Journal of Experimental Psychology:
Learning, Memory, and Cognition
1998, Vol. 24, No. 6, 1397–1410

0278-7393/98/$3.00

Decision Rules for Recognition Memory Confidence Judgments

Vincent Stretch and John T. Wixted
University of California, San Diego

According to the standard signal-detection model of recognition memory, confidence judgments for recognition responses are reached in much the same way that old–new decisions are reached (i.e., on the basis of criteria situated along the strength-of-evidence axis). The question investigated here is how the confidence criteria shift when recognition accuracy is manipulated across conditions. Although several theories assume that the old–new decision criterion shifts when recognition accuracy changes, less is known about how the confidence criteria move. An analysis of data previously reported by R. Ratcliff, G. McKoon, and M. Tindall (1994) and some new data reported here suggest that the confidence criteria fan out on the decision axis as d' decreases. This result is qualitatively consistent with the predictions of a likelihood ratio model, although the data did not support the stronger quantitative predictions of this account.

When recognition accuracy is manipulated, study participants are often assumed to change the criterion they use to decide whether a test item is old or new. If, for example, ample time is given to study a list of words, the target items on the recognition test will seem considerably more familiar than the lures. If participants are aware of that fact, then they might reasonably require that a test item seem quite familiar before declaring it to be "old." On the other hand, if learning conditions are less favorable (e.g., if the list items are presented rapidly), then the targets on the recognition test may seem only slightly more familiar than the lures. Under these conditions, participants might require a lower level of familiarity before calling an item "old" (because a lower level of familiarity is compatible with the item having appeared on the list).

Figure 1 illustrates this idea using the standard assumptions of signal-detection theory (Macmillan & Creelman, 1991). This model assumes that the decision axis represents a continuous strength-of-evidence variable, such as familiarity. According to this account, the familiarity values associated with the target items and lure items are both normally distributed, with the mean of the target distribution being situated higher on the decision axis than the mean of the lure distribution. In this example, the variances of the target and lure distributions are equal, but in practice they differ somewhat (Ratcliff, Sheu, & Gronlund, 1992). To arrive at a recognition decision, participants are assumed to set a decision criterion somewhere along the decision axis. Any test item with a familiarity exceeding the criterion is judged to be old ("yes"); otherwise the item is judged to be new ("no"). The ideal location for the decision criterion is midway between the two distributions because that is the point that maximizes the proportion of correct responses. If participants respond in a more or less optimal way, the criterion will be set at a relatively high point on the familiarity axis in the strong condition (Figure 1, upper panel) and at a relatively low point in the weak condition (lower panel). This criterion-shift mechanism has been used to explain changes in the false alarm rate that tend to accompany changes in the hit rate when recognition accuracy is manipulated. That is, an increase in the hit rate is usually accompanied by a decrease in the false alarm rate (a phenomenon known as the *mirror effect*), perhaps because the decision criterion shifts across conditions (Gillund & Shiffrin, 1984).

A straightforward extension of this basic detection model holds that confidence judgments are reached in much the same way that old–new decisions are reached (e.g., Macmillan & Creelman, 1991). That is, as illustrated in Figure 2, a different criterion for each confidence rating is theoretically placed somewhere along the decision axis. Familiarity values that fall above O_H receive an "old" response with high confidence; those that fall between O_M and O_H receive an "old" response with medium confidence; and those that fall between C and O_M receive an "old" response with low confidence. On the other side of the criterion, familiarity values that fall below C but above N_M receive a "new" response with low confidence; those that fall below N_M but above N_H receive a new response with medium confidence; and those that fall below N_H receive a new response with high confidence.

The relationship between the movement of the old–new decision criterion across conditions and the movement of the remaining confidence criteria is not well specified, and a consideration of this issue raises several interesting questions. For example, if the decision criterion shifts between conditions (as in Figure 1), do the confidence criteria shift as

Vincent Stretch and John T. Wixted, Department of Psychology, University of California, San Diego.

We thank Bennett B. Murdock, Jr., Elliot Hirshman, and two anonymous reviewers for their helpful comments on this article.

Correspondence concerning this article should be addressed to either Vincent Stretch or John T. Wixted, Department of Psychology, University of California, San Diego, La Jolla, California 92093. Electronic mail may be sent to John T. Wixted at jwixted@ucsd.edu.

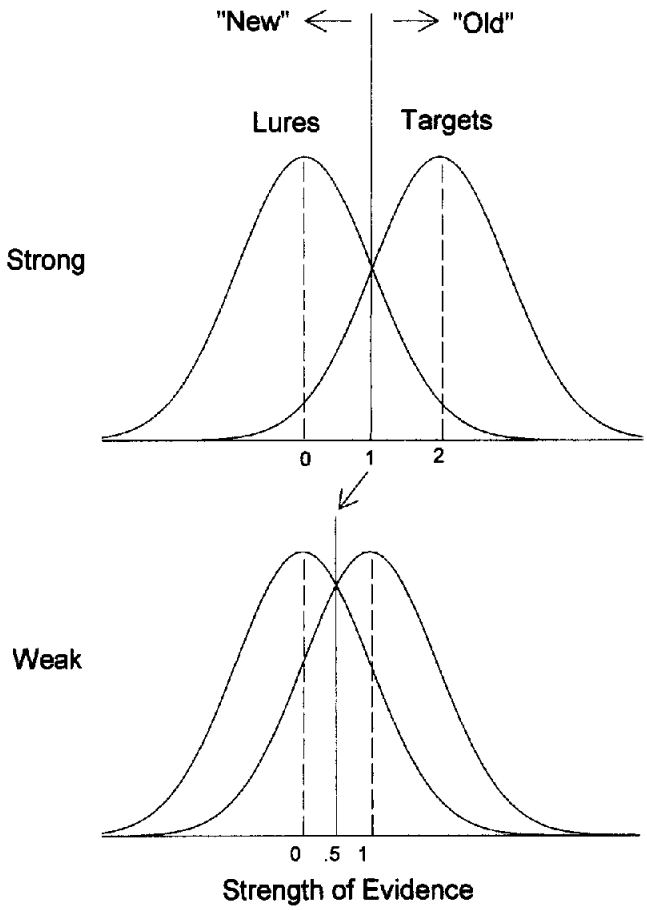

Figure 1. Illustration of a signal detection interpretation of recognition performance in strong and weak conditions. The figure shows the theoretical shift in the placement of the decision criterion across conditions.

well? If so, how do they shift and what accounts for the observed pattern of movement? The next section describes several possible patterns of movement and their implications for theories of recognition memory decision processes.

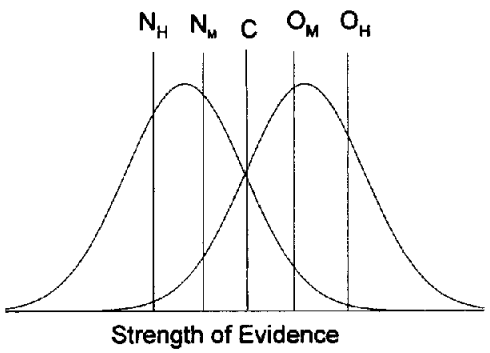

Figure 2. Illustration of the placement of confidence criteria in a signal detection framework. The old–new decision criterion is represented by C, and the remaining criteria represent familiarity values beyond which "old" (O) and "new" (N) responses of higher confidence, medium (M) or high (H), are given.

Decision Rules for Confidence Judgments

Lockstep Model

The simplest model might assume that as the decision criterion moves to the left to stay between the target and lure distributions, the confidence criteria move to the left as well by a corresponding amount. We refer to this model, which is illustrated in the top panel of Figure 3, as the *lockstep model* because all of the criteria move together. This model has also been referred to as the *distance from criterion model* because the position of a particular confidence criterion is always a fixed distance from the old–new decision criterion (Balakrishnan & Ratcliff, 1996). The lockstep model is both computationally simple and intuitively plausible. To use this strategy, a participant only needs to know the means of the target and lure distributions (in order to keep the decision criterion roughly midway between them). When the decision criterion is moved, the confidence criteria are moved by a corresponding amount in the same direction.

Range Model

Alternatively, perhaps the O_H criterion is set some fixed distance above the mean of the target distribution, and the N_H criterion is placed some fixed distance below the mean of the lure distribution. The decision criterion, C, is then placed midway between these two extreme confidence ratings. This model (see middle panel of Figure 3) is similar to a theory that was proposed by Parducci (1984) and recently endorsed by Hirshman (1995). If the participant knows where the mean of the target distribution is relative to the mean of the lure distribution, placing various decision criteria in relation to them should not be difficult. Unlike the lockstep model, which predicts that the various confidence criteria will change in parallel as d' changes, this model predicts that the confidence criteria will converge as d' decreases. That is, the minimum distance between O_H and N_H occurs when $d' = 0$.

Note that if participants were to choose to adopt this strategy they would be behaving in a way that is almost the opposite of ideal. In particular, high-confident errors would increase rather dramatically as d' decreased. A more optimal strategy would involve adjusting the confidence criteria in such a way that high-confident responses would remain highly accurate even as overall performance decreased. The next model predicts that participants will behave in just that way.

Likelihood Ratio Model

If participants maintain constant likelihood ratios for each of the decision criteria, the confidence criteria should diverge rather than converge as d' decreases. According to this account (see bottom panel of Figure 3), the old–new decision criterion, C, is placed in such as way as to maintain a likelihood ratio of 1. Similarly, the O_H criterion is placed in such a way as to maintain a larger likelihood ratio, such as 10 to 1. That is, any item that is 10 or more times as likely to have come from the target distribution than the lure distribu-

tion receives an "old" response with high confidence. Similarly, the N_H criterion is placed in such a way as to maintain a small likelihood ratio, such as 1 to 10 (.1). That is, any item with only a 1 in 10 chance (or less) of having been drawn from the target distribution receives a "no" response with high confidence. As d' decreases, this model assumes that the criteria move in such a way as to maintain these confidence-specific likelihood ratios.

The likelihood ratio model assumes that the participant not only knows the location of the target and lure distributions but also knows their mathematical forms. If the distributions are Gaussian, the target distribution with mean d' and standard deviation σ is given by

$$SN(d', \sigma) = 1/\sqrt{2\pi\sigma^2}e^{-.5(f-d')^2/\sigma^2},$$

and the lure distribution with mean 0 and standard deviation σ is given by

$$N(0, \sigma) = 1/\sqrt{2\pi\sigma^2}e^{-.5(f-0')^2/\sigma^2},$$

where f is a value along the familiarity axis. Using this knowledge, the old–new decision criterion, C, is placed at the point on the familiarity axis where the height of the signal distribution equals that of the noise distribution (i.e., where the likelihood ratio is 1). This is the optimal location of the decision criterion because this is where the odds that the test item was drawn from the target distribution exactly equal the odds that it was drawn from the lure distribution. For any familiarity value greater than that, the odds that the item was drawn from the target distribution (i.e., that it appeared on the list) are better than even, in which case an "old" response makes sense. As indicated earlier, the remaining confidence criteria are placed in such a way as to maintain specific odds ratios that are greater than or less than 1. If O_H is associated with a likelihood ratio of 10 to 1, then that confidence criterion is placed on the familiarity axis at the point where the height of the target distribution is 10 times the height of the lure distribution regardless of the value of d'.

For any given likelihood ratio, L, the criterion is placed on the familiarity axis at the point f that satisfies the following equation:

$$L = [SN(d', \sigma)|f]/[N(0, \sigma)|f]. \quad (1)$$

The right side of the equation is simply the ratio of the height of the target (signal plus noise) distribution to the height of the lure (noise) distribution at the point f. Solving this equation for f yields

$$f = d'/2 + \ln(L)/d'. \quad (2)$$

To determine where a particular criterion is placed, one need only enter the value of L theoretically associated with that criterion. For example, theoretically, the old–new decision criterion is placed at the point where $L = 1$. Substituting 1 for L (and C for f) in Equation 2 reveals that $C = d'/2$, which we already knew to be true. That is, according to this model, the decision criterion should be placed midway between the target and lure distributions no matter what the value of d' is. If O_H is placed at the point where the odds are 10 to 1 in favor of the item having been drawn from the target distribution, then Equation 2 indicates that

$$O_H = d'/2 + \ln(10)/d'$$

$$= d'/2 + 2.30/d'.$$

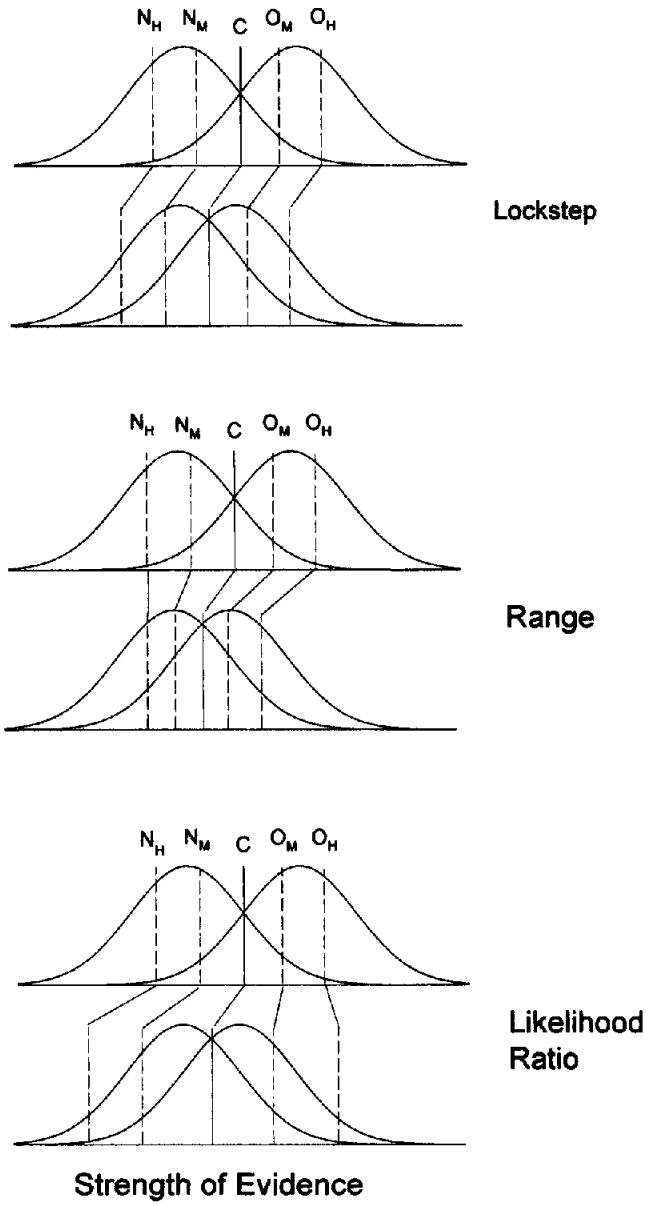

Figure 3. Illustration of the movement of the confidence criteria as a function of d' according to the lockstep, range, and likelihood ratio models (upper, middle, and lower panels, respectively). In each panel, the top figure represents a strong condition and the bottom figure represents a weak condition. N = new; H = high; M = medium; C = decision criterion; O = old.

In other words, O_H is placed at a point higher than $d'/2$ but by an amount that varies with d'. The more general version of this equation is $O_H = d'/2 + k_1/d'$, where k_1 is the constant log likelihood ratio associated with O_H. Similarly, if N_H is placed at the point where the odds are only 1 in 10 in favor of the item having been drawn from the target distribution, then Equation 2 indicates that

$$N_H = d'/2 + \ln(0.10)/d'$$

$$= d'/2 - 2.30/d'.$$

The more general version of this equation is $N_H = d'/2 + k_2/d'$, where k_2 is the constant log likelihood ratio associated with N_H. This model predicts that the distance between O_H and N_H on the familiarity axis will be inversely related to d'. That distance (i.e., O_H minus N_H) is given by

$$O_H - N_H = (d'/2 + k_1/d') - (d'/2 + k_2/d')$$

$$= (k_1 + k_2)/d'.$$

Thus, as d' becomes very large, O_H and N_H should converge to the point that they coincide on the evidence axis (i.e., the distance between them should decrease to 0). By contrast, as d' approaches 0, those two criteria should move infinitely far apart.

The only previous research directly addressing this issue was recently reported by Balakrishnan and Ratcliff (1996). They were mainly concerned with testing whether their data were more accurately predicted by the distance-from-criterion (i.e., lockstep) model or the ideal observer (i.e., likelihood ratio) model when strength was manipulated, and their findings were more consistent with the former account. Our method of analysis differs somewhat from that used by Balakrishnan and Ratcliff in that we use the predictions of the lockstep model as a benchmark to gauge the predictions of the other models. That is, instead of asking which model offers the most accurate quantitative predictions, we ask whether the data exhibit statistically significant deviations from the predictions of the lockstep model and, if so, whether the observed deviations are in the direction predicted by the range model or the likelihood ratio model. In a later section, we consider how the two methods of analysis might lead to different conclusions.

For the time being, our analysis is based on the assumption that the signal-detection decision axis represents a strength-of-evidence variable, such as familiarity (cf. Kintsch, 1967), because that is the simplest and most common assumption. The alternative possibility is that the decision axis represents a log likelihood ratio scale. This is a central assumption of Glanzer's attention likelihood theory (Glanzer & Adams, 1985; Glanzer, Adams, Iverson, & Kim, 1993) and will be discussed in a later section.

The kind of experiment needed to differentiate between the three models shown in Figure 3 is quite simple. Participants need to be exposed to pure strength lists (weak vs. strong) followed by a recognition test in which confidence ratings are taken. As described later, these ratings can be used to estimate the locations of the various confidence criteria in the two strength conditions. Pure strength manipulations are of most interest here because participants presumably are well aware of the effects of that manipulation on the strength of targets. That being the case, they have all the information they need to adjust the decision criterion as shown in Figure 1. Other manipulations that also affect d', such as a word frequency manipulation, may differ in this respect. That is, participants may not appreciate the fact that low frequency (LF) words are more recognizable than high frequency (HF) words (Greene & Thapar, 1994; Wixted, 1992). Indeed, perhaps for that reason, some recent evidence suggests that participants do not use a different decision criterion for HF and LF words (Hirshman & Arndt, 1997; Stretch & Wixted, 1998). If the decision criterion does not shift across conditions, the question of how the confidence criteria shift does not arise.

A recent experiment by Ratcliff, McKoon, and Tindall (1994) provided a wealth of pure-strength data that can be used to test the models shown in Figure 3. That article was not actually concerned with how the confidence criteria shift as a function of strength but instead was concerned with the effect of various manipulations on the slope of the receiver operating characteristic (ROC). Nevertheless, the data presented in the appendix of that article can be used to test the predictions of the lockstep, range, and likelihood ratio models.

Of the six experiments reported by Ratcliff et al. (1994), the first three involved substantial manipulations of strength and a correspondingly large shift in the location of the decision criterion (as evidenced by a substantial difference in the false alarm rates associated with the two pure-strength conditions). The last three involved less substantial manipulations of strength and correspondingly slight movements in the location of the decision criterion (i.e., the false alarm rates in the strong and weak conditions did not differ by much). Thus, we concentrate mainly on the first three experiments and discuss the remaining three in less detail. Experiments 1 and 2 of Ratcliff et al. (1994) were based on data pooled over participants, and we consider those first. In a subsequent section, we analyze the single-participant data from Experiment 3 of that article. In addition, because those results are somewhat ambiguous, we report new single-subject data as well.

Group Data Analyses: Experiments 1 and 2 of Ratcliff et al. (1994)

In the first experiment reported by Ratcliff et al. (1994), 16 participants were exposed to lists of 32 words, some of which were weak and some of which were strong. In pure weak lists, the items were presented for 50 ms each; in pure strong lists, they were presented for 200 ms each. Following list presentation, participants were exposed to a standard yes–no recognition test involving 32 targets randomly intermixed with an equal number of lures. Participants responded to each item on a 6-point confidence scale consisting of 6 = *sure old*, 5 = *probably old*, 4 = *maybe old*, 3 = *maybe new*, 2 = *probably new*, and 1 = *sure new*. The

second experiment was similar to the first except that it involved 15 participants and the presentation times in the pure weak and pure strong conditions were 100 and 400 ms, respectively.

Table 1 shows the group hit and false alarm rates for the weak and strong conditions from these two experiments. A standard mirror effect was observed, which is consistent with the idea that participants set a higher criterion in the strong condition than in the weak condition (as in Figure 1). As described in more detail later, confidence-based ROC analyses were performed on these data to determine how the remaining confidence criteria shifted as well. The values of d shown in Table 1 represent accuracy measures similar to d' that were derived from those ROC analyses.

ROC Analyses

Estimates of where the various confidence criteria are placed along the decision axis in a given strength condition can be obtained in exactly the same way that the placement of the old–new decision criterion is usually obtained. In general, d' is computed from a participant's hit and false alarm rates. Using those values, one can also determine response bias, β (the likelihood ratio above which a yes response is given), and C (the placement of the decision criterion relative to the mean of the lure distribution). For example, symmetrical hit and false alarm rates of .84 and .16 yield a d' of 2.0, a β of 1, and a criterion placement (C) of $d'/2$ (i.e., $C = 1.0$ in this example). As with d', this value of C represents the criterion's distance from the mean of the lure distribution in standard deviation units. It can be computed by simply converting 1 minus the false alarm rate into a z score. If the decision criterion is placed one standard deviation above the mean of the lure distribution (as in this example), then 16% of that distribution falls to the right of the criterion. Hence, a false alarm rate of .16.

This method of computing the placement of C involves all "old" responses to lures with a confidence rating of *maybe old* or greater (which is to say, all "old" responses to lures). To determine where the O_M confidence criterion is placed, one can use only those responses to lures that received a confidence rating of *probably old* or greater. From these responses, a new false alarm rate can be computed and the location of the O_M criterion can be estimated by converting 1 minus that false alarm rate into a z score (exactly as one usually does for "old" responses that exceed the old–new decision criterion). Although we did use this procedure as a check on our analyses, in practice, more precise estimates of the confidence criteria locations were obtained using a maximum likelihood analysis of the ROC curves that also takes account of responses on target trials (rather than just analyzing confidence-specific false alarms). A detailed description of the method we used is provided by Ogilvie and Creelman (1968). This method simultaneously estimates the placement of the various decision criteria (O_H, O_M, C, N_M, and N_H), d (a discriminability measure similar to d'), and r (the standard deviation of the lure distribution relative to the target distribution) using the logistic approximation of the Gaussian distribution.[1] The parameters r and d can be thought of as the slope and intercept of the ROC plotted in z-transformed coordinates. The r parameter is included because prior research clearly shows that the target and lure distributions in memory experiments do not have equal standard deviations (Ratcliff et al., 1992). Note that all of the criteria parameter estimates are scaled with respect to the mean of the lure distribution (which is arbitrarily assigned a value of 0).

Ratcliff et al. (1994) showed that the z-transformed ROC data were essentially linear, indicating that the assumption of Gaussian target and lure distributions is a reasonable one (although other distributions can be found to fit these data as well; Lockhart & Murdock, 1970). For all of the analyses described later, the quality of fit (and the parameter estimates) are virtually identical whether underlying Gaussian or logistic distributions are assumed.[2] The estimates we present are based on the logistic approximation because they are the same values produced by a commonly used program called *EPCROC* (e.g., Ratcliff et al., 1994; see Ogilvie & Creelman, 1968, for program).

Table 2 presents the maximum likelihood estimates of the fitted parameters for the pure-strength manipulations in Experiments 1 and 2 of Ratcliff et al. (1994), respectively. An unsurprising finding is that, for both experiments, the decision criterion, C, was estimated to lie at a higher point on the decision axis in the strong condition than in the weak condition. This merely reflects the higher false alarm rate in the weak condition of both experiments. Of more interest here is how the remaining confidence criteria shifted with recognition accuracy. As shown in Table 2, the strength manipulation produced a fan pattern that more or less conforms to the predictions of the likelihood ratio model. Thus, for example, in Experiment 1, O_H was estimated to lie 1.52 standard deviations above the mean of the lure distribution in the strong condition and 1.64 standard deviations above the mean of the lure distribution in the weak condition

Table 1
Group Hit and False Alarm Rates From Experiments 1 and 2 of Ratcliff et al. (1994)

	Experiment 1			Experiment 2		
Condition	Hit	FA	*d*	Hit	FA	*d*
Weak	0.39	0.35	0.08	0.42	0.31	0.26
Strong	0.59	0.30	0.57	0.64	0.25	0.97

Note. Hit = mean hit rate; FA = false alarm rate; d = discriminability.

[1] The d parameter in these fits is equal to 0.61 times d^*, where d^* is the intercept of the linearized ROC, assuming logistic distributions (Ogilvie & Creelman, 1968). The multiplying factor gives a close approximation of the value that would be obtained assuming Gaussian distributions. The confidence criteria were also multiplied by 0.61 for the same reason.

[2] The Gaussian fits were performed by replacing the equation for the cumulative logistic with an extremely close approximation of the cumulative Gaussian. That approximation is given by Equation 26.2.17 in Abramowitz and Stegun (1965).

Table 2
Group ROC Parameter Estimates From Experiments 1 and 2 of Ratcliff et al. (1994)

Confidence criterion	Experiment 1			Experiment 2		
	Weak	Strong	Difference	Weak	Strong	Difference
O_H	1.64	1.52	+0.12	1.58	1.60	−0.02
O_M	0.97	0.94	+0.03	0.94	1.05	−0.11
C	0.38	0.52	−0.14	0.48	0.67	−0.19
N_M	−0.33	0.00	−0.33	−0.41	0.03	−0.44
N_H	−1.40	−0.91	−0.49	−1.62	−0.80	−0.82

Note. ROC = receiver operating characteristic; O = old; H = high; M = medium; C = decision criteria; N = new.

(a difference of +0.12, strong to weak). That is, participants actually required a higher level of familiarity before giving a high-confident "old" response in the weak condition. The decision criterion, C, moved in the opposite direction from 0.52 to 0.38 (a difference of −0.14), and the N_H criterion moved in the negative direction more than that, from −0.91 to −1.40 (a difference of −0.49). The pattern was much the same in Experiment 2. These difference scores are consistent with the fan pattern predicted by the likelihood ratio model (lower panel of Figure 3).

To see whether this apparent fan pattern differs significantly from a lockstep pattern, the data from each experiment were fit twice, once allowing separate confidence criteria estimates for the strong and weak conditions and once again requiring that the criteria remain a fixed difference apart (that distance being represented by the parameter Δ). For both fits, the d and r values for the strong and weak conditions were fixed at the values obtained from the fits described earlier. Note that the fit of the full model involved 10 free parameters (O_H, O_M, C, N_M, and N_H for the weak condition and a similar set of parameters for the strong condition), whereas the fit of the reduced model involved 6 free parameters (O_H, O_M, C, N_M, N_H, and Δ). For the latter fit, the confidence criteria in the strong condition were constrained to differ from the corresponding criteria in the weak condition by the same amount (Δ), but the parameters were otherwise free to vary. The difference in the chi-square goodness-of-fit between these two fits is itself distributed as chi-square with 4 degrees of freedom. If the addition of the 4 extra parameters in the full model was gratuitous, the expected difference in chi-square would be 4 (which would not be significant). Instead, the chi-square value was 152.7 for Experiment 1 and 190.9 for Experiment 2, both of which are highly significant. Thus, the null hypothesis of a lockstep shift in the criteria can be rejected. The range model described earlier can also be rejected because that model requires that the criteria converge (rather than diverge) as strength decreases.

The interpretation of the foregoing results depends on several important assumptions. As indicated earlier, we assume Gaussian (or logistic) distributions and a strength-of-evidence decision axis. Even more important, we assume that the parameters of the lure distribution in the weak condition match the parameters of the lure distribution in the strong condition. That is, we assume that the mean and standard deviation of the lure distributions in the two conditions are the same (assumptions hereinafter symbolized as $\mu_{Weak} = \mu_{Strong}$ and $\sigma_{Weak} = \sigma_{Strong}$). The assumption that $\sigma_{Weak} = \sigma_{Strong}$ is especially critical because the location of the confidence criteria in the strong and weak conditions are scaled in units equal to the standard deviation of their respective lure distributions. If the standard deviations of the two lure distributions are not equal, then the two sets of estimates cannot be directly compared (it would be like comparing one set of lengths measured in inches with another set of lengths measured in centimeters).

The idea that the strong and weak lure distributions have the same standard deviations is intuitively plausible because the items are physically identical, on average. That is, in both cases, the lures are simply words that did not appear on the list that were drawn randomly from the word pool. This assumption also has some theoretical justification. Specifically, Search of Associative Memory (SAM) assumes that strength manipulations do not affect the characteristics of the lure distribution. Shiffrin, Huber, and Marinelli (1995) recently presented empirical evidence in support of this assumption by showing that when steps are taken to prevent criterion shifts, strengthening target items does not affect the false alarm rate (as it should if that manipulation affected the familiarity characteristics of the lures). Thus, although the preceding analysis depends on a fairly strong assumption, that assumption can be made with some justification.

The parameter estimates just analyzed were based on responses to both targets and lures (i.e., on both hit and false alarm rates). It could be argued that the inclusion of targets in the analysis creates distortions because of the possible role for all-or-none retrieval processes in recognition memory (e.g., Gardiner & Java, 1990). Because a continuous strength-of-evidence axis was assumed in the preceding analyses, the presence of all-or-none retrieval of some of the targets during recognition testing could complicate the analysis in unknown ways. However, as indicated earlier, estimates of where the various criteria are located can also be obtained by using the confidence-specific false alarm rates rather than the maximum likelihood procedure described earlier. Thus, for example, if 2.5% of the lures received a high-confident "old" response (i.e., *sure old*), then the location of the O_H confidence criterion would be 2 standard deviations above the mean of the lure distribution (because that location would leave 2.5% of the lure distribution to the right of the

O_H criterion). This value is arrived at by converting 1 − .025 to a z score. For all of the data considered here, the z score estimates of the locations of the confidence criteria are very similar to those obtained from the full maximum likelihood analysis, and the fan effect predicted by the likelihood ratio account is still plainly apparent. Note that this is true even though the analysis excluded targets from the analysis altogether. Thus, if some of the target items were retrieved in an all-or-none fashion during recognition testing, that fact apparently did not appreciably distort the maximum likelihood parameter estimates.

Unequal Variance

The predictions of the likelihood ratio model were made on the basis of an equal variance model, but as indicated earlier, ROC analyses of recognition memory data almost always suggest that the variance of the target distribution is greater than that of the lure distribution (i.e., r is less than 1). Predictions of the likelihood ratio model in the unequal variance case are much less straightforward than in the equal variance case because, for most likelihood ratios, no unique solution exists. For example (and as described in more detail in the Appendix), if r is less than 1, then to maintain any particular likelihood ratio greater than 1, a confidence criterion could be placed at either of two locations on the familiarity axis. In addition, some likelihood ratios of less than 1 do not exist because the unequal variances (specifically, the greater target distribution variance) places a constraint greater than zero on the lowest possible likelihood ratio. In the equal variance case, by contrast, any likelihood ratio from zero to infinity is possible, and every particular likelihood ratio corresponds to a single point on the familiarity axis. For these reasons, the idea that participants base likelihood ratio decisions on an unequal variance model seems implausible.

The simplest way out of this apparent dilemma is to assume an equal-variance likelihood ratio model even though the slope of the ROC is typically less than 1. Indeed, Yonelinas (1994) explicitly endorsed this assumption and argued that the slope of the ROC is less than 1 not because of unequal variances (as is usually assumed) but because of the contribution of all-or-none retrieval processes. If the target and lure distribution variances are equal in spite of the fact that r is less than 1, then responding on the basis of likelihood ratios seems more plausible. An alternative possibility is that the variances are indeed unequal and that the criteria fan out for reasons having nothing to do with an actual likelihood ratio computation. For example, perhaps participants have simply learned on the basis of everyday experience that more stringent confidence criteria settings are required when memory conditions change for the worse (otherwise high-confident errors would increase). Such experience might not train participants to maintain constant likelihood ratios exactly, but it might teach them to behave more or less in accordance with what a computational likelihood ratio model would predict.

The results shown in Table 2 suggest that the confidence criteria fan out on the decision axis, at least to some extent, as d' decreases. However, one possible concern about these experiments is that all of the analyses were based on group data. Perhaps the apparent likelihood ratio fan is an artifact of averaging over participants. The data from Experiment 3 of Ratcliff et al. (1994) can be used to test the same idea using single-participant analyses. We also analyze single-participant data from an experiment that we performed. This new experiment involved a much greater manipulation of strength than that reported by Ratcliff et al. (1994) and yielded somewhat more definitive results.

Single-Subject Analyses: Experiment 3 of Ratcliff et al. (1994) and New Data

In Ratcliff et al.'s (1994) Experiment 3, seven participants were tested for up to 10 sessions each so that individual ROC analyses could be performed. Participants in this experiment studied lists of 16, 32, or 64 items, and the items were presented at a 1-s rate in the pure weak condition and at a 3-s rate in the pure strong condition. Each list was followed by a yes–no recognition test, and confidence ratings were measured on a 6-point scale as before. We collapsed across the list length manipulation for the analyses described later.

ROC analyses were performed as described earlier, but each participant's data were analyzed separately. That is, for each participant, the maximum likelihood estimation procedure yielded an estimate of d, r, and five criteria (O_H, O_M, C, N_M, and N_H). Table 3 shows the average parameter estimates obtained from the individual participant fits. As in the previous analyses, it appears that the criteria do not move in lockstep, nor do they converge. Instead, they diverge as predicted by the likelihood ratio model. That is, although the criterion, C, shifted left from 0.79 to 0.56 standard deviation units above the mean of the lure distribution (strong to weak), N_H shifted to a greater degree and O_H actually moved slightly in the opposite direction. A repeated measures analysis of variance (ANOVA) performed on the individual participant parameter estimates revealed a marginally significant main effect of strength, $F(1, 6) = 4.79$, $MSE = 0.168$, $p = .071$; a main effect of level of confidence, $F(4, 24) = 48.0$, $MSE = 0.219$, $p < .05$; and (most important for present purposes) a just barely significant interaction between those two factors, $F(4, 24) = 2.80$, $MSE = 0.026$, $p < .05$.

Table 4 shows the difference scores for the five confidence criteria and for the discriminability parameter for each

Table 3
Mean of Individual-Participant Parameter Estimates From Experiment 3 of Ratcliff et al. (1994)

Confidence criterion	Weak	Strong	Difference
O_H	1.77	1.76	+0.01
O_M	1.02	1.20	−0.18
C	0.56	0.79	−0.23
N_M	0.09	0.38	−0.29
N_H	−0.71	−0.33	−0.38

Note. O = old; H = high; M = medium; C = decision criterion; N = new.

Table 4
Difference Scores (Weak–Strong) for Individual-Participant Parameter Estimates From Experiment 3 of Ratcliff et al. (1994)

Parameter estimate	Participant						
	1	2	3	4	5	6	7
O_{HW}–O_{HS}	+0.07	+0.06	−0.05	−0.13	−0.07	−0.02	+0.19
O_{MW}–O_{MS}	−0.39	−0.29	−0.04	−0.23	−0.12	−0.11	+0.06
C_W–C_S	−0.82	−0.20	+0.02	−0.25	−0.10	−0.15	−0.11
N_{MW}–N_{MS}	−1.15	−0.29	+0.01	−0.30	−0.10	−0.09	−0.13
N_{HW}–N_{HS}	−1.54	−0.38	+0.01	−0.40	−0.08	−0.07	−0.20
d_W–d_S	−1.59	−0.90	−0.21	−0.71	−0.31	−0.41	−0.40

Note. O = old; H = high; W = weak; S = strong; M = medium; C = decision criterion; N = new; d = discriminability.

participant (weak minus strong). Of the 7 individual participants, 4 showed evidence of a fan effect (Participants 1, 2, 4, and 7) and 3 did not. The 4 participants exhibiting a fan all performed at least moderately better in the strong condition than the weak (i.e., the strength manipulation had the desired effect on performance). The participant whose performance was affected the most by the strength manipulation (Participant 1) also showed the most robust fan effect (by far). Of the 3 participants who did not exhibit a fan, 2 showed only slightly better performance in the strong condition than in the weak condition (Participants 3 and 5), so their data were not particularly helpful in distinguishing between the various models. The third participant's performance (Participant 6) showed a moderate strength effect, but the shifts in the confidence criteria showed no clear pattern. Thus, although some variability is apparent, on balance, these data support the predictions of the likelihood ratio model. However, for the effect to be clearly observed, it seems that strength must be substantially manipulated so that the decision criterion and the remaining confidence criteria shift enough to detect it.

This issue accounts for why the remaining studies reported by Ratcliff et al. (1994) are not analyzed here. Experiments 4 and 5 both involved relatively small manipulations of strength and little or no shift in the location of the decision criterion as a function of strength. This can be seen by examining the false alarm rates in the pure weak and pure strong conditions of those experiments. In Experiment 4, for example, they differed by less than one percent. In Experiment 5, they differed by only about 1.6%, and in Experiment 6 they differed by less than 3%. This contrasts with the results from the first three experiments, which showed differences in the false alarm rates of 5% or more.

One possible concern about the experiments we did choose to analyze (i.e., Experiments 1 through 3) is that participants were given fairly detailed instructions on how they should use the 6-point confidence scale. More specifically, in these experiments participants were instructed to distribute their responses evenly over all of the confidence categories. In addition, in Experiment 3, participants were given feedback at the end of each list concerning the number of responses they made in each confidence category. These instructions were reasonable in the context of that experiment because they were designed to ensure a more complete description of the ROC. However, for the question we are asking (namely, How do the confidence criteria shift as a function of strength?), these instructions may not have been optimal. That is, conceivably, the significant deviations from the lockstep model resulted from participants' attempts to comply with experimenter instructions concerning the use of the confidence scale, not because of their inherent decision making processes. Although it is easier for us to imagine how such instructions might (if anything) attenuate a fan effect, the question is an empirical one: Would the fan effect be observed if participants were permitted to use the confidence scale as they see fit? We now report the results of a study designed to answer that question. In this experiment, participants were tested for two sessions each. One session involved four strong lists (with each item presented three times), and another session involved four weak lists (with each item presented only once). Confidence-based ROC analyses were performed separately for each participant.

Method

Participants. The participants were 18 undergraduates of the University of California, San Diego, who were enrolled in a lower division psychology course. Participation in the experiment satisfied a course requirement.

Materials and design. The word pool consisted of 576 words drawn from Nelson, McEvoy, and Schreiber (1994), of which 288 were LF (0 to 3 occurrences per million) and 288 were HF (40 occurrences or more per million). A nonoverlapping pool of mixed-frequency words was compiled from the Kučera and Francis (1967) norms for use in a distractor task.

Procedure. Participants were tested individually and were presented with a list of words in one of two encoding conditions that occurred separately in two sessions 1 week apart. A random half of the participants received the weak encoding condition first (Session 1) and the strong encoding condition second (Session 2), and the other half received them in the reverse order. Participants studied four 36-item lists in each session, and each list was followed by a yes–no recognition test. The lists were constructed by randomly drawing words from the word pool without replacement. In the weak encoding condition, list items were presented one at a time for 500 ms each, with an interstimulus interval (ISI) of 250 ms. The strong encoding condition was the same except that the target words were presented three times each, with each presentation occurring randomly throughout the list. Participants were instructed to read each word aloud as it appeared on the screen.

After a list was presented, participants were given a 20-s distractor task in which words from the distractor pool were presented at the center of the screen, one at a time for 500 ms with an ISI of 500 ms, spelled backwards. Participants were instructed to read the words as they would be pronounced if spelled in the correct, forward order.

Immediately after the distractor task, participants were given a yes–no recognition test. The recognition test consisted of the 36 targets randomly intermixed with 36 additional words randomly drawn from the word pool (without replacement). Participants were informed that they would be asked to make a remember–know judgment for each yes response (although these were not relevant to the present analysis) and to make a confidence rating on a 1 to 5 scale (*complete guess* to *absolutely certain*) for each response. For consistency with the preceding analyses, we label the criteria

associated with a high-confident yes response O_H (this corresponds to a yes-5), a medium-confident yes response O_M (this corresponds to a yes-3), and high- and medium-confident no responses N_H and N_M, respectively.

Participants were encouraged to respond quickly to keep them engaged in the task. If they took longer than their mean reaction time on a preliminary test (plus 600 ms) to make their "yes" or "no" response, the computer beeped and a message was displayed requesting a faster decision on the next trial. Because participants were encouraged to respond quickly, if they thought they had mistakenly clicked the wrong mouse button, they were allowed to change their initial yes–no response before giving confidence ratings.

Results and Discussion

As expected, a typical mirror effect was observed. The hit rate in the weak condition was less than that of the strong condition (0.62 and 0.81, respectively), whereas the false alarm rate in the weak condition was considerably greater than that of the strong condition (0.28 and 0.12, respectively). ROC analyses were performed as described earlier, but each participant's data were analyzed separately. That is, for each participant, the maximum likelihood estimation procedure yielded an estimate of d, r, and three criteria (O_H, C, and N_H). A few participants never expressed high confidence in a response. For them, the next closest criterion (e.g., N_L or O_M) was estimated.

The strength manipulation had the expected effect on d (0.93 for the weak condition and 2.09 for the strong condition). Fifteen of the 18 participants produced values of r that were similar in the weak and strong conditions, although for most participants the value of r in the weak condition was slightly greater than that in the strong condition (mean r values of 0.79 and 0.70 in the weak and strong conditions, respectively, for those 15 participants). For 3 participants, however, the r values in the two conditions differed by more than 2 to 1. One of these participants exhibited a dramatic lockstep pattern, 1 showed no clear pattern, and 1 showed a reverse fan (the pattern predicted by the range model). Thus, both in terms of the r parameter and the criteria-shift pattern, these 3 participants appeared to differ from the remaining 15. The average parameter estimates for O_H, C, and N_H obtained from the individual participant fits, excluding the 3 participants with very different r values, are given below. As in the previous analyses, it is clear that the criteria do not move in lockstep, nor do they converge. Instead, as predicted by the likelihood ratio account, the criteria diverge. O_H changes by -0.02 standard deviation units (from strong to weak) (from 2.47 to 2.45), the old–new decision criterion moves -0.48 standard deviation units (from 1.09 to 0.61), and N_H shifts -1.15 standard deviation units (from -0.63 to -1.78). A repeated measures ANOVA performed on the individual-participant parameter estimates revealed a main effect of strength, $F(1, 14) = 42.7$, $MSE = 0.157$, $p < .05$; a main effect of level of confidence (i.e., O_H vs. C vs. N_H), $F(2, 28) = 79.7$, $MSE = 1.27$, $p < .05$; and (most important for present purposes), a significant interaction between those two factors, $F(2, 28) = 14.4$, $MSE = 0.167$, $p < .05$. Including the three outliers does not alter the fact that a fan is observed or the fact that the interaction is highly significant. All that changes is the appearance of the fan: O_H, instead of remaining more or less fixed, shifts by -0.29 standard deviation units, the criterion by -0.64 units, and N_H by -1.21 units.

General Discussion

The present analyses were designed to measure how confidence criteria that are theoretically situated along the signal–detection decision axis shift between conditions that produce different levels of recognition accuracy. The results suggest that these criteria fan out on the evidence axis when strength is manipulated (strong to weak), a finding uniquely predicted by the likelihood ratio model. However, this result should not be taken to imply that participants precisely maintain constant likelihood ratios for each confidence criterion across all conditions. In general, although the criteria do fan out in the weak condition as the likelihood ratio model predicts, they do not fan out as much as they should. This is most easily seen by examining the parameter estimates associated with the data taken from the weak condition in Experiment 1 of Ratcliff et al. (1994). According to Equations 1 and 2, as d' approaches 0, the criteria should fan out infinitely far. Although d' was indeed near 0 in the weak condition of that experiment, the criteria did not fan out to an extreme degree. Thus, the results reported here are qualitatively (but not quantitatively) consistent with the predictions of the likelihood ratio account.

The conclusion that the criteria fan out on the decision axis as d' decreases depends on the assumption that the parameters of the lure distributions were the same in both conditions. Because the lures in the strong condition were physically identical to the lures in the weak condition (i.e., in both cases they consisted of new words randomly drawn from the word pool), this assumption seems reasonable. However, because this assumption is critical, we turn now to a detailed discussion of that issue.

Strong and Weak Lure Distributions

Although the lures in the strong and weak conditions were equivalent in every physical respect (on average), it is theoretically possible that their psychological strength characteristics changed depending on the strength of the targets. If so, the interpretation of the results presented earlier would need to be modified. Empirical evidence bearing on this assumption was recently provided by Shiffrin et al. (1995). In that experiment, participants studied a single long list of words composed of items drawn from many different semantic and orthographic categories (21 categories in all). The various categories differed in length (i.e., number of exemplars drawn from a category) and strength (number of exemplar presentations). The unprovable but seemingly reasonable assumption was that with so many different categories, participants would adopt a single decision criterion and use it throughout (rather than adopting a different decision criterion for each of the 21 categories). Given that assumption, any change in the category-specific false alarm

rate as a function of category strength or category length could be attributed to changes in the properties of the lure distribution.

The results of this experiment were consistent with the predictions of SAM and with a central assumption made here. That is, although the hit rate increased with category strength (obviously), the false alarm rate was unaffected by that manipulation (a finding also reported by Stretch & Wixted, 1998). Note that if category strength affected either the mean or the variance of the lure distribution, the false alarm rate should have changed (assuming a fixed decision criterion). Because the false alarm rate did not change as a function of strength, the results suggest that the familiarity characteristics of the lures remain constant over variations in strength. If so, and if the same is true here, then the fan effect observed for the strength manipulation is a real effect rather than an illusion created by the effect of target strength on the standard deviation of the lure distribution.

Prior Research on Strength and Confidence

Only one previous study has directly addressed the issue under investigation here. Balakrishnan and Ratcliff (1996) recently found that for a strength manipulation the lockstep model (which they termed the "distance from criterion" model) provided a better account of their data than the likelihood ratio model. Their method of analysis differed considerably from the one used here. Instead of assuming underlying Gaussian distributions and estimating the locations of the various confidence criteria in the strong and weak conditions as we did, they compared cumulative frequency distributions for 10 levels of confidence. If participants shift their criteria in lockstep as strength is manipulated, and if the variances of the strong and weak lure distributions are the same, then the cumulative frequency distribution should simply shift to the left or right, depending on which way the criteria are moved. This prediction holds true even if the underlying strength distributions are not Gaussian in form. If participants instead shift their criteria in such a way as to maintain constant likelihood ratios (i.e., if participants behave as ideal observers), then the two cumulative frequency distributions should intersect. The data reported by Balakrishnan and Ratcliff were visually more in line with the predictions of the lockstep model.

Balakrishnan and Ratcliff (1996) compared the predictions of the lockstep model versus an idealized likelihood ratio model, whereas we tested for statistically significant deviations (however small) from the lockstep pattern. Our findings suggest that although the deviations from the lockstep model may not be as extreme as they should be, they are significant and are in the direction predicted by the likelihood ratio account. Thus, Balakrishnan and Ratcliff's results showing that the data do not convincingly rule out the lockstep model are not entirely inconsistent with the results presented here. The lockstep model may provide a reasonable approximation of the data, but if one is willing to assume underlying Gaussian or logistic distributions (thereby adding power to the analysis), the deviations from that model are statistically significant and are in the direction predicted by the likelihood ratio account (which is not to say that participants were behaving as ideal observers).

Unequal Variances and the Likelihood Ratio Decision Rule

The predictions of the likelihood ratio model shown in the lower panel of Figure 3 and represented by Equations 1 and 2 were derived from the assumption of equal target and lure distribution variances. When those variances are not equal and the distributions are Gaussian in form, responding on the basis of a likelihood ratio seems peculiar because no unique solution exists. For example, a likelihood ratio of 1 occurs at the point of intersection between the two distributions (because that is the point where the height of the target distribution equals the height of the lure distribution). For the equal variance case, the distributions intersect at only a single point. Thus, setting the criterion at the point where the likelihood ratio is 1 poses no dilemma. For the unequal variance case, the distributions intersect in two places. As a result, participants who choose to set their decision criterion on the familiarity axis at the point where the likelihood ratio is 1 would have two options. As described in more detail in the Appendix, similar difficulties arise for ratios other than 1. Thus, given Gaussian distributions with unequal variances, the predictions of the likelihood ratio model are ambiguous. Nevertheless, in spite of the fact that the target and lure distributions have unequal variances (given that r is usually less than 1), participants appear to shift their confidence criteria in approximate accordance with the predictions made by the equal-variance likelihood ratio model.

The fact that participants do not shift their confidence criteria to the degree that they should already suggests that they might not be engaging in an actual likelihood ratio computation. If not, then the unequal target and lure distribution variances may not pose a dilemma. For example, perhaps participants adjust their criteria in the manner that they do merely because that strategy has proven to be effective in everyday life. If one is at all sensitive to feedback, one should learn that when memory conditions are unfavorable, more extreme criteria settings should be used before expressing high confidence in a recognition decision. Failure to follow that strategy (as one may have learned through painful experience) too often results in high-confident responses being wrong. According to this idea, experience with the likelihood of being correct under various memory conditions, not a likelihood ratio computation, accounts for the pattern of results reported here. However, experience might not be so effective a teacher as to train participants to behave as ideal observers, and that might explain why the criteria do not fan out as far as they should.

On the other hand, if participants do engage in a likelihood ratio computation after all, there may be other ways around the unequal variance problem. As indicated earlier, one possible explanation is to assume that the equal variance model is correct in spite of the fact that the slope of the ROC (r) is less than 1. Yonelinas (1994), for example,

suggested that the target and lure distribution variances may indeed be equal and that r deviates from 1 because some recognition responses are based on all-or-none retrieval processes. If so, then participants might be able to place their confidence criteria along the familiarity axis according to an equal-variance likelihood ratio model after all. Of course, even this idea would not explain why participants do not shift their criteria in exact accordance with the predictions of a likelihood ratio model. Thus, if one assumes that participants are responding on the basis of an equal variance computation, one must further assume that some inaccuracies creep into the computations. Thus, for example, when d' is close to zero, perhaps participants mistakenly believe that it is significantly greater than zero (thereby explaining why the criteria do not fan out as much as they should).

Another way to avoid the dilemma of an unequal-variance likelihood ratio without introducing all-or-none retrieval processes is to assume that the underlying distributions are not Gaussian in form after all. Indeed, one unusual implication of the unequal-variance Gaussian model is that items with extremely low levels of familiarity are more likely to be targets than lures. Intuition suggests that it should be the other way around. Although some mechanism may be able to account for that curious possibility, it seems simpler to assume that the target distribution is at all points (including in the left tail) situated to the right of the lure distribution. A number of other distributions, such as the binomial distributions assumed by Glanzer et al. (1993) and the positively skewed distributions generated by a new theory of memory proposed by Shiffrin and Steyvers (1997), have this property. These distributions are bounded at zero (which in itself makes more sense than a Gaussian distribution that extends to minus infinity) and can describe distributions with unequal variances that nevertheless exhibit a continuously increasing likelihood ratio function as strength of evidence increases (which is another way of saying that the distributions intersect only once).

Unfortunately, if the underlying distributions are not Gaussian (or Gaussian-like), then the estimates of the locations of the confidence criteria presented earlier may not be valid. What can be said of the present findings under those conditions? One pattern sometimes observed in the relevant data still suggests a fan even if one makes much weaker assumptions about the shape of the underlying distributions. An example of this pattern was observed in the data from Experiment 1 of Ratcliff et al. (1994). In that experiment, the overall false alarm rate increased from strong to weak (from 0.30 to 0.35, as shown in Table 1), but the high-confident false alarm rate actually decreased slightly (from 0.07 to 0.06). The same pattern was also observed in the data of Participants 1, 2, and 7 from Experiment 3 of Ratcliff et al. (1994). Regardless of the shape of the underlying lure distribution, this could occur only if C shifted to the left (thereby increasing the overall false alarm rate) and O_H shifted to the right (thereby decreasing the high-confident false alarm rate). When this pattern is not observed, though (as in their Experiment 2), assumptions about the shape of the underlying distributions are critical.

Glanzer's Attention Likelihood Theory

The analyses discussed to this point have all assumed a strength-of-evidence decision axis because that is a common and intuitively appealing assumption. Nevertheless, it is possible that the decision axis does not represent strength of evidence but instead represents a likelihood ratio scale. According to this idea, participants compute likelihood ratios from a strength variable as described earlier, but now the likelihood ratio itself is the operative psychological variable. Thus, the target and lure distributions no longer represent distributions of familiarity values but instead represent distributions of log likelihood ratios. Similarly, a single point on the decision axis no longer represents a familiarity value but instead represents a specific likelihood ratio. Glanzer's attention likelihood theory (ALT) explicitly assumes a likelihood ratio decision axis (Glanzer et al., 1993) and in doing so explicitly endorses the view that participants know the forms of the relevant strength distributions (binomial in their model) and can compute likelihood ratios from them. A new model of recognition memory proposed by Shiffrin and Steyvers (1997) also assumes a likelihood ratio decision axis.

If the decision axis represents a strength-of-evidence variable (such as familiarity), then to maintain constant likelihood ratios across conditions the confidence criteria must fan out as indicated in the lower panel of Figure 3. If the decision axis represents a log likelihood ratio scale, by contrast, then to maintain constant likelihood ratios across conditions the criteria must remain fixed (movement would imply a changing likelihood ratio). Furthermore, the variances of the target and lure distributions must both decrease as conditions change from strong to weak. Why, then, do the criteria appear to fan out in the data reported here? Because, according to ALT, the lure distribution in the weak condition is less variable than the corresponding strong distribution (i.e., $\sigma_{Weak} < \sigma_{Strong}$). If so, then a central assumption of the preceding analysis is violated, and an apparent fan effect would be observed. As indicated earlier, the location of the confidence criteria are measured in units equal to the standard deviation of the lure distribution. If that value were greater in the strong condition than in the weak condition, then the confidence criteria in the strong condition would appear to be less distant from the mean of the strong lure distribution than the corresponding confidence criteria in the weak condition. Hence, the illusion of a fan effect. Thus, the findings reported here in favor of a fan effect are equally consistent with the idea that the criteria remain fixed across conditions and that $\sigma_{Weak} < \sigma_{Strong}$ (which is what a likelihood ratio decision-axis model would require).

The results of the study performed by Shiffrin et al. (1995), which showed that the false alarm rates for strong and weak categories were the same for a within-list strength manipulation, appear to suggest that the characteristics of the lure distribution do not vary as a function of the strength of the targets. That is, assuming a fixed decision criterion, the equivalent false alarm rates across conditions suggest that $\sigma_{Weak} = \sigma_{Strong}$. However, because so many categories were used, participants presumably did not know whether

the category from which a particular lure was drawn was strong or weak. Under these conditions, they could not compute strength-specific likelihood ratios even if they wanted to. Instead, they would be forced to generate a generic likelihood ratio (based, perhaps, on a theoretical target distribution the mean of which was the average of the various strong and weak conditions). Under those conditions, only one lure distribution would exist (rather than a strong lure distribution and a weak lure distribution), and the false alarm rate would not be expected to change as a function of category strength. Thus, even Glanzer's model predicts that for the procedure used by Shiffrin et al. (1995) the characteristics of the lure distribution should not change as a function of strength. In a pure-strength manipulation (like those examined here), by contrast, participants would be able to form strength-specific likelihood ratios for the lures. Under those conditions, the characteristics of the lure distribution would change, with the variance of the strong lure distribution being greater than that of the weak lure distribution (and an apparent fan effect would be observed).

On the other hand, Stretch and Wixted (1998) recently reported a study like that reported by Shiffrin et al. (1995), except that only two perceptual categories were used. In this study, strong items (which received extra study) were presented in one color, such as red, and weak items in another color, such as green. On the subsequent recognition test, the strong–red and weak–green targets were randomly intermixed with an equal number of red and green lures for a yes–no recognition test. Under these conditions participants should be able to easily form strength-specific likelihood ratios just as they might do for a between-list strength manipulation. As with Shiffrin et al. (1995), however, the false alarm rates to red and green lures were nearly identical. One simple explanation for this result is that the decision axis does indeed represent a strength-of-evidence variable (like familiarity) and that participants prefer to use a single decision criterion throughout the course of a recognition test (rather than alternating between different settings, depending on whether the test item is red or green). This finding is less easily explained by a likelihood ratio model. If participants ordinarily compute likelihood ratios for each item anyway, why would they ignore color information that accurately indicates whether the corresponding target distribution is strong or weak?

In any case, the results of the between-list strength analysis reported here are consistent with the predictions of Glanzer's theory even if the results of the within-list strength analysis reported by Stretch and Wixted (1998) are not. Interpreted in terms of ALT, the findings reported here suggest that for a strength manipulation the variance of the lure distribution in the weaker condition is less than the variance of the lure distribution in the stronger condition (assuming one is prepared to accept the possibility that the decision axis represents a log likelihood ratio scale). This is basically the same finding that Glanzer and Adams (1990) reported with respect to a word frequency manipulation.

Conclusion

The main conclusion to be drawn from the present research is the following: If the decision axis is assumed to represent a strength-of-evidence variable such as familiarity, then for strength manipulations the confidence criteria fan out as recognition accuracy decreases (rather than converging or shifting in lockstep). This is qualitatively what a likelihood ratio model predicts, but the criteria did not fan out as much as they should have. If one instead assumes a likelihood ratio decision axis, as Glanzer does, then the results presented here suggest that the variance of the weak lure distribution is less than that of the strong lure distribution (in accordance with the predictions of ALT).

References

Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* New York: Dover.

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 615–633.

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18,* 23–30.

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

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and 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. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567.

Greene, R. L., & Thapar, A. (1994). Mirror effect in frequency discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 946–952.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.

Hirshman, E., & Arndt, J. (1997). Discriminating alternative conceptions of false recognition: The cases of word concreteness and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1306–1323.

Kintsch, W. (1967). Memory and decision aspects of recognition learning. *Psychological Review, 74,* 496–504.

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

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74,* 100–109.

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1994). *The University of South Florida word association, rhyme and word fragment norms.* Unpublished manuscript.

Ogilvie, J. C., & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology, 5,* 377–391.

Parducci, A. (1984). *Perspectives in psychological experimentation: Toward the year 2000.* Hillsdale, NJ: Erlbaum.

Ratcliff, R., McKoon, G., & Tindall, M. (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, & Cognition, 20,* 763–785.

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

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 267–287.

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

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24,* 1379–1396.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 681–690.

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

Appendix

Computing Likelihood Ratios for the Unequal-Variance Gaussian Model

The equation for the signal distribution with mean d' and standard deviation s is given by

$$SN(d', s) = 1/\sqrt{2\pi s^2}e^{-.5(f-d')^2/s^2}, \quad \text{(A1)}$$

and the equation for the lure distribution with mean 0 and standard deviation 1 is given by

$$N(0, 1) = 1/\sqrt{2\pi}e^{-.5f^2}. \quad \text{(A2)}$$

The likelihood ratio at a particular point f on the familiarity axis is given by the ratio of these two equations:

$$L = \frac{1/\sqrt{2\pi s^2}e^{-.5(f-d')^2/s^2}}{1/\sqrt{2\pi}e^{-.5f^2}}, \quad \text{(A3)}$$

which reduces to

$$L = re^{-0.5r^2(f-d')^2}e^{0.5f^2},$$

or

$$L = re^{-0.5r^2(f-d')^2+0.5f^2}, \quad \text{(A4)}$$

where $r = 1/s$. Taking the log of both sides yields

$$\log(L) = \log(r) - 0.5r^2(f - d')^2 + 0.5f^2$$

or, after squaring,

$$\log(L) = \log(r) - 0.5r^2(f^2 - 2fd' + d'^2) + 0.5f^2. \quad \text{(A5)}$$

By subtracting $\log(L)$ from both sides of the equation, Equation A5 can be rearranged in the form of a quadratic:

$$af^2 + bf + c = 0,$$

where $a = -0.5(r^2 - 1)$, $b = d'r^2$, and $c = \log(r) - 0.5r^2d'^2 - \log(L)$.

To solve for a value of f that satisfies the particular likelihood ratio, L, the standard quadratic formula may be used:

$$f = (-b \pm \sqrt{b^2 - 4ac})/2a.$$

When the foregoing values for a, b, and c are substituted into this equation, it eventually reduces to this rather forbidding-looking expression:

$$f = [d'r^2 \pm \sqrt{(d'r^2)^2 - (r^2 - 1)[(d'r)^2 - 2\log(r/L)]}]/(r^2 - 1). \quad \text{(A6)}$$

To compute where on the familiarity axis a confidence criterion should be placed to maintain a given likelihood ratio, one need only substitute the relevant values and solve the quadratic. For example, assume that in the strong condition, $d' = 2.5$, $r = .80$, and $O_H = 2.0$ (which means that the familiarity value, f, corresponding to the high-confidence "old" criterion is at 2.0 on the familiarity axis). According to Equation A4, the likelihood ratio associated with this confidence criterion is 5.46, which is to say that any item that is 5.46 or more times as likely to have come from the target distribution than the lure distribution receives a "yes" response with high confidence. Now assume that in the weak condition $d' = 1.0$ and $r = 0.90$. Where should O_H be placed in order to keep the same likelihood ratio of 5.46? To find the answer, simply substitute these values for d', r, and L in Equation A6. In this example, two possible values of f are returned: 2.17 and −10.70. Thus, in the weaker condition (with $d' = 1$), the high-confident "old" criterion could be placed at either location in order to maintain a constant likelihood ratio of 5.46. The fact that two solutions exist illustrates one of the peculiarities that arises for an unequal variance

(*Appendix continues on next page*)

likelihood ratio model. Usually, however, one of the solutions is an extreme one, so one could perhaps assume that participants use the less extreme solution (in this case, 2.17).

To illustrate a second peculiarity that can arise, assume that $N_H = -1.0$ in the same strong condition. According to Equation A4, this criterion is associated with a likelihood ratio of 0.074. Where should N_H be placed to maintain that likelihood ratio in the weak condition? It turns out that there is no setting on the familiarity axis that will yield this likelihood ratio in the weak condition. The minimum possible likelihood ratio that can be obtained is 0.107, which occurs when the criterion is placed at −4.26. One could perhaps assume that when faced with a situation like this, participants place the criterion at the point that comes closest to maintaining a constant likelihood ratio (in this case, at −4.26).

With these caveats in mind, exact predictions for the preceding experiments can be generated. Table A1 provides an example from Experiment 1 of Ratcliff et al. (1994). All of the predictions were generated using actual d' values (the distance between the means of the target and lure distributions in units equal to the standard deviation of the lure distribution). This differs from the d values shown in Table 1. The transformation is given by $d' = d(1 + r)/2r$. The likelihood ratios associated with the five criteria in the strong condition were first computed using Equation A4, and then predictions about where the various criteria should be placed in the weak condition in order to maintain constant likelihood ratios were generated using Equation A6. Table A1 presents the results and shows that as conditions change from strong to weak, the criteria in the weak condition should fan out. What is also clear is that in order to maintain constant likelihood ratios, they should fan out considerably more than they actully do.

Table A1
Exact Likelihood Ratio for Experiment 1 of Ratcliff et al. (1994)

Confidence criterion	Strong	Likelihood ratio	Weak predicted	Weak actual
O_H	1.52	2.00	3.87	1.64
O_M	0.94	1.25	1.96	1.07
C	0.52	0.95	−0.28	1.03
N_M	0.00	0.73	−1.22[a]	0.95
N_H	−0.91	0.57	−1.22[a]	0.93

Note. O = old; H = high; M = medium; C = decision criterion; N = new.
[a]Indicates criterion placed at the minimum possible likelihood ratio.

Received April 28, 1997
Revision received January 26, 1998
Accepted February 4, 1998 ■