H_{A(M)}(k) \leq H_{A(PS)}(k), k \leq r^*;$$
$$H_{A(PW)}(k) \geq H_{A(M)}(k) \geq H_{A(PS)}(k), k \geq r^*.$$

$$H_{B_1(PW)}(k) \leq H_{B_1(M)}(k), k \leq r^*;$$
$$H_{B_1(PW)}(k) \geq H_{B_1(M)}(k), k \geq r^*.$$

$$H_{B_2(PW)}(k) \geq H_{B_2(M)}(k), k \leq r^*;$$
$$H_{B_2(PW)}(k) \leq H_{B_2(M)}(k), k \geq r^*,$$

where H denotes the cumulative frequency distribution of the Rating Response R, the subscripts represent the condition (i.e., A = short line, B_1 = medium line, and B_2 = long line; P = pure, W = weak, S = strong, and M = mixed), and r^* is a rating-response value that depends on where the B_1 and B_2 strength distributions intersect, Point w).

The assumption that the likelihood ratio functions are monotone is not crucial for the crossover predictions of the model—a more general result is given in the Appendix. To illustrate the result for both assumptions (monotone and nonmonotone likelihood ratio functions), Figure 3 shows the cumulative distribution functions of E for the normal, equal variance strength model (i.e., the likelihood ratio functions are monotone; upper panel) and the normal, unequal variance strength model (i.e., the likelihood ratio functions are nonmonotone; lower panel).

Some other types of dominance violations predicted by the optimal model are also illustrated in Figure 3. Notice, for example, that the cumulative distribution functions of E when Stimulus B_1 is presented in the pure weak condition and when Stimulus B_2 is presented in the pure strong condition also cross over. Unfortunately, although we can give a proof of this prediction for strength distributions like those in Figure 1 (i.e., the A stimulus distribution is shifted to the right by some amount to obtain the B stimulus distributions) and for some general types of nonmonotone likelihood ratio functions, we do not have a general proof of it for the case of the monotone likelihood ratio models. Thus, we cannot prove that this particular test of the optimal decision model is as strong as the others.

Predictions of the Distance-From-Criterion Model

Recall that if the participant uses the distance-from-criterion rule, then in each condition, a criterion (T) must be

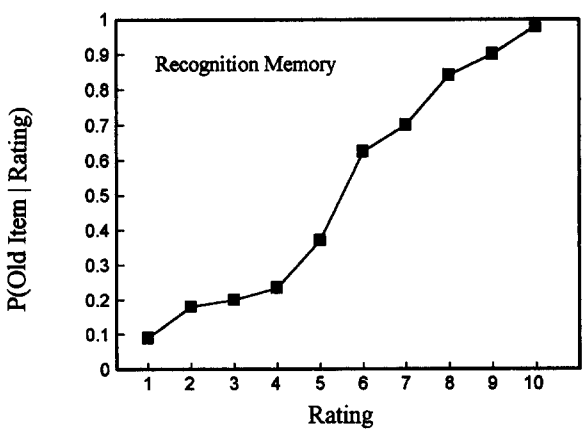

Figure 5. The probability that the test word is a studied item, given that the rating response is k, for each k (each possible confidence rating) in the recognition-memory experiment (data combined across participants and mixed–pure conditions). Note that the middle regions of the abscissa represent the *least confident* values (5 and 6) and the extremes the *most confident* values (1 and 10).

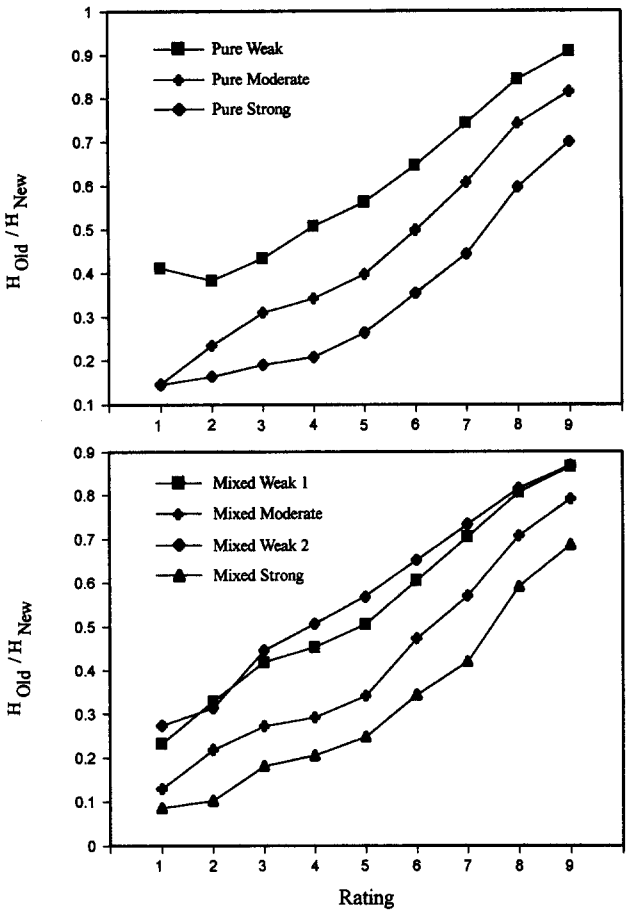

Figure 6. The ratio of cumulative frequency distributions of confidence under A versus B stimulus conditions (convexity test) for each condition of the recognition-memory experiment (data combined across participants).

chosen to divide the strength values into A and B responses. The perceived likelihood ratio, E, is then an increasing function of the signed distance of S from this criterion, that is, $E = g(S - T)$. If T and $g(.)$ are the same in all conditions, then obviously there would be no difference between the cumulative frequency distributions of the rating responses when the same stimulus is presented, for example, $H_{A(PW)}(k) = H_{A(M)}(k) = H_{A(PS)}(k)$, for all k. If the criterion is affected by the context but the $g(.)$ function is the same across conditions, then the cumulative distributions of E will simply shift in one direction or the other—shifting the distributions implies stochastic dominance between them. If the shift in the value of T is small, then the shift in the distributions will be small.

For the distance-from-criterion model to predict the dominance violations that the optimal model predicts, the $g(.)$ function would need to have a very special kind of dependence on the mixed–pure stimulus condition. Predictions of this model when $g(.)$ is a linear function and the response criterion is placed optimally (i.e., maximizing the percentage of correct choice responses) are shown in Figure 4 for the same strength distribution model used to illustrate the predictions of the optimal classifier (upper panel of Figure 3).

The complete set of empirical tests is summarized in Table 1. For the earlier group (the AB tests), the optimal model predicts that dominance will hold, whereas for the AA and BB comparisons involved in the context sensitivity test, the model predicts that dominance should be violated. The distance-from-criterion model does or does not predict dominance in the AB tests, depending on the strength distributions. For the AA and BB comparisons, however, it predicts that dominance should hold. The direction of the dominance (which distribution is larger) depends on the relative placements of the criteria (T). The pattern represented in Table 1 is based on the assumption that the criterion moves toward the location that maximizes percentage correct, which means that $T_{\text{pure weak}} < T_{\text{mixed}} < T_{\text{pure strong}}$.

The empirical data reported below do not exhibit the dominance violations predicted by the optimal model. The cumulative frequency distributions of the rating responses are not strongly affected by the stimulus condition. Their

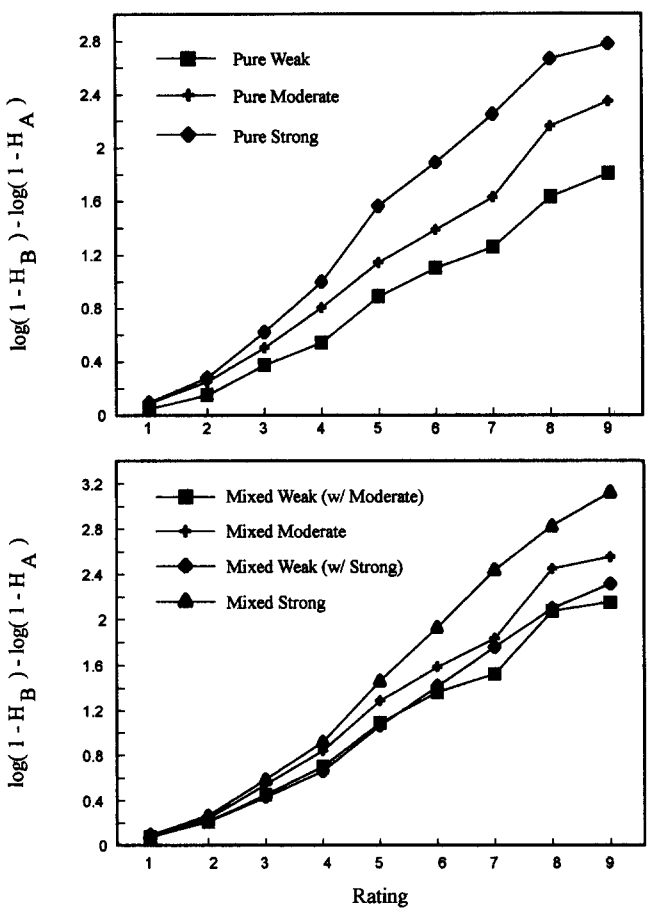

Figure 7. The log of the ratio of survivor functions (hazard rate dominance test) plotted against the Rating Response k for the different conditions of the recognition-memory experiment.

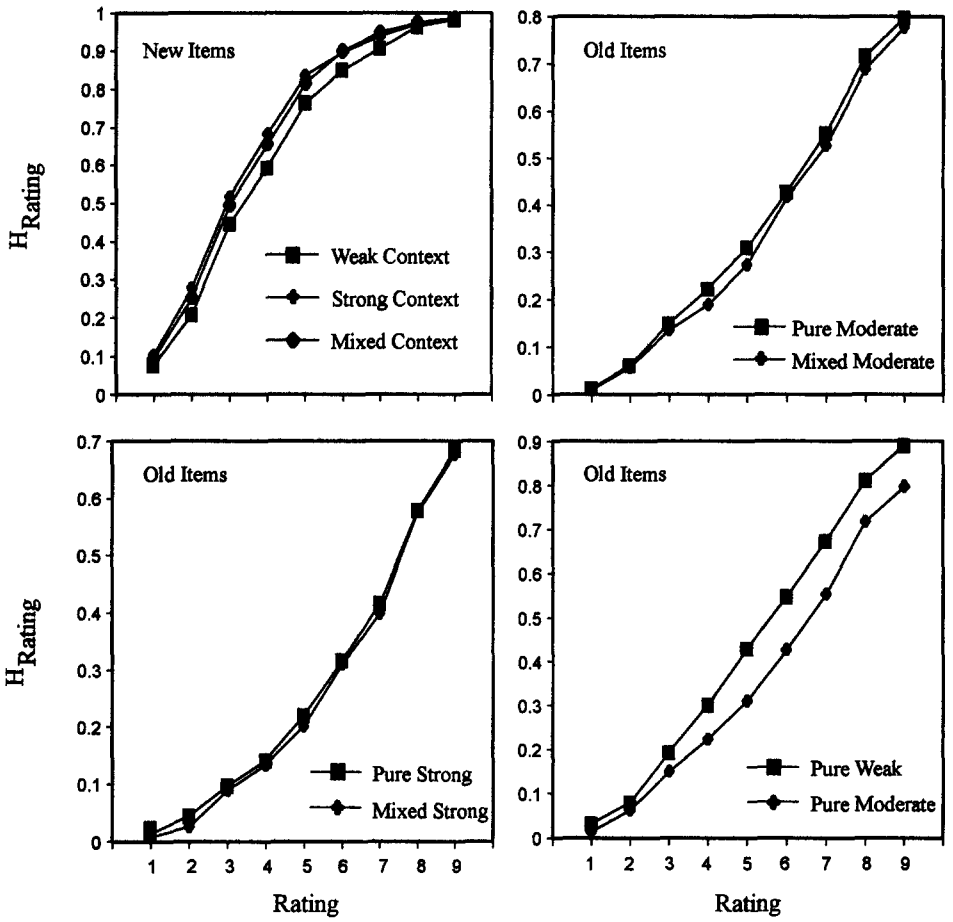

Figure 8. Cumulative frequency distributions of confidence ratings by condition in the recognition-memory experiment.

order pattern is instead consistent with a distance-from-criterion model in which the choice criterion value (T) changes by small amounts with the stimulus condition while $g(.)$ is the same across conditions. Thus, a simple version of the distance-from-criterion model provides an efficient summary of the results.

Experiment 1: Study–Test Recognition Memory

Method

To test the predictions of the optimal decision model for recognition memory, we used a version of the mixed–pure paradigm, with single-word stimuli and three levels of study strength. The five separate conditions were (a) pure weak, (b) pure moderate, (c) pure strong, (d) mixed: weak + moderate, and (e) mixed: weak + strong. For word recognition, the strength effect of the stimulus represents the familiarity effect of the item. This familiarity effect was manipulated by increasing the number of repetitions of an item within the study list.

Participants. Thirty students from an introductory psychology course at Northwestern University participated, in partial fulfillment of a course requirement. Individual sessions lasted about 50 min.

Procedure. Word stimuli were presented on CRT screens controlled by a PC. The three levels of the strength manipulation were one, two, and four repetitions for weak, moderate, and strong conditions, respectively. The total number of different items in each study list was held constant at 22, with the first and last 3 items excluded from sampling at test. All other studied items, plus an equal number of new items, were presented once at test. Each study item was presented for 250 ms, with 250 ms between items. The sequencing of the study items was random; however, no immediate repetitions of the same word were allowed. Time between test items was participant-determined (self-paced testing). Although the average number of items intervening between study and test of a given item was always the same, the average time between study and test varied for a given item type (see Murnane & Shiffrin, 1991, for a discussion of the effects of timing and sequencing in this kind of design). All items were randomly sampled from Kucera and Francis's (1967) word pool, with inclusion in the sample conditioned on length of the word (between 5 and 10 characters) and frequency (5 or more occurrences per million).

The confidence rating scale was defined by the numbers 1 to 10 on the top of the keyboard, with 1 indicating *most sure new* and 10 indicating *most sure old.* Attention was drawn to the fact that the cutoff between what would be considered new and old responses on the scale was between the 5 and 6 key responses. Participants

Table 3
Hit and False-Alarm Rates for Discrimination and Categorization Data

Condition	Participant 1 Hit	Participant 1 False alarm	Participant 2 Hit	Participant 2 False alarm
Discrimination				
Pure weak	.69	.33	.65	.32
Pure strong	.84	.22	.85	.26
Mixed				
Weak (B_1)	.66		.69	
Strong (B_2)	.92		.89	
A		.36		.41
Categorization (lines)				
Pure weak	.64	.32	.67	.45
Pure strong	.76	.24	.81	.36
Mixed				
Weak (B_1)	.54		.65	
Strong (B_2)	.80		.86	
A		.32		.41
Categorization (numbers)				
Pure weak	.48	.24	.67	.40
Pure strong	.68	.15	.88	.35
Mixed				
Weak (B_1)	.44		.68	
Strong (B_2)	.72		.89	
A		.23		.39

Note. Hit = responded "yes" to an old item; false alarm = responded "yes" to a new item.

were also asked to be conservative in their use of the extremes of the scale, so that differences between relatively high levels of confidence could be distinguished, if possible. The purpose of this instruction, which was successful (see below), was to increase the chances that the criteria would extend into the tails of the confidence distributions. Instructions emphasized accuracy of performance, and no pressure was induced on response time or on the total number of lists completed during a session.

Results and Discussion

Table 2 lists the percentages of correct old (or hits) and incorrect old (or false alarms) responses for the five different conditions of the experiment (pure weak, pure moderate, pure strong, mixed weak + moderate, mixed weak + strong). The pattern of results (e.g., relatively small effects of mixed versus pure conditions) is typical of recognition performance (e.g., Ratcliff et al., 1990). The analyses that follow are based on averages of the estimated functions for each participant and distribution test (aggregating in this way does not affect predictions about dominance). The total sample sizes ranged from 1,900 to 2,000 per stimulus condition.

Objective certainty test. In Figure 5, the probability that the stimulus is an old item, conditioned on the participants' rating response, is plotted for each rating category, with the data collapsed across participants and the five conditions. This function is clearly monotone increasing. At the highest level of confidence, the probability that the response is correct is roughly 90%. For the least confident response (the middle of the abscissa in Figure 5), the probability drops to about 60% (50% is the chance level). Thus, the confidence reports are very good predictors of accuracy.

Stochastic dominance: AB comparisons. The distribution dominance and convexity tests described above can be applied simultaneously by plotting the ratio of the cumulative frequency distributions, H_B/H_A. If this function is always between 0 and 1, then the distribution dominance property is satisfied; if it is also increasing, then the convexity dominance test is satisfied as well. Dominance of the hazard rate functions is tested by plotting the logarithm of the $U(k)$ function defined above (otherwise, the range of the ordinate axis is extremely large, making the shape of the functions difficult to identify). The estimated empirical functions are presented in Figures 6 and 7 for old versus new items. For each of the seven conditions, all of the dominance predictions of the optimal model are clearly supported. To make the same predictions about confidence, the distance-from-criterion model must assume that the strength distributions exhibit this dominance pattern (e.g., they are all normal with equal variances).

Context sensitivity test. Because there are many conditions and the conclusions are consistently the same, Figure 8 shows the results of the context sensitivity test for some representative cases. Recall that the optimal model predicts that when the cumulative frequency distributions are compared across conditions (e.g., mixed–pure) for the same stimulus type, crossovers should occur according to a specific pattern (see above). Instead of following any crossover pattern, a much better description of these results is that the functions are not very different and they are ordered. In the new item conditions (upper left panel of Figure 8), for example, the order is: pure weak ≤ mixed ≤ pure strong. This is the pattern predicted by the distance-from-criterion rule if the choice response criterion shifts by small amounts toward its optimal position for each condition. The bottom two panels compare the old item cumulative frequency

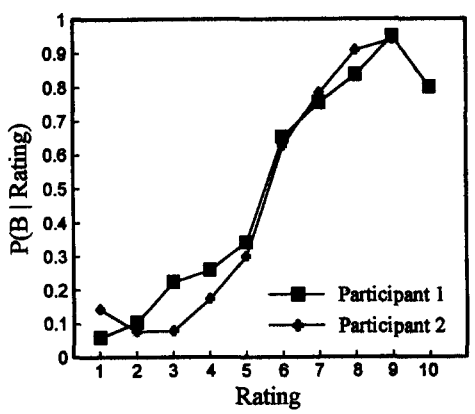

Figure 9. The proportion of trials that the stimulus was a B, when the rating response was k, for each rating category k in the categorization and discrimination experiments combined.

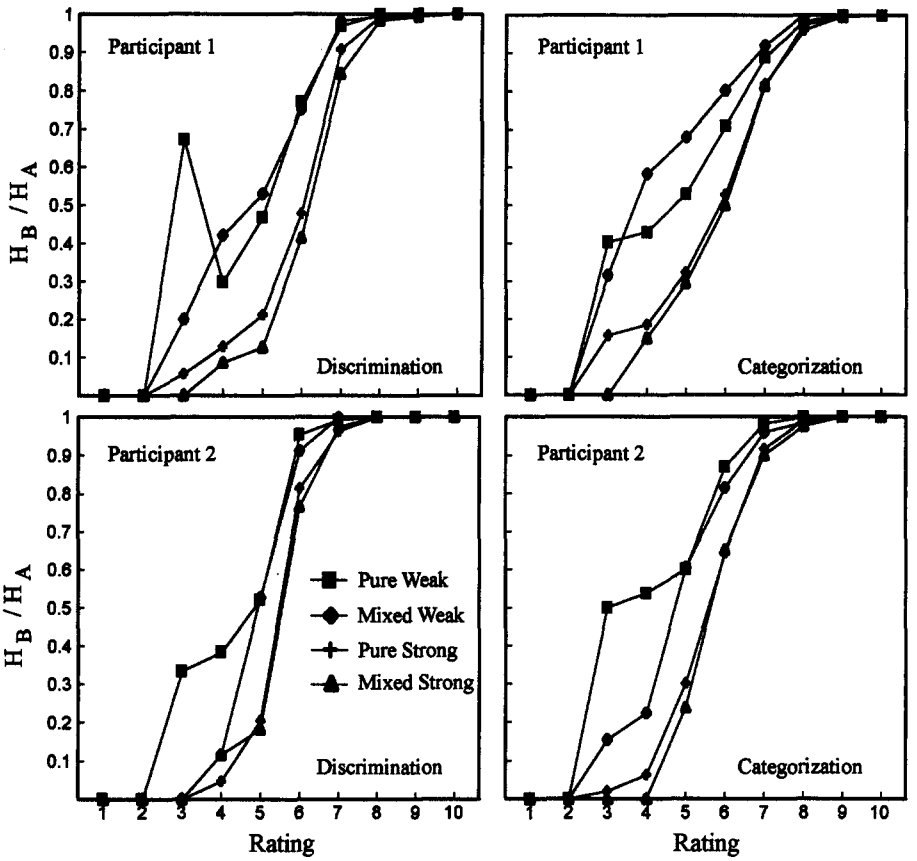

Figure 10. Ratios of cumulative frequency distributions of confidence for line-length discrimination and categorization experiments.

distributions for weak and moderate study items. Instead of crossing over, these are also strongly ordered.

To summarize, the data exhibit the strongest testable forms of stochastic dominance when A and B stimulus conditions are compared, but there is no evidence for the violations of dominance predicted by the optimal model when A-to-A and B-to-B comparisons are performed between conditions of the mixed–pure paradigm. The distance-from-criterion rule suggests an immediate and simple explanation: The $g(.)$ function above is constant across conditions, and the criterion (T) changes with condition.

Experiments 2 and 3: Discrimination and Categorization

Method

A similar mixed–pure design was used to apply the empirical tests to elementary perceptual tasks. Two participants were paid for their participation in 13 hr of total session time each, providing relatively large samples of single participant data that could be analyzed individually. There were three types of conditions, with the following order reversed for Participant 2: (a) line-length discrimination, (b) line-length categorization, and (c) number categorization.

The categorization tasks used the randomization technique developed recently by Ashby and colleagues (e.g., Ashby & Gott, 1988). In its application here, the computer was used to generate samples from distributions like those in Figure 1. The magnitude of the sample became the magnitude of the stimulus on a given trial. In the line-length categorization task, the magnitude of the sample was used to determine the length of a line presented on the screen, and in the number categorization task, the sample value itself was presented on the screen. In both cases, the participant was asked to decide which distribution the stimulus was sampled from, A or B.

In the line-length discrimination and categorization tasks, the stimuli were horizontal, single pixel lines presented on a liquid crystal diode (LCD) video display. Lengths of the lines in the discrimination task were 6.60 cm, 6.75 cm, and 6.90 cm, with a viewing distance of approximately 45 cm. Horizontal location of the lines was randomized, with the line appearing within a centered 15-cm region of the display. The response was a confidence rating on a scale from 1 to 10 (same as that in the recognition experiment), with keys 1 and 10 representing *highest confidence* short and long line length, respectively, and 5 and 6 representing *lowest confidence* short and long line length, respectively.

In the perceptual categorization task, length of the line was a random sample from a normal distribution with means equal to 6.7 cm, 7.1 cm, and 7.6 cm for the three stimulus conditions. Standard deviation was a constant 0.7 cm. The same distribution models

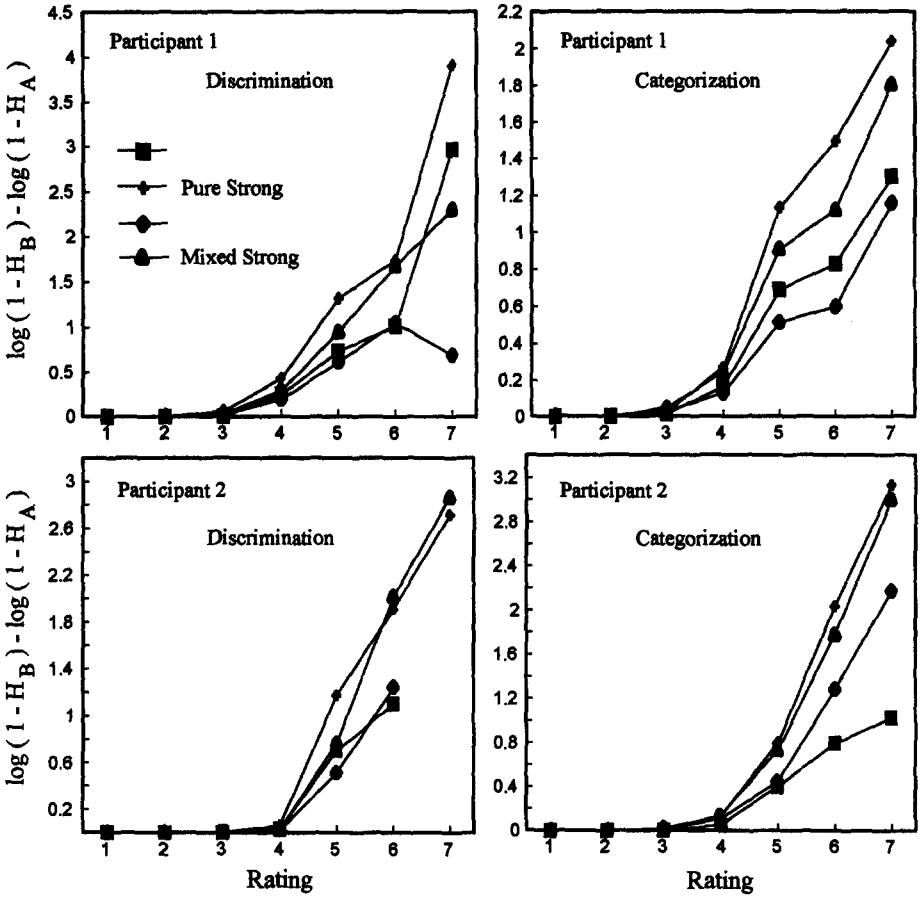

Figure 11. The log of the ratio of survivor functions (hazard rate dominance test) for the discrimination and categorization experiments. As in Figure 7, the functions are shown for rating responses up to 7; beyond this point, they continued to increase.

were used to generate stimuli in the numerical categorization task; however, instead of line length, the actual integer value representing the number of pixels used to generate the line stimuli was presented as a stimulus. In all other respects, the methods were identical in all three conditions, as detailed below.

Participants. The 2 participants were undergraduate students at Northwestern University. They were paid $7 per hour for a total of 13 hr (each) of participation.

Procedure. For each participant, the first session lasted 1 hr and was counted as a warm-up. The stimuli were blocked according to the mixed–pure design defined above. Each block was identified to the participant before it began as a mixed, a pure weak, or a pure strong block of trials, and the meaning of this language was carefully explained. Each block began with a demonstration series of trials sampled using the appropriate stimulus mixture and requiring no response. A stimulus was presented for 1 s, and the correct response to it was then displayed. The demonstration consisted of 9 trials for the discrimination task and 32 trials for the categorization tasks. In all three tasks, the demonstration was followed by 32 response trials in which feedback was given on each trial. Performance was self-paced, with short breaks allowed at any time and a required 10–15-min break after the first 55 min. Total time for individual sessions was 2 hr.

Results and Discussion

Table 3 lists the percentage of correct B responses (hits) and incorrect B responses (false alarms) for each of the conditions of the experiment. For most of the analyses, results of the discrimination and perceptual categorization conditions are presented together, following the same order that was used for the recognition-memory data, apart from the additional analyses to be included. The corresponding results from the numerical categorization condition were omitted because they merely replicated those of the perceptual categorization condition. In fact, overall, the major conclusions are the same for all three tasks: There was substantial support for stochastic dominance in AB comparisons, but there was no evidence for the intersection patterns predicted by the optimal model in AA or BB comparisons.

Objective certainty test. The estimated $P(A \mid R = k)$ functions are shown in Figure 9 for the data-combined conditions for each participant. Although there were two violations of monotonicity, they occurred in the extremes of the response scale and represented small proportions esti-

mated from extremely small sample sizes (15 samples for Participant 1 and 7 samples for Participant 2, out of a total of more than 3,000 responses per participant). Thus, the results strongly favor models predicting monotonicity of this function in its estimable range.

Stochastic dominance: AB comparisons. The convexity and hazard rate tests for discrimination and categorization are shown in Figures 10 and 11, respectively. Once again, the functions are monotone increasing, with the only exceptions occurring in the extremes, where the sample sizes make the estimates unreliable. In Figure 10, there is a striking similarity between the shapes of the functions in the two experiments, adding some prima facie support for the idea that physical noise in the categorization task has the same qualitative effect that internal noise has in discrimination tasks.

Stochastic dominance: AA and BB comparisons. The cumulative frequency distributions for each stimulus by stimulus condition (pure weak, pure strong, and mixed) are shown in Figure 12. The results are very similar to those of recognition memory; that is, the functions are very close together with no well-defined intersections, and without the pattern of a single intersection followed by increasing separation that would be indicative of an optimal decision rule.

The conclusions from all three experiments are therefore the same. Dominance between the cumulative frequency distributions from A versus B stimulus trials is strongly supported, but violations of dominance for the same stimulus in different B conditions (e.g., A in the pure weak condition versus A in the mixed condition) is not.

Direct estimates of the confidence by strength functions. One final analysis that was applied to the categorization data served to illustrate why the optimal model failed to make the correct predictions about the distributions. Recall that the tests that counterindicate the optimal decision rule are based on the effect that the stimulus condition (pure weak, pure strong, or mixed) should have on the feeling of confidence caused by a given piece of perceptual informa-

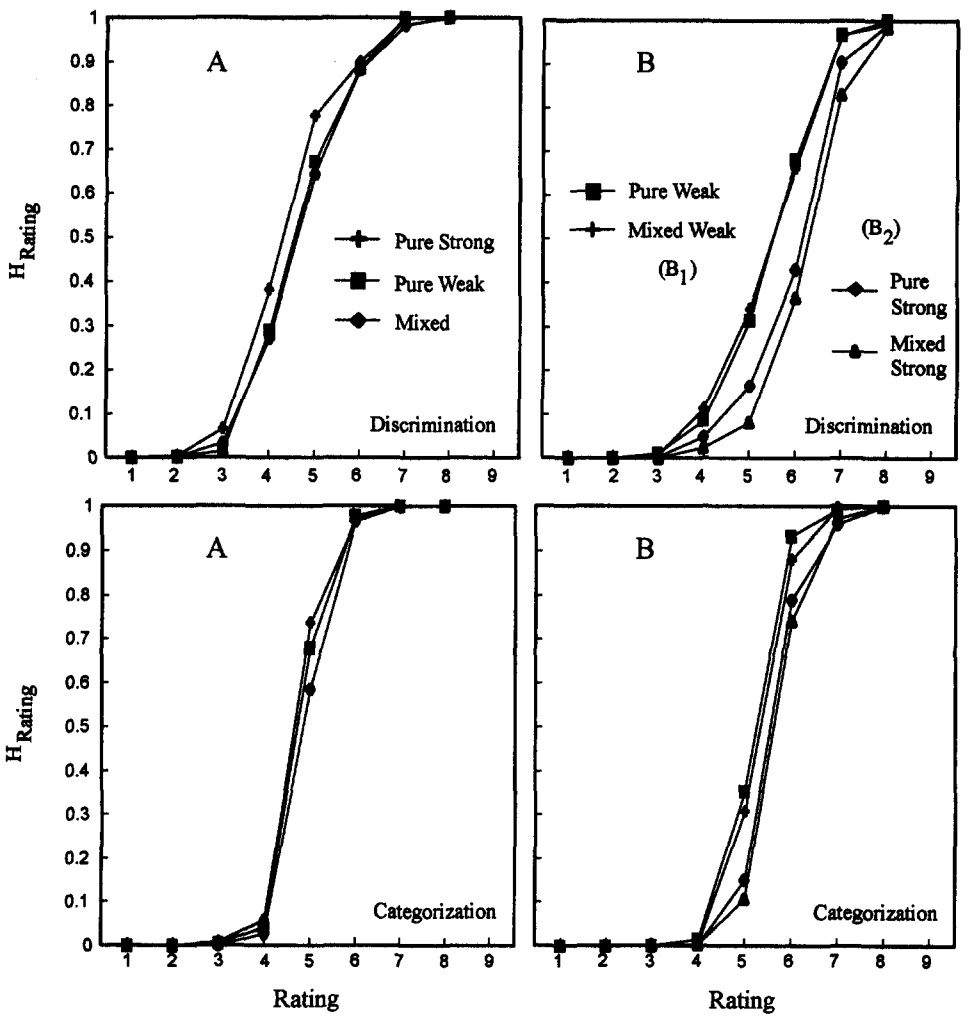

Figure 12. Cumulative frequency distributions of confidence ratings by stimulus (A or B) and mixed–pure condition for the discrimination and categorization experiments (data combined across participants).

tion about the stimulus (S). In the categorization tasks, physical variability in the stimuli is presumably large relative to the expected perceptual noise level. Thus, except for a negligible error, the physical size of the stimulus itself should be a good measure of perceived size.

In Figure 13, the mean confidence rating as a function of the stimulus size (line length or number size) is plotted for the three conditions of the mixed–pure categorization tasks. Empirical results are shown in the four upper panels, and predictions of the optimal and distance-from-criterion models are shown in the two lower panels. For the distance-from-criterion model, the choice criterion (T) was set at its optimal point for each condition (i.e., the point that maximizes percentage correct), and for both models, additive normal noise was added to the confidence criteria to mimic the effects of perceptual or criterial noise in the mapping from confidence to ratings. All functions, both empirical and theoretical, were smoothed by a 25-point Hamming window (i.e., the plotted value at Abscissa Point X becomes the mean of the 25 unsmoothed values to the left and the right of X) and truncated in the extreme tails (i.e., where small sample sizes cause the estimates to become erratic).

The main result of this analysis is straightforward: The functions are fairly close together, indicating that the effect

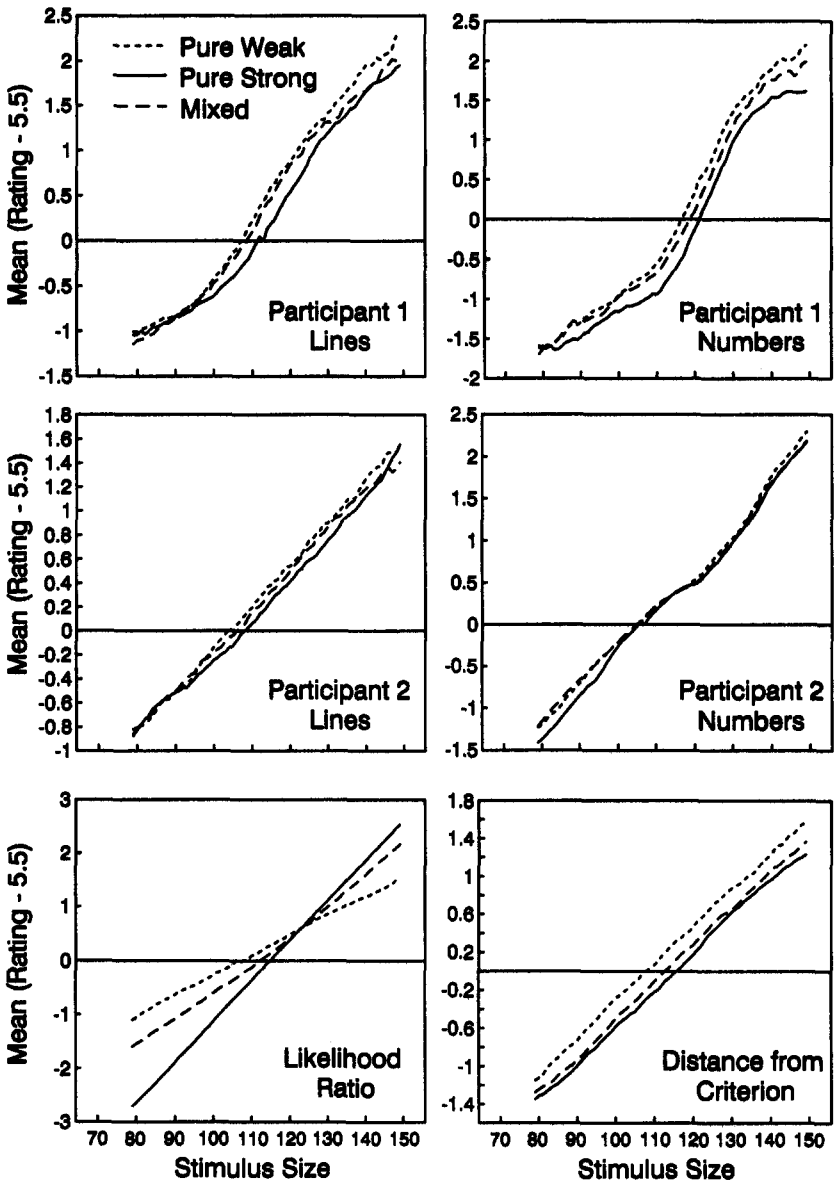

Figure 13. Mean confidence rating as a function of the stimulus value in categorization. The four upper panels represent performance of the participants in categorizing lines (left side) and numbers (right side). The two lower panels illustrate the predictions of the optimal and distance-from-criterion models.

of context (mixed vs. pure) on confidence is not strong. A close analysis, however, shows that a dominance pattern does exist. To illustrate this result more clearly, Figure 14 shows the values of the confidence functions for a more central range of stimulus values. Notice that all three functions are different, the order being: pure weak ≥ mixed ≥ pure strong. Once again, this is the pattern predicted by a distance-from-criterion model that assumes that the choice criterion is adjusted by the participant in the direction of the optimal location (i.e., the location that maximizes the percentage of correct responses). Thus, a simple version of this model, in which the criterion is affected by context but not the transformation function, $g(.)$, seems to account for all of the major results of this study.

General Discussion

Many theories of perception do not make a strong commitment about the nature or the contribution of decision processes in laboratory perception tasks. The main purpose of the theory is to explain how the internal perceptual representations of the stimuli are obtained, and therefore the implicit assumption is made that a stimulus is classified incorrectly when it is perceived incorrectly, or incorrectly enough within the context of the experiment. When a signal detection theory analysis is used to separate decision-making effects from encoding-level effects on performance, the idea is still that participants base their judgments directly on the percept, and thus sensitivity is represented by a pair of perceptual distributions.

A somewhat different, and ultimately more general, approach to modeling classification performance is to begin with the assumption that participants choose the response that they believe is most likely to be correct. In this case, the proportion of errors represents the proportion of trials in which an incorrect response seemed more likely to the participant, that is, for which the perceived likelihood value was largest. This value could depend on the perceptual information in a number of different ways, and so the problem is to find a test of the decision-making process without knowing how the perceptual encoding process works.

Evidence for the Distance-From-Criterion Decision Model

The idea that classification data measure the perceived likelihood ratio is not inconsistent with classical signal detection theory; this model merely adds an assumption about the decision process (i.e., the distance-from-criterion model), which causes the response proportions to measure, in effect, both the perceptual and perceived likelihood ratio effects of the stimuli. However, the emphasis on the perceptual distributions in signal detection theory tends to disguise the importance of the likelihood ratio in the original sources of the model (i.e., statistical decision theory). For example, to predict some very simple facts, including the lawful effects that physical similarity of the stimuli has on performance level, the decision maker must have sub-

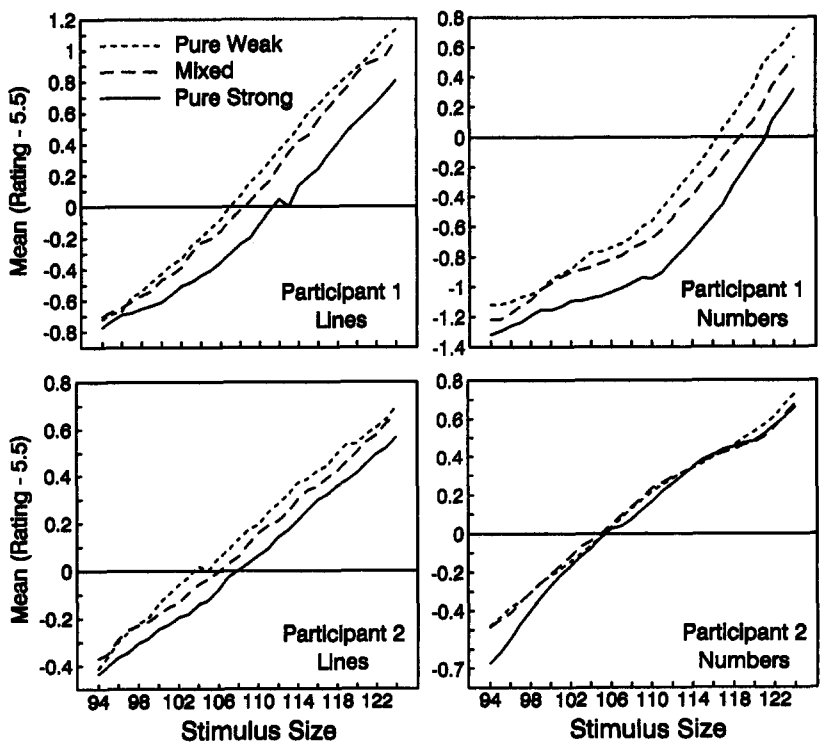

Figure 14. Central regions of the empirical functions shown in Figure 13.

stantial knowledge about the relationship between perceptual experiences and the objective likelihoods of the stimuli. That is, the perceived likelihood ratio must be at least reasonably close to the true likelihood ratio, even though the function used to transform the perceptual effect into a response must change radically whenever the stimuli change.

The main conclusion of the empirical analyses reported here is that the perceived likelihood ratio and the objective likelihood ratio of an optimal decision system are close but not identical. In short, the participants did not use an optimal decision rule. If they did use such a rule, then a very specific pattern of crossovers should have been observed between the cumulative frequency distributions of confidence ratings obtained from a participant when the same stimulus was presented under different discrimination conditions (i.e., what the other stimuli in the discrimination task were, see Table 1 and Figure 3). There was no sign of this crossover pattern in the empirical data we examined.

A better model for the decision process is the distance-from-criterion rule. Specifically, the participant chooses a single perceptual state, which serves as a cutoff value between the two discrimination responses, and perceived likelihood ratio is assumed to be some increasing function of the distance of the incident percept from this criterion. The superiority of this model goes beyond the issue of whether any response bias exists or not, because the data also rule out the possibility that participants use the correct formula when computing the likelihood ratio but insert an incorrect estimate of the a priori probabilities of the stimuli (i.e., the β value in signal-detection theory). This kind of bias changes the quantitative predictions of the model but not its qualitative predictions (i.e., violations of stochastic dominance in the mixed–pure paradigm).

Final Comments: Implications of Stochastic Dominance Between Perceived Likelihood Ratio Distributions

Although the optimal model can be rejected, it is worth noting that this model did correctly predict, without any data fitting or any assumptions about the encoding effects, that stochastic dominance will always be satisfied whenever the rating distributions for the two stimulus conditions are compared (see Table 1). To explain these dominance properties of the data, together with the predictiveness of the confidence level (i.e., the objective certainty test), the distance-from-criterion rule must assume that the objective likelihood ratio function is monotone (e.g., Green & Swets, 1966). This ensures that the objective likelihood ratio is a monotone function of distance from the criterion, which ensures that the highest level of dominance in the stochastic dominance hierarchy will hold for both the perceptual effects and perceived likelihood ratio (confidence) distributions (e.g., Townsend, 1990; Townsend & Ashby, 1983).

In principle, there is no reason why the perceptual effects distributions should not have a monotone likelihood ratio function. Ultimately, this type of constraint implies that the encoding system makes the decision problem much easier than it might be, because setting a single criterion on the percept under these circumstances allows the decision process to perform at a level arbitrarily close to that of the optimal decision rule. Therefore, the ultimate conclusion of this study is that decision processes are not optimal in the pure sense of statistical decision theory, but they are reasonably sophisticated and efficiently matched with the encoding process.

References

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 50–71.

Balakrishnan, J. D. (1994). Simple additivity of stochastic psychological processes: Tests and measures. *Psychometrika, 59,* 217–240.

Dzhafarov, E. N., & Rouder, J. N. (in press). Empirical discriminability of two models for stochastic relationship between additive components of response time. *Journal of Mathematical Psychology.*

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

Glanzer, M., Adams, J. K., & Iverson, G. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

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

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory & Cognition, 19,* 119–130.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–168.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195.

(Appendix follows on next page)

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin, 108,* 551–567.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In N. J. Castellan, Jr., & F. Restle (Eds.), *Cognitive theory* (Vol. 3, pp. 200–239). Hillsdale, NJ: Erlbaum.

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

Appendix

Derivations of the Stochastic Dominance Tests

All of the model predictions described in the text are stated with respect to the cumulative frequency distributions of the observable rating responses. To prove the results, we show that the necessary relationships hold for the cumulative distribution functions of the feeling of confidence (E); in each case, the application to rating distributions follows from the assumed relationships between confidence and confidence ratings discussed in the text.

Proof of Stochastic Dominance Predictions of the Optimal Decision Model

A useful result in statistical decision theory is that the likelihood ratio of two random variables that are themselves the product of a likelihood ratio transformation of two other random variables (e.g., L_A and L_B are computed from S_A and S_B) is necessarily monotone (since "the likelihood ratio of the likelihood ratio is the likelihood ratio," Green & Swets, 1966, p. 26). Townsend and Ashby (1978) showed that a monotone likelihood ratio for two variables implies that their hazard rate functions are ordered, which implies that their cumulative distribution functions are ordered.

To prove that the convexity functions are ordered, note first that

$$\frac{f_{L_B}(t)}{F_{L_B}(t)} \geq \frac{f_{L_A}(t)}{F_{L_A}(t)}$$

if and only if

$$f_{L_B}(t)F_{L_A}(t) - f_{L_A}(t)F_{L_B}(t) \geq 0.$$

This quantity can be rewritten as

$$f_{L_B}(t)\int_0^t f_{L_A}(s)\,ds - f_{L_A}(t)\int_0^t f_{L_B}(s)\,ds$$

$$= \int_0^t f_{L_B}(t) f_{L_A}(s)\,ds - \int_0^t f_{L_A}(t) f_{L_B}(s)\,ds$$

$$= \int_0^t [f_{L_B}(t) f_{L_A}(s) - f_{L_A}(t) f_{L_B}(s)]\,ds .$$

The fact that the likelihood ratio function,

$$\frac{f_{L_B}(t)}{f_{L_A}(t)},$$

is increasing with t guarantees that the integrand in the last expression is positive for all values of s, which establishes the result.

To show that the optimal classifier model predicts that the two ratios

$$\frac{1 - F_{E_B}(t)}{1 - F_{E_A}(t)}$$

and

$$\frac{F_{E_B}(t)}{F_{E_B}(t)}$$

must be increasing if the participant is using the optimal decision rule, note that

$$\frac{d}{dt}\log\left[\frac{1 - F_{E_B}(t)}{1 - F_{E_A}(t)}\right] = \frac{f_{E_B}(t)}{1 - F_{E_A}(t)} - \frac{f_{E_B}(t)}{1 - F_{E_B}(t)}$$

and

$$\frac{d}{dt}\log\left[\frac{F_{E_B}(t)}{F_{E_A}(t)}\right] = \frac{f_{E_B}(t)}{F_{E_B}(t)} - \frac{f_{E_A}(t)}{F_{E_A}(t)}.$$

Because $d/dt \log[g(t)]$ has the same sign as $d/dt\ g(t)$, these equalities imply that the two ratio functions must be increasing if the hazard rate and convexity function dominance properties hold.

Finally, to show that the distance-from-criterion model does not necessarily make any of these dominance predictions, suppose that

$$F_B(t) = F_A(\alpha t + \beta).$$

This makes the strength distributions a location-scale pair, which implies that dominance of their cumulative distribution functions will be violated if α is not equal to 1. Violation of lower levels necessarily implies violation of all higher levels of the stochastic dominance hierarchy, and monotonicity of the $g(.)$ transformation to E_A and E_B ensures that the same violations occur for the cumulative distribution functions of these variables.

Proof That the Optimal Classifier Model Predicts Violations of Stochastic Dominance in AA and BB Comparisons of the Mixed–Pure Paradigm (Context Sensitivity Test)

The Strength Likelihood Ratio Function, L_s Is Monotone and the B_1 and B_2 Strength Distributions Intersect

Let $K_{PW}(t)$, $K_M(t)$, and $K_{PS}(t)$ be the likelihood ratio functions under the three different mixed–pure stimulus conditions (PW = pure weak, M = mixed, and PS = pure strong). Let t^* be the

strength value at which the strength distributions of the B_1 and B_2 stimuli intersect, that is,

$$f_{B_1}(t^*) = f_{B_2}(t^*).$$

Let v^* be the value of the likelihood ratio functions at this point, that is,

$$v^* = K_{PW}(t^*) = K_M(t^*) = K_{PS}(t^*).$$

Because the likelihood ratio functions are monotone, their inverse functions exist. Therefore,

$$P(L_{A(PW)} \leq v) = P[S_A \leq K_{PW}^{-1}(v)],$$

$$P(L_{A(M)} \leq v) = P[S_A \leq K_M^{-1}(v)],$$

$$P(L_{A(PS)} \leq v) = P[S_A \leq K_{PS}^{-1}(v)],$$

where the superscript -1 denotes the inverse function. From the fact that the B_1 strength distribution must be larger than that of B_2 for strength values less than v^*, and smaller for values greater than v^*, it follows that

$$K_{PW}^{-1}(v) < K_M^{-1}(v) < K_{PS}^{-1}(v), \qquad v < v^*$$

$$K_{PW}^{-1}(v) > K_M^{-1}(v) > K_{PS}^{-1}(v), \qquad v > v^*.$$

From this the statement about the cumulative distribution functions of the three noise conditions follows directly. Essentially the same arguments apply for the mixed–pure comparisons of E under B_1 and B_2 stimulus conditions.

The Strength Likelihood Ratio Function Is Nonmonotone

Because the term *nonmonotone* is not a specific statement about the shape of the function, there are many ways to proceed. We add the assumptions that (a) the likelihood ratio functions $K_{PW}(t)$ and $K_{PS}(t)$ decrease then increase, with $K_{PS}(t)$ initially greater than, subsequently less than, and finally greater than $K_{PW}(t)$ (i.e., the functions intersect twice), and (b) the minimum value of $K_{PS}(t)$ is less than that of $K_{PW}(t)$. This particular set of assumptions was chosen because it describes the normal, unequal variance model of Figure 3.

First, note that the second assumption immediately implies that

$$F_{L_{A(PW)}}(v) \leq F_{L_{A(M)}}(v) \leq F_{L_{A(PS)}}(v)$$

for sufficiently small values of v. The first assumption implies that for large enough values of v, the interval of strength values such that $K_{PW}(t)$ is less than v is larger than that of $K_M(t)$, which is larger than that of $K_{PS}(t)$. This implies that

$$F_{L_{A(PW)}}(v) \geq F_{L_{A(M)}}(v) \geq F_{L_{A(PS)}}(v)$$

for sufficiently large values of v, which proves that the functions must intersect. The proofs for the B_1 and B_2 stimulus conditions follow virtually the same steps.

The predictions of the distance-from-criterion model stated in the text depend on simple results about the effects of monotone transformations of random variables. For example, consider the case in which $g(.)$ is the identity function, that is, $g(S - T) = S - T$. The perceived likelihood ratio distributions are all shifted by the amount T; hence, they are effectively identical to those of the strength distributions.

Received December 15, 1993
Revision received March 7, 1995
Accepted April 13, 1995 ■