Journal of Experimental Psychology:
Learning, Memory, and Cognition
2000, Vol. 26, No. 3, 582–600

0278-7393/00/$5.00 DOI: 10.1037//0278-7393.26.3.582

ROC Curves and Confidence Judgments in Recognition Memory

Trisha Van Zandt
Johns Hopkins University

Most models of recognition memory rely on a strength/familiarity-based signal detection account that assumes that the processes giving rise to a confidence judgment are the same as those giving rise to an old–new decision. Confidence is assumed to be scaled directly from the perceived familiarity of a probe. This assumption was tested in 2 experiments that examine the shape of confidence-based *z* receiver operating characteristic (*z*ROC) curves under different levels of response bias induced by changing stimulus probabilities (Experiment 1) and payoffs (Experiment 2). Changes in the shape of the *z*ROC curves with bias indicate that confidence is not scaled directly from perceived familiarity or likelihood. A model of information accumulation in recognition memory is proposed that can account for the observed effects.

In 1958, Egan applied the considerable power of signal detection theory to the problem of recognition memory, where he realized that the distinction between a new, unstudied item and an old, studied item could be equated to a situation in which an observer attempted to detect signals presented in noise. This work began the long and fruitful investigation of strength theories of memory (Atkinson & Juola, 1973; Murdock, 1965; Parks, 1966), which continues to this day. The idea of a continuum of familiarity together with a criterion for responding "old" has become the heart of most successful models of recognition and recall.

In particular, the *global memory* models, so called because they each assume that a probe is compared with all items in memory at once, describe how familiarity values are computed for old and new items. There are three major models of this type that have experienced more or less the same degrees of success at explaining various memory phenomena. These are TODAM (Murdock, 1982), MINERVA 2 (Hintzman, 1988), and SAM (Gillund & Shiffrin, 1984). Despite differences among the models, the recognition decision process proceeds in exactly the same way for each. A single familiarity value is computed that indicates the degree to which the probe matches one or more items in memory and a decision is made on the basis of a comparison between that value and a criterion. In these models, confidence judgments arise from the setting of several criteria along the familiarity axis; that is, confidence is scaled directly from the perceived familiarity of a probe.

This model of confidence judgments, the target of this article, permits the efficient construction of receiver operating characteristic (ROC) curves (Egan, 1958; Ratcliff, Sheu, & Gronlund, 1992), the plots of hit ("old" responses to old words) versus false-alarm ("old" responses to new words) rates. Such curves are important because, as Ratcliff et al. (1992) noted, different memory models make different predictions about the shapes of the curves (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Yonelinas, 1994). The points of the ROC curve are often transformed to *z* scores (under the assumption of normally distributed familiarity), and then the *z* scores are plotted as a *z*ROC curve. The *z*ROC curve is convenient because (if the normality assumption holds) it is a straight line, the slope of which is the ratio of the new-to-old probe familiarity standard deviations, and the intercept of which is proportional to d', the separation between the familiarity distributions. One strong basis of support for the criterion hypothesis of confidence judgments is the fact that both Egan (1958) and Ratcliff et al. constructed *z*ROC curves using confidence judgments and compared them with the *z*ROC curves constructed by manipulating bias in simple yes–no recognition. There were no differences between the curves constructed in these different ways.

The criterion hypothesis of confidence makes a simple prediction: Confidence-based *z*ROC curves should be invariant under different levels of response bias if confidence is scaled directly from familiarity. For a confidence scale that ranges from 1 (*certain old*) to 6 (*certain new*), five criteria are placed along the familiarity axis (see Figure 1). The effect of bias is to shift these criteria toward either end of the axis, as shown in the top and bottom panels of the figure. Because the familiarity distributions do not change with bias, the *z*ROC curves constructed under bias should fall on the same line, shown in the center panel of Figure 1. The points represented by squares on the *z*ROC curves are those that might be generated by a bias to respond "new," whereas the circles are those that might be generated by a bias to respond "old." Therefore, the *z*ROC curves constructed

This research was funded by National Science Foundation Grant SBR-9702291. Portions of this study were presented at the 1996 annual meeting of the Society for Mathematical Psychology, University of North Carolina, Chapel Hill. I thank Christina Eyster for help in running the experiments reported here, and Roger Ratcliff and David Scrams for helpful discussion in the early stages of this work. Thanks also are due to Doug Hintzman, Derrick Parkhurst, Tom Wallsten, and Steven Yantis for comments that improved an earlier version of this article.

Correspondence concerning this article should be addressed to Trisha Van Zandt, Department of Psychology, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218.

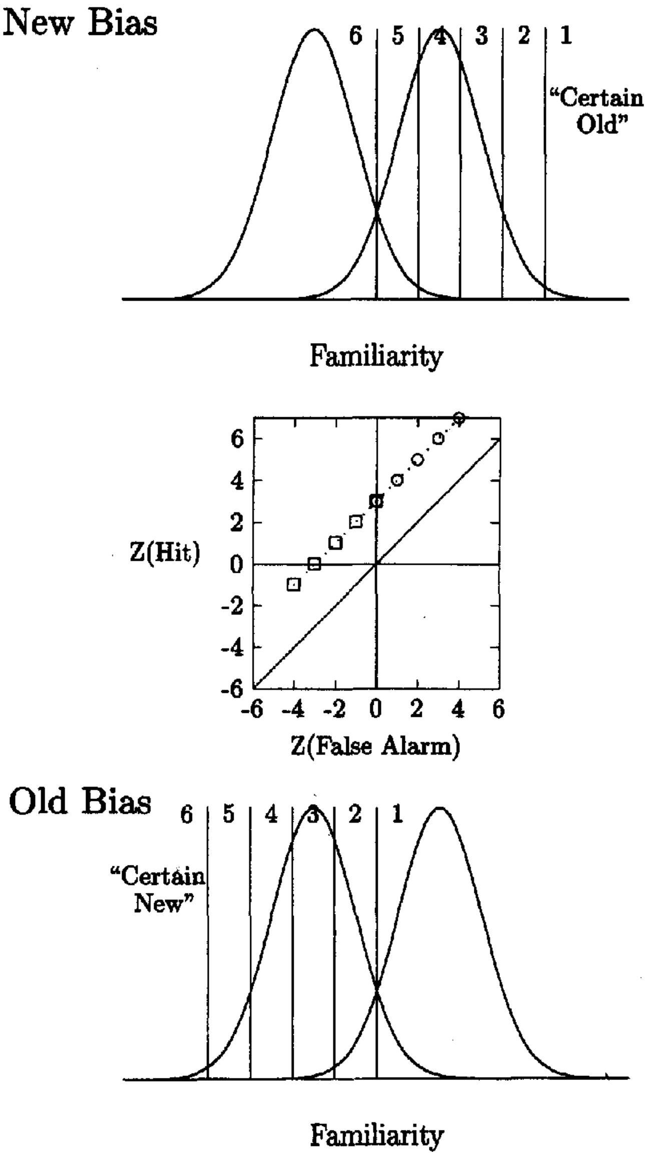

Figure 1. The *z*ROC curve under bias. The top panel shows the placement of confidence criteria under a bias to respond "new." The points on the *z*ROC curve corresponding to these placements are shown as squares (center panel). The bottom panel shows the placement of confidence criteria under a bias to respond "old." The points on the *z*ROC curve corresponding to these placements are shown as circles (center panel). The straight-line *z*ROC curve is identical for both the new and old bias points.

from confidence judgments under different response biases should have the same slopes and intercepts regardless of how the criteria are placed and regardless of the variances of the distributions.

Several more recent global models of recognition memory and confidence judgments have supposed that old–new decisions are based on the posterior odds that a probe is old or new given that it evoked some level of familiarity (Glanzer & Adams, 1990; McClelland & Chappell, 1995; Shiffrin & Steyvers, 1997).[1] The appeal of the posterior-odds approach is that it allows rememberers to select a response according to their perception of the response that is most likely to be correct given some level of familiarity. The posterior odds is defined using the likelihood ratio

$$L(x) = \frac{P(x|Old)}{P(x|New)} = \frac{f_O(x)}{f_N(x)},$$

where $P(Old)$ and $P(New) = 1 - P(Old)$ are the experimenter-determined stimulus probabilities and $f_j(x)$ is the probability density of perceiving x given that the probe (j) is old or new. As discussed above, f_j is often assumed to be the normal density, but it may take on any shape. The posterior odds is

$$O(x) = \frac{P(Old|x)}{P(New|x)} = \frac{P(Old)\,f_O(x)}{P(New)\,f_N(x)} = \frac{P(Old)}{1 - P(Old)}L(x).$$

For convenience, the log odds function is referred to; it is

$$\ln O(x) = \ln L(x) + \ln \frac{P(Old)}{1 - P(Old)}. \qquad (1)$$

If confidence is scaled from perceived odds or log odds instead of from familiarity, the *z*ROC curve must still be invariant under bias. To see that this must be true, consider the relationship between log odds and familiarity given in Equation 1. Familiarity (x) is sampled from one of two distributions with density functions f_N and f_O. There are, then, two distributions for the possible values of ln $L(x)$: one for when x is sampled from an old probe and the other for when x is sampled from a new probe. The distributions of ln $L(x)$ are invariant with respect to $P(Old)$, and the distributions of ln $O(x)$ are shifted up or down depending on $P(Old)$. Therefore, the variance and the separation between the distributions of ln $O(x)$ are constant across different values of $P(Old)$, and it is only the relative locations of the distributions that change. Thus the *z*ROC curves must all be equal across different values of $P(Old)$.

The invariant *z*ROC prediction for both familiarity- and odds-based models depends on the assumption that the underlying distributions are normal; that is, that the transformation from hit and false-alarm rates to *z* scores is appropriate. All of the global memory models produce normally

[1] Although Glanzer and Adams (1990) referred explicitly to the likelihood ratio, test lists in their experiments were always composed of equal numbers of old and new words. There is no difference between the likelihood ratio (the ratio of the probabilities of observing some level of familiarity given that a test word is old or new) and the posterior odds (the ratio of the probabilities of the test word being old or new given some perceived level of familiarity) when the prior probabilities of the test word being old or new are equal to .5.

distributed familiarity distributions. However, the theory of signal detection is far more general. The distributions of perceived magnitude may take on any shape, and the normal distributions are only a very special (and restrictive) case. The linearity of the zROC is very robust to violations of the normality assumption (Murdock, 1965). An example is shown in Figure 2, where the new and old familiarity values are exponentially distributed. The "new" distribution starts high for low familiarity values, whereas the "old" distribution has a higher tail for high familiarity values. The zROC curve for these two distributions is shown in the center panel. The best-fitting line through these points accounts for more than 99% of the variance and has a slope of .75 and an intercept of .71. Under bias to say "new," shown in the top panel, the lower points on the zROC curve, shown as squares, are swept out. Under bias to say "old," shown in the bottom panel, the upper points on the zROC curve, shown as circles, are swept out.

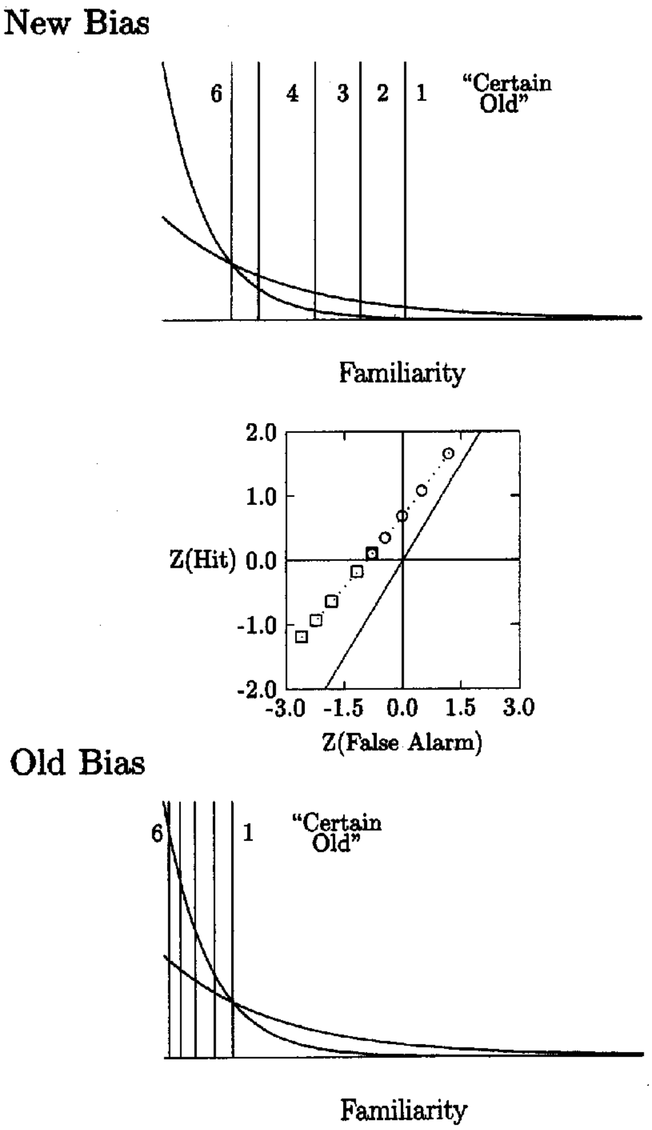

Figure 2. The zROC curve under bias for nonnormal distributions of familiarity. Points on the curve swept out under new bias are shown as squares, whereas points on the curve swept out under old bias are shown as circles (center panel). Despite the violation of the normality assumption, the zROC curve is still linear, and invariant under bias.

Violations of the normality assumptions may lead to slightly different slopes across bias because of picking out separate parts of a nonlinear zROC curve and fitting a straight line to them. It is hard to imagine what sort of familiarity distributions might lead to gross nonlinearities in the signal detection model,[2] although Yonelinas (1994) proposed a dual-process model of recognition that predicts nonlinear, U-shaped zROCs that are due to an increase in high-confidence responses derived from recollection-based responding. To avoid spurious slope differences arising from a nonlinear zROC, the different zROC curves constructed under bias must overlap, or span the same range, so that any differences between them cannot easily be attributed to nonlinearity. Visual inspection of the zROC curves is also important to rule out collinearity. Regardless of the nature of the distributions, and of whether the zROCs are linear, the prediction of the criterion hypothesis is unchanged: The curves should be invariant with respect to bias. The nonlinear zROCs produced by the dual-process model should also be invariant with bias, unless a mechanism is proposed whereby the changes in bias also change the proportion of recollection-based responses or how those responses are distributed to different levels of confidence.

In the following sections I present the results of two experiments designed to test the invariance of the zROC curves under bias. In both experiments participants studied lists of words. They were then given information about the test list designed to influence their response bias. In the first experiment participants were told the probabilities that probes would be old or new. In the second experiment participants were given different payoffs for hits and correct rejections. Because in each experiment this information was not provided until after participants had studied the word lists, encoding conditions were constant across all levels of bias for each participant. Participants then performed recognition judgments by estimating their confidence that probes were old or new. These confidence judgments were used to construct zROC curves. In both experiments, the zROC curves changed with bias, which is not consistent with the criterion model of confidence judgments. After describing the experiments I present an alternative model, based on a process of information accumulation, that can explain the results.

[2] I explored a large number of distributions in an attempt to answer this question. None were able to produce nonlinear zROCs. I then explored shapes of the ROC curves themselves, to determine what kind of ROC would transform into a nonlinear zROC. Nonlinear zROCs were produced by highly aberrant ROC curves that formed a U-shaped function in ROC space. It is unknown what sort of distributions might produce U-shaped ROC curves.

Experiment 1: Probabilities

In this experiment two groups of participants gave confidence judgments under fast and slow study conditions.

Method

Participants. Ten English-speaking Johns Hopkins University students participated in this study for pay. All participants reported normal or corrected-to-normal vision.

Materials and apparatus. The words for all blocks were chosen randomly without replacement from a pool of 2,337 common English words (mean Kučera–Francis, 1967, written frequency = 49 per million), four to eight letters long ($M = 6.4$ letters). Nouns and their plurals (e.g., *book, books*) and verb conjugates (e.g., *walk, walking*) did not appear together in the pool. Words were presented in uppercase in the center of a computer monitor, appearing light on a dark background. Participants responded by pressing the keys of the computer's keyboard. Timing of the stimuli and responses were controlled by computer software.

Procedure. Participants were presented with a list of 32 words for study. The first and last pairs of words in the list were tested but not included in the analysis. After the last item on the study list was presented, the participant was told how many words in a subsequent test list would be old (i.e., previously seen in the study list). Old words were selected randomly and without replacement from the study list. Additional new words were selected randomly and without replacement from the pool for a total of 40 words in the test list. Studied words and distractors were used only once per session.

In the slow study condition, the words in the study list were presented one at a time for 550 ms, separated by 200-ms pauses (750 ms per word). Immediately after the presentation of the last word in the study list and before the test list, the participants were informed that 8, 14, 20, 26, or 32 of the 32 words on the study list would be included in the test list. In the fast study conditions, the words in the study list were presented one at a time for 200 ms, separated by 200-ms pauses (400 ms per word). Immediately after the presentation of the last word in the study list and before the test list, the participants were informed that 20%, 35%, 50%, 65%, or 80% of the words on the test list were "old." The modification in the way that pretest information was displayed was made to make the changes in stimulus probabilities more obvious, because 2 pilot participants in the fast study condition misinterpreted the pretest information. No confusion was reported or observed in the slow study condition.

Presentation of the test list began 3 s after the frequency/probability information. Each item on the test list was presented in the center of the screen, preceded by a 50-ms warning stimulus (a hyphen [-] two spaces to the left of the first letter in the word) and a 200-ms pause. Following the participant's response to the word, there was a 250-ms pause before the next warning stimulus.

Participants were instructed to respond to each test word using a 6-point scale varying from "certain old," "probably old," "maybe old," and so on, to "certain new." These responses were mapped onto the X, C, V, B, N, and M keys on the computer keyboard and were executed with the first three fingers of the participants' left and right hands. Response times were measured from the onset of the test word. Participants were instructed to take their time, think carefully about their level of confidence, and try to distribute their responses over all six keys. After responding to the test list, each participant was given feedback about overall response accuracy. One point was awarded for every correct response (e.g., responding either "certain old," "probably old," or "maybe old" when the stimulus was in fact old), and 1 point was deducted for every incorrect response (e.g., responding either "certain new," "probably new," or "maybe new" when the stimulus was in fact old). At the end of each session, participants were paid an additional half cent for every point they earned—an average $1.50 bonus per session.

There were 12 study–test blocks in each session. The first two blocks of each session were deemed practice trials and were not included in the analyses. These blocks were also unbiased in that the test lists contained 20 old and 20 new words. Each level of bias (8, 14, 20, 26, and 32 old words) was presented twice in the remaining 10 blocks in random order. Participants performed in 10 sessions over a period of approximately 2 weeks.

Results

The *z*ROC curves were constructed from the confidence judgments for each participant for each of the five levels of bias in each experiment. The overall response frequencies were corrected by adding ⅙ to each frequency and increasing the total number of responses for each bias condition by 1 (Snodgrass & Corwin, 1988). This was done to avoid having to collapse low-frequency confidence levels together to fit *z*ROC curves. (Although the Snodgrass and Corwin [1988] correction has the effect of artificially increasing r^2, it has little effect on the estimated *z*ROC slopes.) Linear functions were fit to each *z*ROC curve for each level of bias using a maximum-likelihood estimation technique (Dorfman & Alf, 1969).[3] Simple linear regressions were also performed for each curve, and the results were similar. All reported effects were significant at the $\alpha = .05$ level.

The *z*ROC data for each participant are presented in Figure 3, along with the best-fitting lines for the 20% (solid line) and 80% (dotted line) conditions. Part A of the figure shows the data for the 5 participants in the fast condition (Participants 1–5), and Part B shows the data for the 5 participants in the slow condition (Participants 6–10). Although the pattern of effects is not easy to see in this figure, there are significant differences among the slopes for each participant and bias condition. The size of the effects ranges from very small for some participants to very large for others. Overall, the slopes tended to be larger under "old" bias: Of the 10 participants, 7 showed larger average slopes in the 65% and 80% old conditions than in the 20% and 35% old conditions. The remaining 3 participants (3, 6, and 8) showed no systematic differences across levels of bias.

If one examines the *z*ROC curves in Figure 3, one can see that the points for the different bias conditions indicate that, overall, the slope differences are not due to selection of different parts of a single, nonlinear *z*ROC curve (see especially Participants 2, 7, and 10). Using *z*(Hit) as the dependent variable, regression onto *z*(False Alarm) as well as *z*(False Alarm)^2 and *z*(False Alarm)^3 showed that quadratic and cubic regression terms accounted for significantly more variance than did the linear term alone when collapsing over all bias conditions. However, when entering each

[3] The equations given by Dorfman and Alf (1969) contain some errors, which were corrected for these analyses. (See Grey & Morgan [1972].)

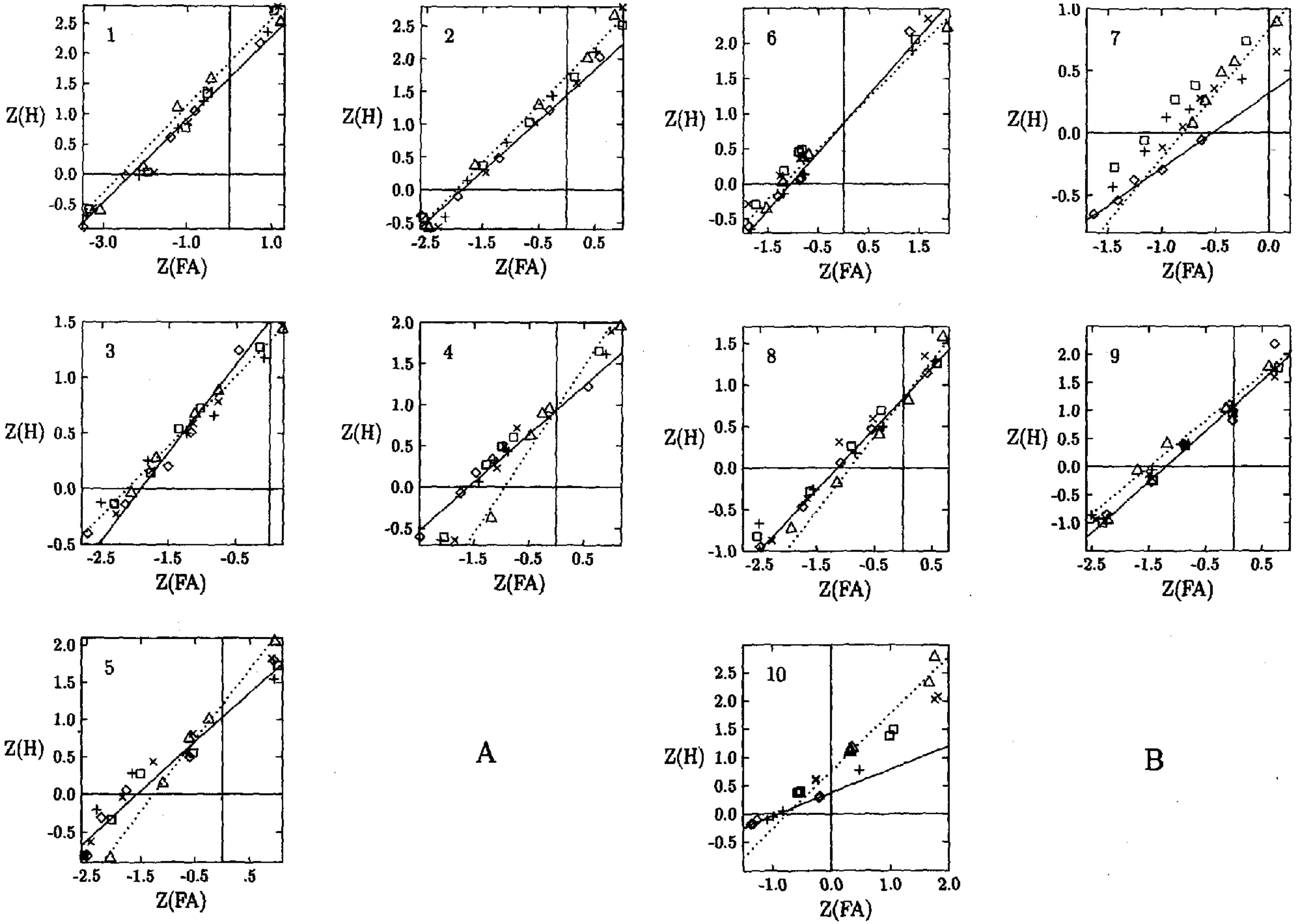

Figure 3. The zROC curves for each participant in Experiment 1. The slow condition is shown in Part A (Participants 1–5), and the fast condition is shown in Part B (Participants 6–10). Conditions of response bias vary from 20% old words (diamonds, solid lines) to 80% old words (triangles, dotted lines). The lines were calculated using the maximum-likelihood estimation procedure. Plus signs (+), boxes, and crosses (×) indicate 35%, 50%, and 65% old words, respectively. H = hit; FA = false alarm.

bias condition into a regression analysis separately, only occasionally were there significant nonlinear trends (13 of the 50 regressions). In no case did the linear component account for less than 93% of the variance. Therefore, the separate zROCs are highly linear. The range spanned by the different bias conditions is the same for all bias conditions (with the exception of Participant 5), and so the slope differences are less likely to have arisen from fitting subsections of a single nonlinear curve.

The mean slopes and intercepts calculated across participants are presented in Figure 4 as a function of bias. The slopes and intercepts of the best-fitting lines were entered into an analysis of variance. The main effect of bias on slope was significant, $F(4, 32) = 3.80$, $MSE = 0.013$. Orthogonal polynomial analyses showed a significant linear trend in the slopes, $F(1, 32) = 13.95$, $MSE = 0.013$, but no higher order trends, indicating a tendency for the slope to increase with increased bias to say "old." The effect of presentation rate on the slope was not significant and did not interact with bias. The main effects of bias and presentation rate on the intercepts were significant, $F(4, 32) = 4.20$, $MSE = 0.014$, and $F(1, 8) = 11.06$, $MSE = 0.321$, for bias and rate, respectively. The intercepts were larger for the longer presentation rate, reflecting an overall increase in d'. Across bias, the intercept showed a significant linear increase, $F(1, 32) = 14.70$, $MSE = 0.014$, but there were no higher order trends.

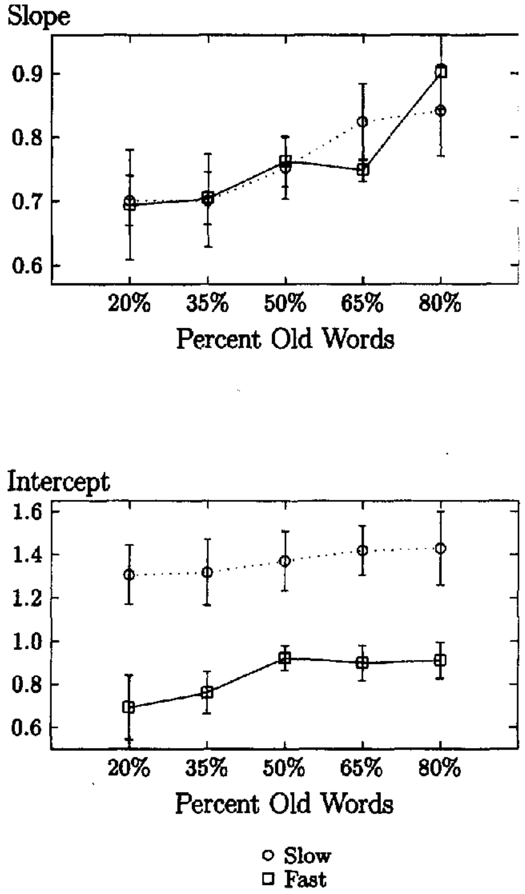

Figure 4. Mean slopes and intercepts (with error bars) as a function of the percentage of old probe words in the test list for the zROC curves shown in Figure 3. The slow condition is plotted with circles, and the fast condition is plotted with squares.

Discussion

The results of Experiment 1 demonstrate that a change in response bias, induced by changing the a priori stimulus probabilities, results in a reliable change in the intercepts and slopes of the zROC function. It is the changes in the slope that are of most interest, because strength- or familiarity-based models of recognition must conclude from this that bias is changing the shapes of the familiarity distributions. The size of this effect differed across participants, but overall there was a significant increase in the slope with increasing bias to respond "old." A similar effect was reported recently by Hirshman and Hostetter (in press), who noted that the slope of the zROC curve tended to increase when criteria were placed lower on the familiarity axis, as would be expected under a bias to say "old."

The small size of the effect is to be expected, for a number of reasons. First of all, the range of slopes varies empirically from around .7 to 1.0 (Glanzer & Adams, 1990; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, 1994). Thus, any variations in slope will be constrained by this range. Second, it often has been difficult to find procedural manipulations that will produce an effect on the slope (Ratcliff et al., 1992, 1994). The most striking failure of this kind, the *null list strength effect,* has proved problematic for all of the global-memory models. The null list strength effect refers to the invariance of the zROC slope with changes in item strength, a finding that was replicated in this experiment (although there are a number of other variables that do result in changes in slope, e.g., Glanzer et al., 1999; Gronlund & Elam, 1994). The fact that the slope changes at all is inconsistent with strength-based models of recognition memory using the criterion hypothesis as a mechanism for confidence.

Experiment 2: Payoffs

The dependence of the zROC curve on probabilities observed in Experiment 1 might be expected if recognition decisions were based on some aspect of perceived odds or likelihood, although, as argued above, such dependence cannot be predicted by the magnitude of perceived odds alone. However, if the changes in the zROC curves are associated with response bias more generally, the same dependencies should be observed when bias is manipulated with payoffs. By using payoffs, in Experiment 2 the odds ratios were held constant across bias conditions, and therefore the results of Experiment 1 were generalized.

Method

All methods, except where noted, were identical to those used in the fast condition of Experiment 1.

Participants. Five English-speaking Johns Hopkins University students participated in this study for pay. All participants reported normal or corrected-to-normal vision, and none had taken part in Experiment 1.

Procedure. Before viewing the test list, the participants were informed that correct "new" responses would be worth 0, 1, 2, 3, or 4 points, whereas correct "old" responses would correspondingly be worth 4, 3, 2, 1, or 0 points (the number of points for correct "old" and "new" responses always summed to 4). No penalty was given for incorrect responses. Participants were instructed to take their time and think carefully about their level of confidence and to try to distribute their responses over all six keys. They were also urged to pay careful attention to the payoff schedule and were told that they should make maximizing points their goal.

Rather than using points to determine a bonus, as in Experiment 1, payment was based solely on points earned, at a rate of 0.78 cent per point. This change from Experiment 1 helped ensure that participants paid attention to the payoff schedule at the beginning of each test list. After completing each test list, the participants were given feedback about overall response accuracy and the number of points earned.

Four of the participants participated in 10 sessions over approximately 2 weeks. One participant performed in 9 sessions over the same interval.

Results and Discussion

The *z*ROC curves were constructed from the confidence judgments for each participant for each of the five levels of bias as described in Experiment 1. Linear functions were fit to each *z*ROC curve for each level of bias using the same maximum-likelihood estimation technique as in Experiment 1. (Simple linear regressions gave similar results.) All reported effects were significant at the $\alpha = .05$ level.

The *z*ROC data for all individual participants are presented in Figure 5, along with the best-fitting lines for the 0 "old points" (solid lines) and four "old points" (dotted lines). Again, the pattern of effects is not easy to see in this figure, but there are significant differences among the slopes for each participant and bias condition. For each participant the slopes tended to be larger under "old" bias.

The changes in slope are not due to selection of different segments of nonlinear *z*ROC functions. Using *z*(Hit) as the dependent variable, regression onto linear, quadratic, and cubic components of *z*(False Alarm) analyses indicated that the higher order trends accounted for significantly more variance than did the linear term alone when collapsing over all bias conditions. However, when entering each bias condition into a regression analysis separately, only 6 of 25 regressions indicated significant nonlinear trends. In no case did the linear component account for less than 96% of the variance. Therefore, the separate *z*ROCs are highly linear. The points along the curves for different bias conditions span the same range for the most part, and so the slope differences are less likely to have arisen from fitting subsections of a single nonlinear curve.

The mean slopes and intercepts calculated across participants are presented in Figure 6 as a function of bias. The main effect of bias on slope was significant, $F(4, 16) = 4.12$, $MSE = 0.006$. Orthogonal polynomial analyses showed a significant linear trend in the slopes, $F(1, 16) = 13.80$, $MSE = 0.006$, but no higher order trends, indicating (as in Experiment 1) a tendency for the slopes to increase with increased bias to say "old." For the intercepts, there was also a significant main effect of bias, $F(4, 16) = 5.03$, $MSE = 0.007$, reflecting the tendency for the intercept to be higher for intermediate levels of bias. This tendency was evident in a significant quadratic trend, $F(1, 16) = 14.86$, $MSE = 0.007$. The linear trend also was significant, $F(1, 16) = 4.65$, $MSE = 0.007$.

The presence of slope and intercept differences in this experiment cannot be explained by models of recognition in which confidence is based on perceived odds or likelihood, because the odds were unchanged over levels of bias. In the next section I outline a mechanism for confidence, the *balance-of-evidence hypothesis,* that accounts not only for the changes in the *z*ROC curve under bias but also for the changes in confidence with response time (RT). The processes hypothesized by the global-memory models establish distributions of familiarity which then drive this mechanism for confidence.

A Model for Confidence

Models designed to account for confidence simultaneously with other behavioral variables (RT and accuracy) were developed as early as the 1940s (Cartwright & Festinger, 1943; Festinger, 1943a, 1943b; Irwin, Smith, & Mayfield, 1956). These models, and the model proposed here, depend on the notion of *sequential sampling* (e.g., Juslin & Olsson, 1997). Sequential sampling models assume that an observer samples a perceptual strength from a stimulus and compares that sample with a criterion, either internal or external. For some models, the difference between the sample and the criterion may be used to make an immediate judgment if it is sufficiently large; otherwise, an additional sample is taken (Cartwright & Festinger, 1943; see also Atkinson & Juola, 1973). Other models assume that the difference is stored and accumulated until the sum of differences becomes large enough to select a response. The accumulation process is very important for most dynamic models of simple (two-choice) decision making, including the random-walk models (Laming, 1968; Link & Heath, 1975; Ratcliff, 1978) and the race models (Audley, 1960; LaBerge, 1962; Pike, 1973; Vickers, 1979).

Both the random-walk and race classes of models assume that evidence arrives at the latest stages of processing supporting one response or the other. The evidence is stored in one of two ways. For the random-walk models the differences between relative amounts of evidence for either response are summed over time. Positive and negative differences push the level of evidence up or down toward one of two response thresholds. For the race models information is stored separately on response "counters." A response is made as soon as the evidence on any counter exceeds a threshold. Both classes of models have been very successful in linking accuracy with RT. By assuming that the accumulation process is noisy, that evidence is identified inappropriately on occasion, and that the time required to integrate the information varies, the relation between RT and

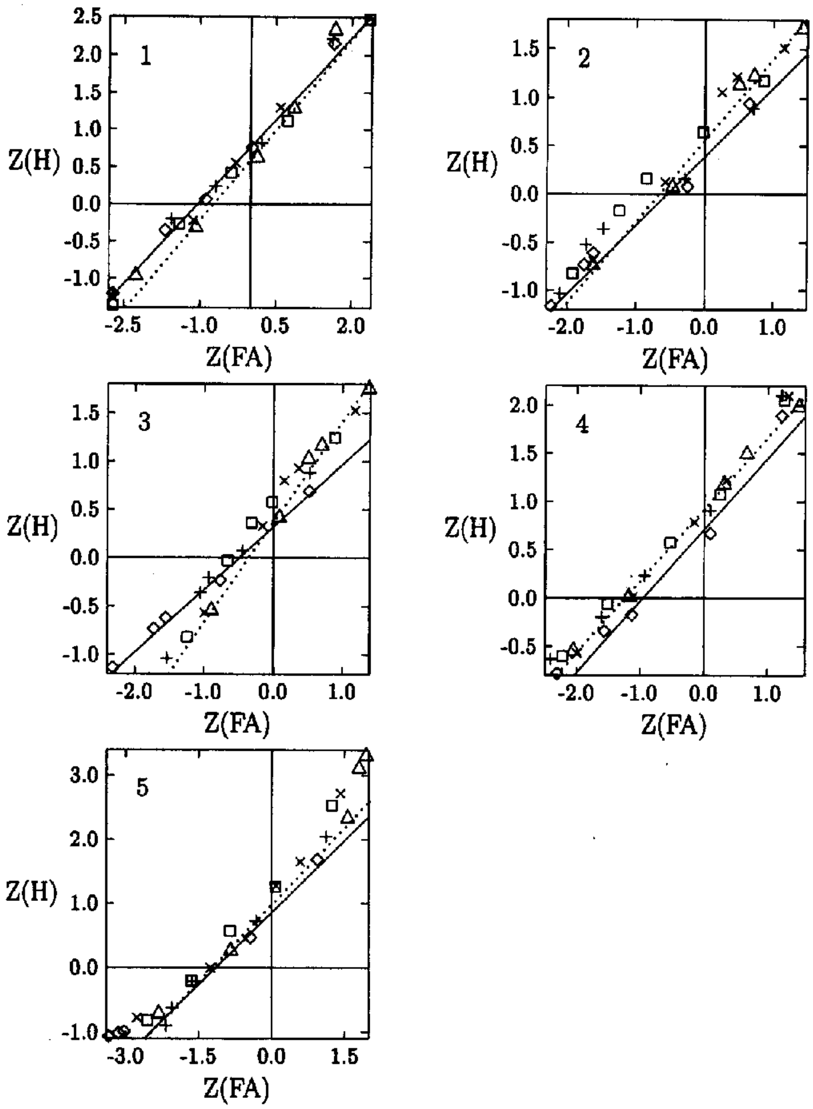

Figure 5. The zROC curves for each participant in Experiment 2. Conditions of response bias vary from 0 points for a correct "old" response (diamonds, solid lines) to 4 points for a correct "old" response (triangles, dotted lines). The lines were calculated using the maximum-likelihood estimation procedure. Plus signs (+), boxes, and crosses (×) indicate 1, 2, and 3 points for correct "old" responses, respectively. H = hit; FA = false alarm.

accuracy can be explained. Low thresholds lead to fast RTs and some errors, whereas high thresholds lead to slower RTs and fewer errors, because of the smaller probability that noise will cause information to accumulate erroneously to a high threshold.

Two proposals have been made to explain confidence in the random walk. One is that confidence is inversely related to the time required to make a response (e.g., Audley, 1960; Volkmann, 1934), a finding that is typical in speeded choice–RT tasks. However, in "expanded-judgment tasks" (Irwin et al., 1956; Vickers, 1979), confidence increases with RT. Also, evidence from experiments using general knowledge tasks, presented by Nelson and Narens (1990), is inconsistent with the idea that confidence is determined by RT. An alternative is that the stimulus intensity sampled, which drives the random walk toward either response threshold, is the basis of the confidence judgment (Link, 1992). It predicts the relationship between low confidence and slow RTs, but it cannot predict the finding of high confidence and slow RTs found in the expanded-judgment task. Also, because this mechanism is functionally equivalent to that specified by the criterion hypothesis, it cannot predict changes in the zROC curves under bias.

Balance-of-Evidence Hypothesis

Vickers and his colleagues (Smith & Vickers, 1988; Vickers, 1979; Vickers, Burt, & Smith, 1985; Vickers &

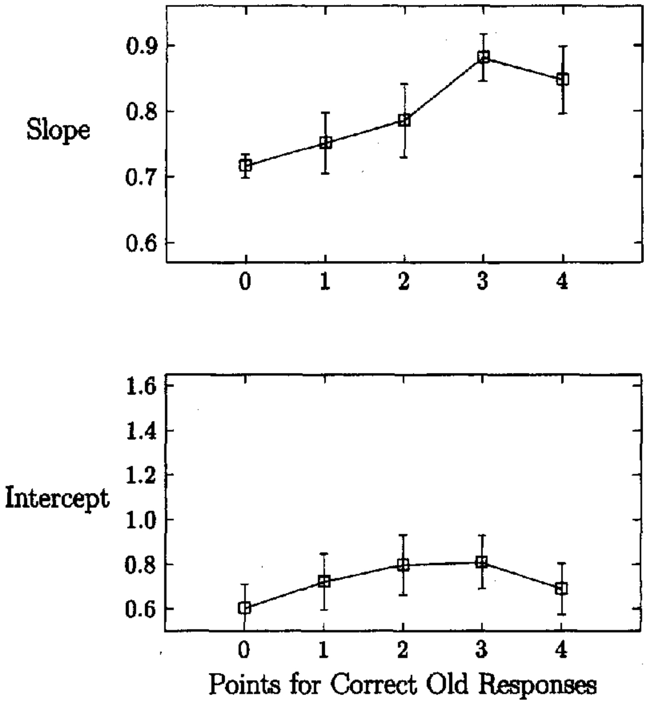

Figure 6. Mean slopes and intercepts (with error bars) as a function of the number of points for a correct "old" response for the zROC curves shown in Figure 5.

Packer, 1982; Vickers, Smith, Burt, & Brown, 1985) have investigated extensively the ability of Vickers's (1979) accumulator model to account for confidence judgments, using what they called the *balance-of-evidence hypothesis.* In the accumulator model perceived stimulus intensities are compared with an internal referent and, if the difference between them is negative, the amount of the difference accumulates on a counter for the negative response; otherwise, it accumulates on a counter for the positive response. A decision is made when one of the counters exceeds a threshold. The balance-of-evidence hypothesis relates the level of confidence to the difference between the two counters at the time of the decision. If both counters have accumulated large amounts of information, the difference between them will be very small, and confidence should be low. On the other hand, if only the winning counter has accumulated a lot of information, the difference between them will be very large, and confidence should be high. Vickers and his colleagues have shown that the balance-of-evidence hypothesis can account for both increases and decreases in confidence with increasing RT, as well as the relationships among RT, accuracy, and confidence.

A Race Model of Recognition Memory

My goal was to incorporate the balance-of-evidence hypothesis into a continuous-time model of recognition decisions (Pike, 1973; Townsend & Ashby, 1983; Van Zandt, Colonius, & Proctor, in press). This model presumes, as does Vickers's (1979) accumulator model, that two counters are established to keep track of evidence building toward the alternative responses "old" and "new." Unlike Vickers's model, which is best suited for the expanded-judgment task, this race model assumes that information is delivered to the counters in discrete units at exponentially distributed interarrival times.[4] The two counters race toward threshold levels of information, and the first to reach threshold determines the response. The time to completion of the race determines RT. This model, sometimes called the *Poisson race model,* was outlined clearly by Townsend and Ashby (1983, chapter 9). It has not previously been applied to recognition memory.

Figure 7 shows the relationship among bias, confidence, and RT in the race model. According to the race model, the decision about whether a probe is old or new is determined by the counter that exceeds threshold first. The top panel of Figure 7 shows evidence being delivered to the accumulator process. Over time, discrete units arrive at exponentially distributed intervals and accumulate on either the "old" counter or the "new" counter. The center panel of Figure 7 shows the state of each counter over time. When an "old" evidence unit arrives, it is added to the accumulated total on the "old" counter. Similarly, the arrival of a "new" evidence unit is added to the accumulated total on the "new" counter. The totals on each counter are described by the two time-dependent random variables, $X_O(t)$ and $X_N(t)$. When one counter exceeds its threshold (K_i, $i = O$ or N), shown here as five units, a response is triggered. In the example shown in Figure 7, the "new" counter exceeds its threshold K_N first, resulting in the response "new," at the time shown on the abscissa. The thresholds may not be equal.

The confidence with which the "new" decision is made is determined by the balance of evidence. At the time (RT) that one of the two counters exceeds threshold, the difference between the two counters, given by $X_N(RT) - X_O(RT)$, determines confidence. Confidence is also a time-dependent random variable, and over the course of a trial it fluctuates up and down as shown in the bottom panel of Figure 7. If the "old" counter had accumulated very few counts, the difference between $X_N(RT)$ and $X_O(RT)$ would be large—close to 5—and hence confidence would be high. If, as is shown in the center panel of Figure 7, the "old" counter is only a count or two shy of exceeding threshold itself, the difference $X_N(RT) - X_O(RT)$ would be smaller, and hence confidence would be lower.

One goal of this modeling effort is to tie the strength-based global-memory models to a dynamic representation of the decision process. The global-memory models explain how distributions of familiarity are established for a memory task. The extension of these models to account for RTs has been left to another stage of processing. In particular, many researchers have suggested that the diffusion process investigated so extensively by Ratcliff (1978, 1980, 1981) might provide a satisfactory means to incorporate the distributions of familiarity with a dynamic decision process (e.g., Gillund

[4] Evidence is assumed to accrue in discrete units for mathematical convenience only. The predictions of the model hold for real-valued increments as well (see, e.g., Rumelhart, 1970).

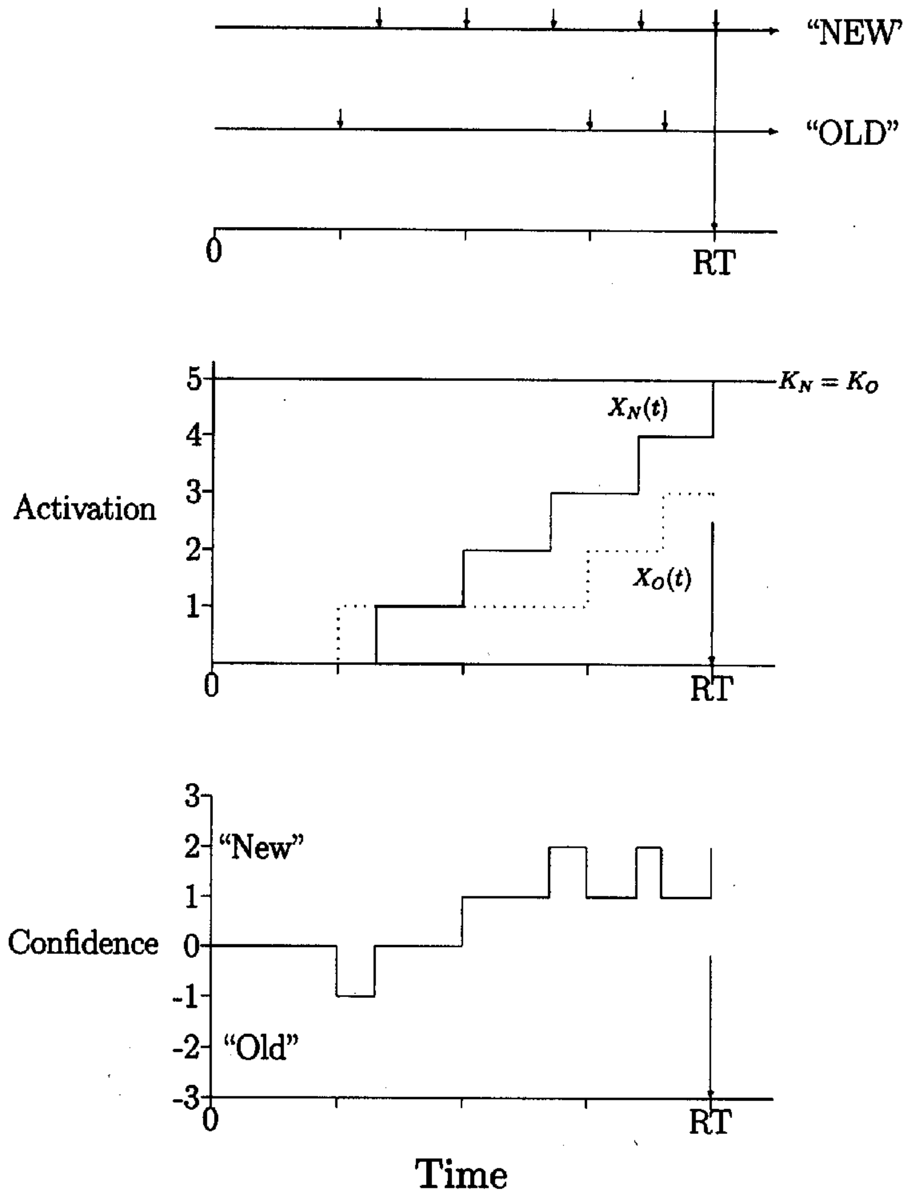

Figure 7. The Poisson race model of recognition memory. In the top panel, "counts" arrive at the old and new counters over time. In the middle panel, the evidence level on each counter is shown as a function of time. As new counts arrive, the variables $X_N(t)$ and $X_O(t)$ are incremented. When one counter [$X_N(t)$] exceeds threshold, a response is emitted at time *RT*. The bottom panel shows the evolution of confidence over time, as given by the balance-of-evidence hypothesis. Positive values represent greater confidence that the probe is new, and negative values represent greater confidence that the probe is old. At the time that the new counter exceeds threshold, the difference between the evidence accumulated on the two counters is 2, leading to a final confidence judgment of 2. RT = response time.

& Shiffrin, 1984; McClelland & Chappell, 1995; Shiffrin & Steyvers, 1997). However, as noted above, because the diffusion process is a member of the random-walk class of models there is no basis for confidence in the diffusion process except probe strength, which cannot explain changes in the *z*ROC with bias.

The units of information accumulated on each counter are presumed to derive from a process in which the familiarity of a probe (or mental representation of a probe) is sequentially sampled. Sampled familiarity derives from a global matching process between the probe and the contents of memory. The sampled strength of the probe determines the rate at which information accumulates on each counter. High levels of familiarity result in rapid accumulation on the "old" counter and slow accumulation on the "new" counter. Similarly, low levels of familiarity result in rapid accumulation on the "new" counter and slow accumulation on the "old" counter.

Regardless of the accumulation rates on the two counters, they are statistically independent from each other. This means that the accumulation process on one counter in no way depends on the amount of information accumulated on the other counter. This characteristic of the race model can be contrasted with the random-walk class of models. Because the random-walk models keep track of evidence differences, the random walk can be represented as two correlated (nonindependent) counters. When evidence is accumulated on a random walk, it means that some amount of evidence was added to one counter and simultaneously subtracted from the other. The independence assumption makes computation of the race model very simple and allows the model to explain confidence judgments.

The Distribution of Confidence

Let C be the confidence level experienced by the rememberer, and define $C = X_N(RT) - X_O(RT)$, where RT indicates the time that the winning counter exceeds threshold. Because the registration of counts forms a Poisson process and the two counters are independent, the distribution of $C = X_N(t) - X_O(t)$ can be determined. For fixed accumulation rates λ_O and λ_N on the old and new counters (respectively), the probability distribution of C can be shown to be (see Appendix):

$$P(C = k|\lambda_N, \lambda_O)$$

$$= \begin{cases} \binom{2K_N - k - 1}{K_N - 1}\left(\frac{\lambda_N}{\lambda_N + \lambda_O}\right)^{K_N}\left(\frac{\lambda_O}{\lambda_N + \lambda_O}\right)^{K_N - k} & \text{if } K_N - K_O + 1 \leq k \leq K_N \\ \binom{2K_O + k - 1}{K_O - 1}\left(\frac{\lambda_O}{\lambda_N + \lambda_O}\right)^{K_O}\left(\frac{\lambda_N}{\lambda_N + \lambda_O}\right)^{K_O + k} & \text{if } -K_O \leq k \leq K_N - K_O - 1 \\ 0 & \text{otherwise} \end{cases} . \quad (2)$$

To link the race model mechanism for confidence judgments to the strength-based global memory models, we must specify the relationship between probe strength and the accumulation rates λ_O and λ_N. To do this, we make use of the superposition of the two processes shown in Figure 7, that is, the Poisson process that results if all counts, "old" and "new," are sent to a single counter. The accumulation rate of this process is equal to $\lambda_N + \lambda_O$, the sum of the accumulation rates for the "old" and "new" counters. If a single count on this new counter is observed, the probability that this count is, say, old, is given by $p_O = \lambda_O/(\lambda_N + \lambda_O)$ (Karlin & Taylor, 1975; Townsend & Ashby, 1983). The terms $\lambda_N/(\lambda_N + \lambda_O)$ and $\lambda_O/(\lambda_N + \lambda_O)$, raised to powers in Equation 2, are the probabilities that an individual count was accumulated on the "new" or "old" counter, respectively. When familiarity x is very large, p_O should be large, and when familiarity is very small, p_O should be very small.

A transformation is sought that maps the unbounded familiarity measure (x) to $p_O \in [0, 1]$. Setting the criterion level of familiarity equal to 0 (Ratcliff, 1978) imposes the additional constraint that a familiarity value of 0 maps to $p_O = 0.5$ (that is, $\lambda_O = \lambda_N$; an ambiguous level of familiarity should result in equal accumulation rates on each counter). A natural choice for this transformation is the logistic function

$$p_O = \frac{1}{1 + e^{-x}}.$$

To make computation simpler, we assume that the value $\lambda_N + \lambda_O = r$ is fixed for all familiarity values. This assumption, which may or may not be reasonable, might reflect a limited amount of capacity in the system or a physical constraint in the maximum firing rate in a population of neurons.[5] Because familiarity is a random variable (X), the accumulation rates also are random variables. The logistic transformation, together with the assumption of a constant r, defines the accumulation-rate variables Λ_O and Λ_N:

$$\Lambda_O = \frac{r}{1 + e^{-X}}, \Lambda_N = r - \Lambda_O.$$

If X is normally distributed with mean μ and standard deviation σ, the density of Λ_O is given by

$$g_{\Lambda_O}(\lambda) = \frac{r}{\lambda(r - \lambda)}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}([-\ln[(r-\lambda)/\lambda]-\mu]/\sigma)^2}. \quad (3)$$

The derivation of Equation 3 is given in the Appendix. The effect of the logistic transformation on the normally distributed familiarity X is shown in Figure 8. In the upper panel are two normal densities, representing the distributions of familiarity. A criterion is placed at $X = 0$, indicating the point at which the familiarity level is ambiguous. The point $X = 0$ is mapped to the point $\Lambda_O = .5r$. In the lower panel are two densities for Λ_O derived from each distribution of familiarity. When the probe presented is new, familiarity values are distributed as in the leftmost density in the upper panel, and Λ_O peaks at accumulation rates less than $.5r$. When the probe presented is old, familiarity values are

[5] Making the assumption that the rates sum to a constant is beneficial for a number of reasons. First, it makes possible direct comparisons between the race model and Ratcliff's (1978) diffusion model of recognition memory, in which familiarity is mapped to a single diffusion rate. Second, allowing both rates to vary would require the specification of a not-insignificant degree of theoretical machinery describing how r varies and how the accumulation rates correlate with each other, if at all. This would also require the specification of at least one more parameter to describe the variation in r. Until such machinery is demanded by the data, I will not speculate about it.

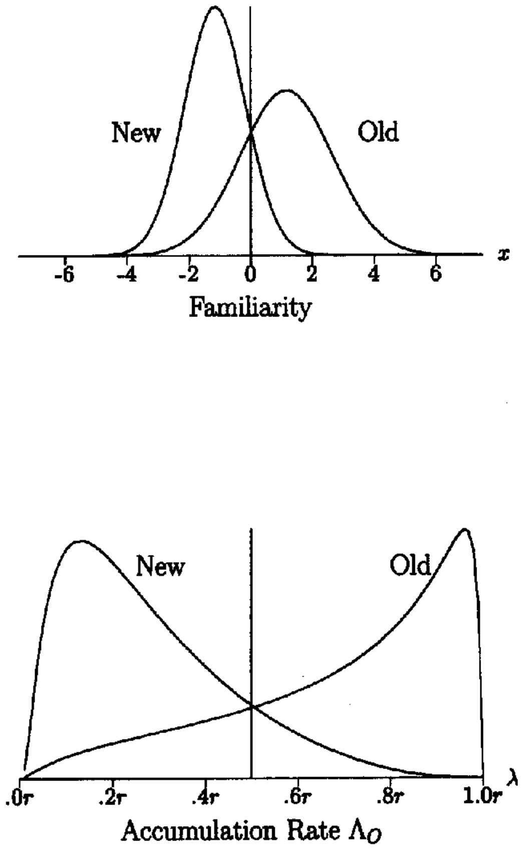

Figure 8. The density functions for Λ_O. When the probe is new, the logistic transformation takes familiarity values from the normal distribution in the top panel labeled "New" to the rates from the distribution in the bottom panel labeled "New." Similarly, if the probe is old, the familiarity values from the normal distribution in the top panel labeled "Old" are mapped to the rates for the distribution in the bottom panel labeled "Old."

distributed as in the rightmost density in the upper panel, and Λ_O peaks at accumulation rates greater than $.5r$. Corresponding mirror-image densities can be constructed for $\Lambda_N = r - \Lambda_O$.

The assumption that the rates sum to a constant r means that only the density of either Λ_O or Λ_N must be considered. Therefore, the marginal probability of observing some confidence level k is given by a mixture distribution:

$$q(k) = \int_0^r P(C = k|\lambda_N, \lambda_O)g_{\Lambda_O}(\lambda_O)\, d\lambda_O$$

$$= \int_0^r P(C = k|r - \lambda, \lambda)g_{\Lambda_O}(\lambda)\, d\lambda.$$

This expression can be integrated numerically to generate the predictions of the model.

Two important facets of this model should be noted. First, under extreme response bias, the direction of the confidence judgment is not necessarily consistent with the counter that won the race. For example, in the center panel of Figure 7, suppose that the "old" threshold K_O were dropped to 2, representing a strong bias to call stimuli "old." In this situation, the response would be "old," and the RT would be reduced to approximately three quarters of *RT*, the response time shown in the figure. However, at this point in time the "new" counter has accumulated one more count than the "old" counter. Therefore, the balance of evidence, shown as a positive difference between $X_N(t)$ and $X_O(t)$ in the bottom panel of Figure 7, favors the "new" decision. If confidence judgments alone are collected, as is done to construct ROC curves, the confidence elicited on this trial would not reflect the counter that actually won the race.[6]

Second, the variable C takes values between $-K_O$ and K_N. This is not, of course, the scale on which rememberers are asked to measure their confidence. Therefore, some transformation f between the *covert* level of confidence C and the experimenter-determined *response* scale R (the confidence judgment actually emitted by the rememberer) must be assumed: $R = f(C)$. For ease of computation and clarity of presentation, no distinction is made between C and any alternative scale: $f(C) = C$. In fact, none is needed. As long as the transformation from C to R is monotonic, preserving the ordinal information in C, and many-to-one, that is, a range of values for C, is mapped to one unique value of R, the zROC predictions of the model based on the variable C must hold. Mappings of C to different confidence scales result in using different points on the cumulative probability function of C to determine the zROC curve. It is important to note that the probability distribution of emitted confidence judgments (R), which is estimated by the data collected in a confidence-judgment procedure, cannot be used to test the truth of Equation 2, because the shape of that distribution is determined by the mapping f. The zROC analysis is independent from the mapping f and is therefore particularly useful in the present context.

The Predictions

To generate predictions for the model, we made a number of assumptions to decrease the amount of necessary computation. One assumption, discussed above, was that of a constant sum of the accumulation rates, r, which was arbitrarily set to .01. The thresholds for each counter were varied from 3 to 12. Changes of bias were modeled by opposing shifts of the thresholds. Under "new" bias, the threshold for counter N was set to 3, whereas for counter O it was set to 12. This relationship was reversed for "old" bias. The accumulation rates λ_O and λ_N depend on the parameters of the old and new familiarity distributions, which have means μ_O and μ_N, respectively, and standard deviations σ_O and σ_N, respectively. The means were assumed to be symmetric around zero; that is, $\mu_O = -\mu_N$. The standard

[6] This implicitly assumes that positive differences in information will be mapped to confidence in one response and negative differences in information will be mapped to confidence in the alternative response. Under bias, zero information difference need not map into ambivalent confidence judgments. For the present, however, I will assume this to be the case. Relaxing this assumption does not change the predictions of the model.

deviation of the new distribution was set equal to 1. There were, then, three parameters manipulated to fit the data: μ_O for the slow and fast study conditions, and σ_O. For the predictions shown, μ_O = .700 for the fast condition (low d') and 1.550 for the slow (high d') condition. The standard deviation of the old distribution, σ_O, was 1.5. The variance difference between the old and new distributions reflects the behavior of the global memory models and captures the idea that whereas items that are not studied have a rather restricted range of familiarity, studied items have a larger range because of encoding variability and differences among items.

The probabilities $q(k)$ over all confidence values were computed for both old and new items and for all possible values of the thresholds K_O and K_N. For each pair of threshold values the probabilities were cumulated across confidence levels and transformed to zROC plots using a normal table, just as would be done if the probabilities were observed in a recognition memory experiment. Even though the confidence distribution is not normally distributed, the percentage of variance accounted for by all the linear, least-squares regressions through the points on the zROC curves were always very high (greater than 99%).

The top panel of Figure 9 shows the slopes of the zROC curves as a function of the threshold for the new counter (the abscissa) and the threshold of the old counter (varying from 3 to 12, as indicated by the label on each separate curve, μ_O = .775). As bias shifts from a tendency to say "new" to a tendency to say "old," the slope increases. How "old" or "new" bias is manifested in the settings of these thresholds is an open question and in all likelihood varies from participant to participant. The point of this plot is to demonstrate that with increasing old bias, whether achieved as the "new" threshold increases and the "old" threshold is fixed or as the "old" threshold decreases and the "new" threshold is fixed, or downward movements of the "old" threshold countered by upward movements of the "new" threshold, the model predicts an increase in slope. Under unbiased conditions, defined here as equal thresholds for each counter, the slope is constrained between .80 (for thresholds equal to 3) and .74 (for thresholds equal to 12). This result coincides nicely with the empirical finding that the zROC curves constructed from confidence judgments in unbiased conditions have slopes around .8.

Using the parameter values as described earlier, the slopes and intercepts for each level of bias were plotted as solid lines in Figure 10 together with the data from Experiments 1 and 2. The extent and direction of change in slopes and intercepts under bias are captured well by the model.

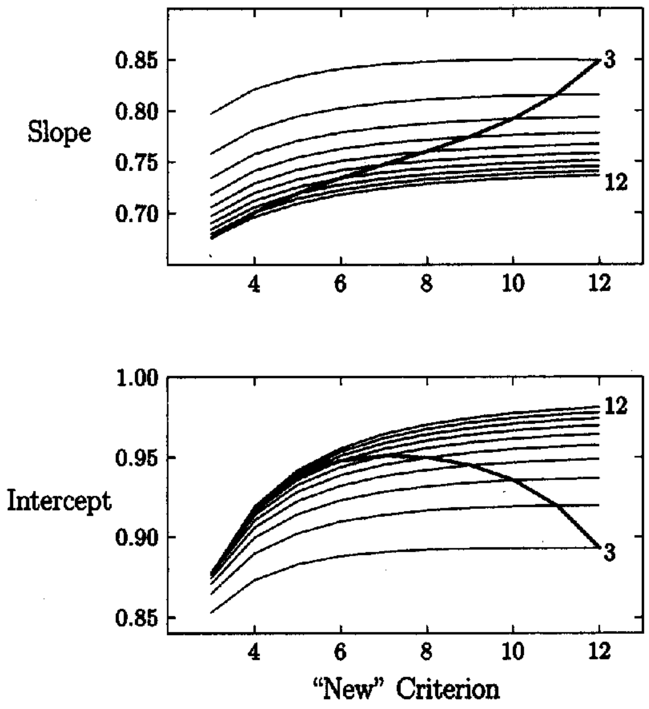

Figure 9. The slopes (top panel) and intercepts (bottom panel) of the zROC curves predicted by the Poisson race model as a function of bias. Bias is indicated by differences between the threshold for the new counter, presented on the abscissa, and the threshold for the old counter, which determines each line on the graph. Each line is ordered as labeled from threshold values of 3 to 12. The dark line on each graph is the change in slope and intercept that would be predicted if response bias results in opposing shifts of the criteria, that is, upward movements of K_N together with downward movements of K_O.

General Discussion

A distinction can be drawn between two types of research involving judgments of confidence in various choice response tasks. This distinction separates research devoted to investigating confidence and its relation to other behavioral variables from research devoted to the details of perception or memory (for which confidence is assumed to exist). Studies examining confidence for its own sake have inspired several recent models of how confidence arises in general knowledge and perceptual tasks (Baranski & Petrusic, 1998; Ferrell, 1995; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin & Olsson, 1997; Wallsten & González-Vallejo, 1994). These models acknowledge the functional differences between a task that requires people to estimate their confidence that a response is correct and a task that requires only the response. Despite a long history of research delineating these differences, many experimental paradigms use confidence only as a tool for measurement; any special characteristics of the information-processing system that confidence judgments might reveal have been ignored.

Research in recognition memory in particular uses confidence as a tool to sweep out ROC and zROC curves. For modeling reasons (i.e., tractibility), the perceived strength or familiarity of a word is assumed to be the sole determinant of confidence. Several criteria are placed along the familiarity or likelihood axis, and confidence is then determined by the region from which a strength of familiarity or likelihood falls.

This simplification of a doubtless-complex process has been quite successful. Models incorporating the criterion hypothesis have been able to explain the relationships

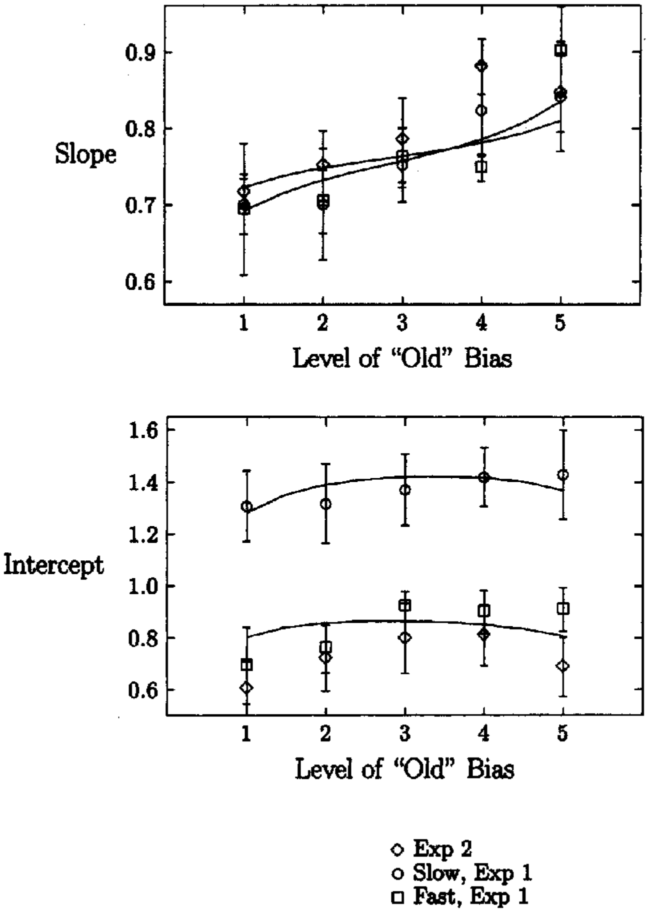

Figure 10. Mean slopes and intercepts (with error bars) over participants for all three experiments as a function of level of "old" bias. For Experiment 1, level of "old" bias is the percentage of old probes, and for Experiment 2 it is the number of points for a correct "old" response. For Experiment 1 the slow condition is plotted with circles, and the fast condition is plotted with squares. Experiment 2 is plotted with diamonds. Exp = Experiment.

among speed, accuracy, and confidence while accounting for an impressive range of data. Any inaccuracies caused by the obvious simplification of confidence have been undetectable relative to the overall success of the models.

The goal of this article was to examine more closely the criterion hypothesis of confidence. A simple prediction of the criterion hypothesis was presented. If confidence ratings are to be used to construct *z*ROC curves, which is an important technique in the analysis of recognition memory data, then bias—the placement of the criteria—should not have any influence on the shape of the curves. Most important, the slopes of the *z*ROC curves, which reflect the ratio of new to old standard deviations, should remain constant across all manipulations of bias. This prediction was tested in two experiments.

In Experiment 1, participants gave confidence judgments under different levels of bias induced by changing the probability that a probe item would be old. Under conditions of high and low discriminability, the slopes of the *z*ROC curves constructed from the confidence ratings were generally lower under "new" bias than "old" bias. These data are problematic for the strength-based, global memory models (Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1982). In Experiment 2 response bias was manipulated by varying payoffs, so that the log odds function was constant across conditions. As in Experiment 1, the slopes of the *z*ROC curves constructed using confidence judgments increased as the bias to call an item "old" increased. Therefore, the criterion hypothesis applied to perceived odds (Glanzer & Adams, 1990; Shiffrin & Steyvers, 1997) is also untenable.

I presented an alternative to the criterion hypothesis, one espoused by Vickers and his colleagues in the context of psychophysical discrimination, called the *balance-of-evidence hypothesis* (Vickers, 1979; Vickers & Packer, 1982; Vickers, Burt, & Smith, 1985; Vickers, Smith, et al., 1985). This hypothesis was embedded in a race model of recognition memory, in which two counters are established to keep track of information favoring "old" and "new" responses. The balance-of-evidence hypothesis states that confidence is determined by the relative difference between the amounts of evidence accrued on both counters at the time that a response is made. This model predicts that the slope of the *z*ROC curve should increase with increasing bias to call an item "old," consistent with the results observed in the experiments.

Biased Versus Unbiased Responding

The present findings confirm that equating confidence with perceived familiarity or likelihood oversimplifies how people estimate confidence. However, the construction of memory *z*ROC curves has a long history in investigations of recognition memory and has been quite successful in accounting for a wide range of effects. Early confirmations of the method showed that *z*ROC curves constructed by way of changing response bias for absolute (old–new) judgments were little different from *z*ROC curves constructed from confidence judgments (cf. Swets, Tanner & Birdsall, 1961). How are past successes with this technique to be explained? The answer is that, as long as rememberers produce confidence judgments in unbiased conditions, confidence and familiarity in recognition memory are correlated highly enough that the simplification is probably warranted.

I will use a signal detection theory metaphor to explain this position. Consider again the Poisson race model shown in Figure 7. When a rememberer is biased to respond "old," the threshold K_O is adjusted downward. If it is pushed as low as two counts, the "old" counter will win the race. If an absolute judgment were required of the rememberer, the response "old" would be emitted. However, if a confidence judgment were required instead, because the "new" counter has accumulated more counts that the "old" counter the confidence judgment would favor the "new" response regardless of the winning counter. From a signal detection theory perspective, it is equivalent to a situation in which a rememberer perceives a level of familiarity to the right of the criterion, which should demand an "old" response, but the rememberer says "new" instead. The effect of bias, then, is to take probability mass from the right of the criterion and redistribute it to the left of the criterion. In essence, the movement of the threshold downward results in a change in the shape of the distributions of familiarity, and so the slope and intercepts of the zROC change under bias.

In contrast, when the rememberer is unbiased, the thresholds should be equal. If $K_O = K_N$, then the level of confidence will always be consistent with the counter that won the race. No redistribution of probability mass will occur, because observations to the right of the criterion will always be given an "old" response, and observations to the left of the criterion will always be given a "new" response. The unbiased case is, therefore, most consistent with the signal detection, multiple-criterion view of confidence. All previous confidence-based zROC curves have been computed from unbiased response conditions, which helps to explain why the multiple-criterion hypothesis has worked so well. However, it is important to recognize that confidence judgments may provide important information about the mechanisms underlying memory that the multiple-criterion hypothesis masks, despite the usefulness of that hypothesis.

Response Times

In this study not only has a new effect been demonstrated that provides a challenge to the strength-based theories of recognition memory, but also a mechanism has been proposed that can simultaneously explain confidence, accuracy, and RT. Because the purpose of this study was to demonstrate that confidence is not scaled directly from familiarity, RTs have not been discussed, and they were not a critical component of the experiments conducted. Participants were instructed to take as much time as they needed to accurately estimate their confidence. Nonetheless, I compared the empirical and predicted RTs.

In this analysis, all individual participant RTs were collected over experiment, bias, probe identity, and confidence level into one group and transformed into z scores. The judgment scale for new words was reversed, so that the scale from 1 to 6 indicated not confidence in probe identity but confidence as a function of accuracy. Therefore, highly confident "old" responses to new words are grouped with highly confident "new" responses to old words, and these two response types are grouped over the label "Incorrect" in Figure 11. There were no significant interactions within bias, confidence level, or discriminability on mean RT, although there were significant main effects of these variables, so grouping the RTs in this way did not influence the pattern shown in Figure 11. The distributions of z scores for each bias condition are shown as box-and-whisker plots in Figure 11 and labeled *O*—the left member of each pair of plots. The same procedure was performed for a simulation of the Poisson race model using the low d' parameters given in the text. The results are shown as box-and-whisker plots in Figure 11 and labeled *M*—the right member of each pair of plots.

The purpose of analyzing the z scores in this way is twofold. First, the pattern of RTs across bias can be compared with the model for all participants as a group, without concern for individual differences in overall RT. Second, the pattern of RTs across bias for the model can be compared with the empirical data, without engaging any elaborate model-fitting scheme to bring the theoretical RTs to precisely the same scale as the observed RTs. The reader should be cautioned, however, that the effects of extraneous variables (such as item and participant effects), and potential

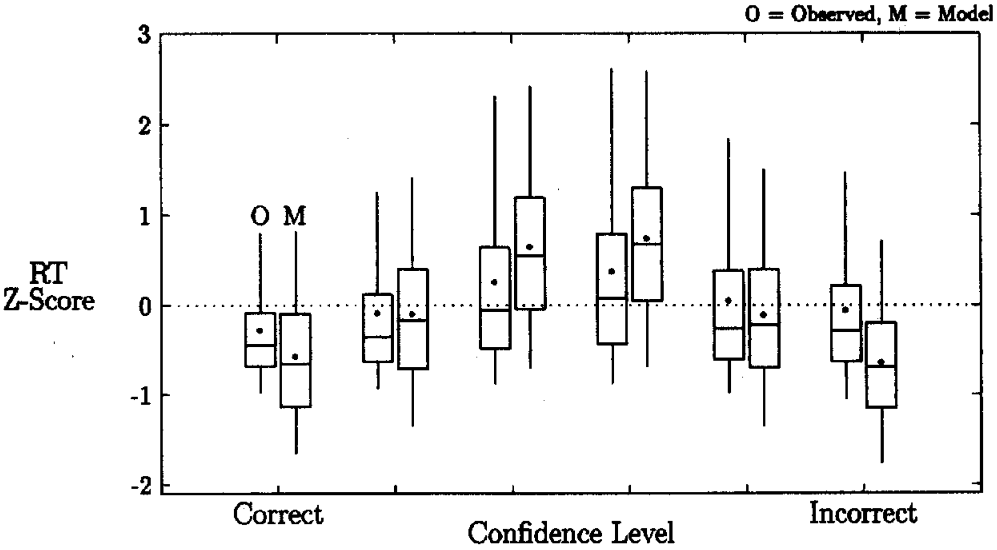

Figure 11. Box-and-whisker plots for the response time (RT) distributions observed in each experiment (O) and as predicted by the Poisson race model (M). The whiskers extend to the 5th and 95th percentiles. The dot represents the mean.

separations between recognition and confidence processes, are glossed over in this procedure.

The box-and-whisker plots in Figure 11 give a feel for how the model accommodates this aspect of the data. The model captures the changes in mean RT over bias, showing faster mean RTs for the extremes of the confidence scale and the slowest RTs for the intermediate levels of confidence. There are no RT differences between extremely confident correct and incorrect judgments for either the model or the data. Also, the variance predicted by the model is approximately that observed in the data. One interesting aspect of the data should be noted in passing: The hypothesis that confidence is based on the time taken to reach a decision (Audley, 1960) can be ruled out on the basis of the large amount of overlap between the RT distributions. If time were the deciding factor in confidence estimation, then these distributions should show very little overlap.

These results are consistent with those of Schreiber, Nelson, and Narens (in Nelson & Narens, 1990), who measured RTs and feeling-of-knowing judgments in a general knowledge task. The inverted-U relationship between RT and the judgment was taken as evidence for two counters. In Nelson and Narens's (1990) model, the two counters store information regarding feeling of knowing (rather than the response to be given to a question). One counter stores evidence for knowing the answer to a question, and the other stores evidence for "knowing not." The knowing-not counter is necessary to explain fast feeling-of-knowing judgments in situations where the rememberer is certain that the answer to a question is not known. Nelson and Calogero (described in Nelson & Narens, 1990) collected data concerning the relationship between confidence and recall latency. Their data showed that confidence levels for correct and incorrect recalls given with the same RTs were quite different, arguing strongly against the hypothesis that confidence is determined by time to respond.

Definitive tests of the model will require investigating the time course of processing in confidence judgments and more careful exploration of the parameter space of the model. Overall, however, the ability of the model to account for RTs and confidence is better than expected, given that a very small number of parameters were varied to produce the *z*ROC predictions (μ_O and σ_O), none of the parameters were selected with the intent of fitting the RT distributions and, indeed, the participants were urged to take as much time as they felt they needed to accurately estimate their confidence.

The Diffusion Process and the Race Model

The difference between the diffusion process and the race model lies in what information rememberers use to select their final decision. In the diffusion process, the first boundary crossed by a sum of information differences determines the response, and the time of that crossing determines RT. In the race model, the first counter to accumulate a threshold absolute level of information determines the response, and the winning time determines RT. Although these two response rules are very different (one is based on differences, and the other is based on overall levels of information), it is not a simple matter to distinguish between the two models on the basis of RT and accuracy alone (Marley & Colonius, 1992).

Confidence judgment data provide a way to distinguish between the two models. If, in the diffusion process, confidence is scaled according to the strength of the probe, then the diffusion process must also predict constant *z*ROC curves under bias. The present results may therefore be taken as evidence against the standard diffusion process model of simple choice in recognition memory. However, there is a simple way to reconcile the tremendous success of the diffusion process model as a model of memory retrieval and its current inability to produce confidence judgments. The difference between the two counters in the race model is, over time, described by a random walk. This particular random walk occurs in continuous time over discrete states representing the difference in the number of counts. When the process stops is determined by the time that one counter exceeds threshold. The response, however, is not necessarily determined by the winning counter. If a confidence judgment is required, then the response is determined by the state of the random walk. Therefore, the present data do not argue against the random-walk mechanism, they require it to be able to explain behavior in both the absolute judgment task and the confidence judgment task.

An alternative model, one that I have not discussed to this point, is the sampling model recently suggested by Juslin and Olsson (1997) for sensory discriminations. This model, similar to Vickers's (1979) accumulator model, has characteristics shared by both the race model and the diffusion model. As in Vickers's model, a perceived intensity difference is sampled from two stimuli (a standard and a comparison stimulus). This difference is a normally distributed random variable with mean $\mu > 0$ when the comparison stimulus is greater than the standard and $\mu < 0$ when the comparison stimulus is less than the standard. In Vickers's model these differences are accumulated on different counters. In the sampling model, however, a fixed number of these differences are averaged or summed, as in a random walk or diffusion. A discrimination response is made when the absolute value of the average exceeds a threshold. The decision is positive or negative depending on the sign of the average. Confidence is based on the proportion of positive (or negative) samples in the average rather than on a balance of evidence between two counters.

Juslin and Olsson (1997) proposed an intriguing new alternative for the basis of confidence in psychophysical judgments, which may be feasible in the context of any random-walk model. However, as Vickers and Pietsch (2000) pointed out, the sampling model fails in a number of fundamental ways. It has difficulty predicting the basic speed–accuracy tradeoff, and it cannot predict the inverse relationship between RT and confidence. Whether the confidence-as-proportion mechanism that Juslin and Olsson proposed is viable in the context of the sampling model will depend on any modifications made to the model to address these problems.

Calibration and Feeling of Knowing

The present results are consistent with findings in other areas of confidence research demonstrating that confidence, as well as feeling of knowing and estimates of future memory performance, is not always based on aspects of information relevant to absolute (old–new) judgment tasks. Calibration research has shown that people are consistently biased to over- or underestimate their performance, and several models have been proposed to explain this effect (Gigerenzer et al., 1991; Wallsten & González-Vallejo, 1994). However, these models do not provide explicit mechanisms for how confidence arises, and they provide no decision-making mechanisms to link confidence, accuracy, and RT.

As an example, Wallsten and González-Vallejo's (1994) stochastic judgment model discriminates between a covert feeling of confidence and the overt confidence judgment in sentence verification tasks. The overt response is based on the level of covert confidence; multiple criteria are placed along the confidence continuum. These criteria are variable, however, which causes variability in the final response that is not due to the judgment process (Wallsten, Bender, & Li, 1999). By modeling criterion location and variance, Wallsten and his colleagues have been able to explain failures of calibration. The race model suggests a mechanism whereby covert confidence is produced and does not speak to issues of calibration. In fact, the issue of scaling between covert and overt responses has been ignored in the present project. The stochastic judgment model picks up where the race model leaves off.

Feeling-of-knowing studies have demonstrated that feeling of knowing is based to some extent on the amount of information that comes to mind about a problem but not on the accuracy of that information (Koriat, 1993). Metcalfe (1993) proposed that feeling of knowing arises from the strength of match between the probe and memory, a hypothesis identical to the multiple-criterion hypothesis of confidence tested here. Nelson and Narens (1990), discussed above, suggested a counter model of feeling of knowing. Although it is tempting to draw parallels between confidence and calibration or feeling of knowing, the tasks in which these variables are measured are considerably different from the simple recognition memory task for which the present model was designed.

Conclusions

Strength-based models of recognition memory that assume that confidence is scaled from perceived familiarity cannot predict changes in the zROC curve observed under bias, whether induced by changes in stimulus probabilities or in payoffs. I propose a dynamic model of decision making in recognition memory that can accommodate these changes and patterns of RTs. According to the model, familiarity drives a process of information accumulation on two independent counters, and confidence arises from the balance of evidence, or difference between the counts at the time that a threshold level of evidence is reached. This hypothesis differs from the diffusion-process explanation in that the decision process does not stop when the accumulated differences cross a boundary but rather when one counter exceeds threshold. This article demonstrates that, apart from being a device to construct ROC curves, confidence judgments can be a useful source of information about structures and processes in memory.

References

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing the speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance IV* (pp. 583–612). New York: Academic Press.

Audley, R. J. (1960). A stochastic model for individual choice behaviour. *Psychological Review, 67,* 1–15.

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

Cartwright, D., & Festinger, L. (1943). A quantitative theory of decision. *Psychological Review, 50,* 595–621.

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

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Rep. No. AFCRC-TN-58-51). Bloomington: Indiana University, Hearing and Communication Laboratory.

Ferrell, W. R. (1995). A model for realism of confidence judgments: Implications for underconfidence in sensory discrimination. *Perception & Psychophysics, 57,* 246–254.

Festinger, L. (1943a). Studies in decision: I. Decision time, relative frequency of judgment, and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology, 32,* 291–306.

Festinger, L. (1943b). Studies in decision: II. An empirical test of a quantitative theory. *Journal of Experimental Psychology, 32,* 411–423.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

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., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500–513.

Grey, D. R., & Morgan, B. J. T. (1972). Some aspects of ROC curve fitting: Normal and logistic models. *Journal of Mathematical Psychology, 9,* 128–139.

Gronlund, S., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355–1369.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Hirshman, E., & Hostetter, M. (in press). Using ROC curves to test models of recognition memory: The relationship between presentation duration and slope. *Memory & Cognition.*

Irwin, F. W., Smith, W. A. S., & Mayfield, J. F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology, 51,* 261–268.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344–366.

Karlin, S., & Taylor, H. M. (1975). *A first course in stochastic processes* (2nd ed.). New York: Academic Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100,* 609–639.

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

LaBerge, D. A. (1962). A recruitment theory of simple behavior. *Psychometrika, 27,* 375–396.

Laming, D. R. (1968). *Information theory of choice reaction time.* New York: Wiley.

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

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

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology, 36,* 1–20.

McClelland, J. L., & Chappell, M. (1995). *Familiarity breeds differentiation: A Bayesian approach to the effects of experience in recognition memory* (Tech. Rep. No. PDP.CNS.95.2). Pittsburgh, PA: Carnegie Mellon University, Department of Psychology and the Center for the Neural Basis of Cognition.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. *Psychological Review, 100,* 3–22.

Murdock, B. B. (1965). Signal detection theory and short-term memory. *Journal of Experimental Psychology, 70,* 443–447.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26,* 125–173.

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73,* 44–58.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review, 80,* 53–68.

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

Ratcliff, R. (1980). A note on modelling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology, 21,* 178–184.

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88,* 552–572.

Ratcliff, R., McKoon, G., & Tindall, M. H. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Rumelhart, D. E. (1970). A multicomponent theory of the perception of briefly exposed visual displays. *Journal of Mathematical Psychology, 7,* 191–218.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin and Review, 4,* 145–166.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology, 32,* 135–168.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34–50.

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

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

Van Zandt, T., Colonius, H., & Proctor, R. W. (in press). A comparison of two reaction time models applied to perceptual matching. *Psychonomic Bulletin & Review.*

Vickers, D. (1979). *Decision processes in visual perception.* New York: Academic Press.

Vickers, D., Burt, J., & Smith, P. (1985). Experimental paradigms emphasizing state or process limitations: I. Effects on speed-accuracy tradeoffs. *Acta Psychologica, 59,* 129–161.

Vickers, D., & Packer, J. S. (1982). Effects of alternating set for speed of accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. *Acta Psychologica, 50,* 179–197.

Vickers, D., & Pietsch, A. (2000). *Decision-making and memory: A critique of Juslin & Olsson's (1997) sampling model of sensory discrimination.* Manuscript submitted for publication.

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.

Volkmann, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin, 31,* 672–673.

Wallsten, T. S., Bender, R. H., & Li, Y. (1999). Dissociating judgment from response processes in statement verification: The effects of experience on each component. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 96–115.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504.

Yonelinas, A. P. (1994). Receiver operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1341–1354.

(Appendix follows)

Appendix

Derivation of Equations

Equation 1

As described in the text, two counters for the "old" and "new" responses are established. Each counter has a threshold number of counts required to initiate a response: K_N and K_O for the new and old counters, respectively. The presentation of a probe causes counts to begin accumulating on the new and old counters with rates λ_N and λ_O, respectively. Let C be the confidence level emitted by the participant, and define C arbitrarily as $X_N(RT) - X_O(RT)$, where RT indicates the time that the winning counter exceeds threshold. (Using $C = X_O[RT] - X_N[RT]$ implies that the confidence scale must be reversed. All calculations and predictions are identical for this alternative definition.)

For fixed rates λ_N and λ_O, the probability that C takes on any value is conditioned on the counter to win the race. If counter N wins, C may take on any value from K_N to $K_N - K_O + 1$. Conversely, if counter O wins, C may take on any value from $K_N - 1 - K_O$ to $-K_O$. Using the law of total probability, the marginal probability of C is

$$P(C = k) = P(C = k \cap O \text{ wins}) + P(C = k \cap N \text{ wins})$$

$$= \begin{cases} P(C = k \cap O \text{ wins}) + 0 & \text{if } -K_O \le k \le K_N - 1 - K_O \\ 0 + P(C = k \cap N \text{ wins}) & \text{if } K_N - K_O + 1 \le k \le K_N \\ 0 & \text{otherwise} \end{cases} .$$

Because the value that C may assume depends on the counter that wins the race, the problem reduces to determining $P(C = k \cap O$ wins) and $P(C = k \cap N$ wins). These probabilities can be computed by looking at the race process in terms of both counters simultaneously. In this situation, the counts arriving on each counter are recorded, and the event that O wins the race is equivalent to waiting for K_O successes in a sequence of $Y = X_N(RT) + X_O(RT)$ Bernoulli trials. When counter O wins, the variable Y has a negative binomial distribution, given by

$$P(Y = y \cap O \text{ wins}) = \binom{y-1}{K_O - 1} p^{K_O}(1-p)^{y-K_O},$$

where p is the probability that a count is old. A similar equation can be written for the distribution of Y when counter N wins by substituting K_N for K_O and $1 - p$ for p.

Given the identity of the winning counter, the random variable C is a linear function of Y. If O wins, $C = X_N(RT) - K_O = Y - 2K_O$. If N wins, $C = K_N - X_O(RT) = 2K_N - Y$. Therefore,

$$P(C = k \cap O \text{ wins}) = P(Y - 2K_O = k \cap O \text{ wins})$$

$$= P(Y = k + 2K_O \cap O \text{ wins}),\ -K_O \le k \le K_N - 1 - K_O.$$

Similarly,

$$P(C = k \cap N \text{ wins})$$

$$= P(Y = 2K_N - k \cap N \text{ wins}),\ K_N - K_O + 1 \le k \le K_N.$$

Substituting $y = k + 2K_O$ and $y = 2K_N - k$ into the expressions for $P(Y = y \cap O$ wins) and $P(Y = y \cap N$ wins), and noting that p is given by $\lambda_O/(\lambda_O + \lambda_N)$ (Karlin & Taylor, 1975; Townsend & Ashby, 1983), the probability distribution of C is given by

$$P(C = k) = \begin{cases} \binom{2K_N - k - 1}{K_N - 1} \left(\frac{\lambda_N}{\lambda_N + \lambda_O}\right)^{K_N} \left(\frac{\lambda_O}{\lambda_N + \lambda_O}\right)^{K_N - k} & \\ & \text{if } K_N - K_O + 1 \le k \le K_N \\ \binom{2K_O + k - 1}{K_O - 1} \left(\frac{\lambda_O}{\lambda_N + \lambda_O}\right)^{K_O} \left(\frac{\lambda_N}{\lambda_N + \lambda_O}\right)^{K_O + k} & \\ & \text{if } -K_O \le k \le K_N - K_O - 1 \\ 0 & \text{otherwise} \end{cases}$$

where the dependence of C on λ_N and λ_O is implicit.

Equation 2

The rate variable Λ_O is related to familiarity X by the logistic function $l(X)$:

$$\Lambda_O = l(X) = \frac{r}{1 + e^{-X}},$$

where r is the sum of the accumulation rates Λ_O and Λ_N. Familiarity is assumed to be normally distributed with means μ_N and μ_O, and standard deviations σ_N and σ_O, depending on whether the stimulus presented is new or old. Let f_X be the normal density, and the rate density is given by

$$g_{\Lambda_O}(\lambda) = f_X(l^{-1}(\lambda)) \left| \frac{d}{d\lambda} l^{-1}(\lambda) \right| .$$

The inverse of $l(X)$ is given by

$$l^{-1}(\lambda) = -\ln(r/\lambda - 1),$$

and its derivative is

$$\frac{d}{d\lambda} - \ln(r/\lambda - 1) = \frac{r}{\lambda(r - \lambda)}.$$

The derivative of l^{-1} is positive for all values of λ. Therefore, substitution into the normal density function gives

$$g_{\Lambda_O}(\lambda) = \frac{r}{\lambda(r-\lambda)} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}([-\ln[(r-\lambda)/\lambda]-\mu]/\sigma)^2}.$$

The normal parameters μ and σ are determined by whether the probe is old or new.

Received January 13, 1999
Revision received August 18, 1999
Accepted August 18, 1999 ■