Journal of Experimental Psychology:
Learning, Memory, and Cognition
2001, Vol. 27, No. 3, 701-722

0278-7393/01/$5.00 DOI: 10.1037//0278-7393.27.3.701

On the Nature of Associative Information in Recognition Memory

Robert Kelley and John T. Wixted
University of California, San Diego

In a typical associative-recognition task, participants must distinguish between intact word pairs (both words previously studied together) and rearranged word pairs (both words previously studied but as part of different pairs). The familiarity of the individual items on this task is uninformative because all of the items were seen before, so the only way to solve the task is to rely on associative information. Prior research suggests that associative information is recall-like in nature and may therefore be an all-or-none variable. The present research reports several experiments in which some pairs were strengthened during list presentation. The resulting hit rates and false alarm rates, and an analysis of the corresponding receiver operating characteristic plots, suggest that participants rely heavily on item information when making an associative-recognition decision (to no avail) and that associative information may be best thought of as a some-or-none variable.

An intriguing issue in the study of recognition memory concerns how people recognize associations between two items beyond recognition of the items themselves. In a typical associative-recognition task, participants study a list of word pairs and are later asked to decide whether pairs presented on a recognition test are intact or rearranged. Intact word pairs consist of two items that appeared together on the list, whereas rearranged word pairs consist of two items that appeared on the list as part of different pairs. An interesting feature of this design is that all of the individual items have been seen before, so item information (e.g., item familiarity) does not help the participant to decide whether a given pair was seen before. In spite of this, and as detailed below, associative-recognition decisions may involve both item and associative information.

The Role of Item and Associative Information in Associative Recognition

Glenberg and Bradley (1979; Bradley & Glenberg, 1983) and Nairne (1983) both showed that maintenance rehearsal of a word pair tends to affect item information without changing the strength of the associative bond. In Nairne's experiment, participants received trials in which they were presented with a three-digit number to remember, followed by a distractor task in which they were asked to repeat a given word pair throughout the retention interval. The amount of maintenance rehearsal a word pair received was manipulated by varying the duration of the retention interval. At the end of the experiment, participants received a surprise associative-recognition test for the word pairs that had served as distractors. Nairne found that lengthening the duration of maintenance rehearsal increased both the hit rate ("yes" responses to intact word pairs) and the false-alarm rate ("yes" responses to rearranged word pairs) but did not affect participants' ability to distinguish intact from rearranged word pairs (i.e., d' was unaffected). Thus, although maintenance rehearsal did not affect the associative connection between word pairs, it apparently did influence item familiarity, and the more familiar the individual items were as a result of maintenance rehearsal, the more likely participants were to declare the pair as having been seen before. These results suggest that participants have faith in the utility of item information even though such information is not at all diagnostic in this procedure.

Using a continuous recognition procedure, Hockley (1991, 1992) investigated the loss of item and associative information as a function of retention interval, and he arrived at similar conclusions. In this procedure, pairs of items were presented for study in sequential fashion and, occasionally, the participant was asked whether a particular pair had been presented earlier in the session. The lag between study and test (i.e., the retention interval) was manipulated by varying the number of intervening presentations. Hockley found that as the retention interval increased, both the hit rate to intact word pairs and the false-alarm rate to rearranged word pairs decreased in tandem, but d' was unaffected. Thus, manipulating the duration of the retention interval does not appear to influence associative information very much, but it may affect the familiarity of the individual items. As the familiarity of the items that composed the intact or rearranged pair decreased with lag, participants apparently became less confident that the pair was previously encountered. Once again, these results suggest that item information plays a role in an associative-recognition procedure even though such information is not indicative of whether a test pair is intact or rearranged.

These ideas can be summed up in a signal-detection model proposed by Hockley (1992). This model states that item information (I) and associative information (A) are both continuously distributed random variables (such as familiarity), and are additive in their effects. In other words, item information and associative information sum to produce a strength-of-evidence value for a given word pair. If the combined evidence ($I + A$) exceeds a decision criterion (c), then the participant concludes that the words

Robert Kelley and John T. Wixted, Department of Psychology, University of California, San Diego.

Correspondence concerning this article should be addressed to John T. Wixted, Department of Psychology, 0109, University of California, San Diego, La Jolla, California 92093-0109. Electronic mail may be sent to jwixted@ucsd.edu.

were presented together.[1] Thus, according to this model, the hit rate (H) for an intact pair is equal to the probability that the summed evidence exceeds a criterion, or

$$H = p[(I + A) > c]. \quad (1)$$

Associative information is lacking for rearranged pairs, so the false-alarm rate is a function of item information only. Thus, the false-alarm rate (FA) for a rearranged pair is equal to the probability that item familiarity exceeds a criterion, or

$$FA = p(I > c). \quad (2)$$

According to this model, the average strength of evidence associated with an intact pair (μ_{intact}) equals average item information (μ_{item}) plus average associative information (μ_{assoc}). Thus, $\mu_{intact} = \mu_{item} + \mu_{assoc}$. By contrast, the average evidence value for a rearranged pair (μ_{rearr}) derives only from item information because no associative information was stored during list presentation (i.e., $\mu_{rearr} = \mu_{item}$). Thus, the difference between the mean evidence values for the intact and rearranged distributions (i.e., $\mu_{intact} - \mu_{rearr}$) is equal to $\mu_{item} + \mu_{assoc} - \mu_{item}$, which is simply equal to μ_{assoc}.

Hockley's (1992) model is illustrated graphically in Figure 1. The upper panel shows the intact and rearranged distributions following a short retention interval, and the lower panel shows the corresponding distributions following a long retention interval. In terms of this simple model, μ_{item} is affected by the delay since the pair was seen, but μ_{assoc} is not. Thus, if μ_{item} decreases as the retention interval increases, the strength of evidence for both intact and rearranged word pairs declines (thereby decreasing both the hit rates and false-alarm rates), but the average difference between the intact and rearranged distributions remains unchanged (i.e., $\mu_{intact} - \mu_{rearr} = \mu_{assoc}$, regardless of whether μ_{item} is large or small). As a result, d', a standardized measure of the distance between the intact and rearranged distributions, is unaffected by this manipulation.

Also shown in both panels of Figure 1 is a distribution labeled *p-lures* (i.e., pair lures). The pair-lure distributions represent hypothetical familiarity values associated with entirely new pairs, that is, pairs in which neither word was previously studied. Hockley (1992) did not actually consider how pair lures might fit in, but the extension of his model to include them seems straightforward (and is relevant to the research we describe later). The difference between the rearranged distribution and the pair-lure distribution is solely a function of item information because in neither case were the items studied together. That is, associative strength equals zero in both cases, but the average familiarity of the individual items is high in the case of rearranged pairs (because the items were seen earlier as part of a different pair) and low for pair lures (because the items have not been encountered previously in the experiment setting). Although neither maintenance rehearsal nor retention interval affects the distance between the intact and rearranged distributions, such manipulations would affect the distance between these two distributions and the pair-lure distribution. As illustrated in Figure 1, the mean of the pair-lure distribution is unaffected by the size of the retention interval, because these items were not presented earlier in the experiment.

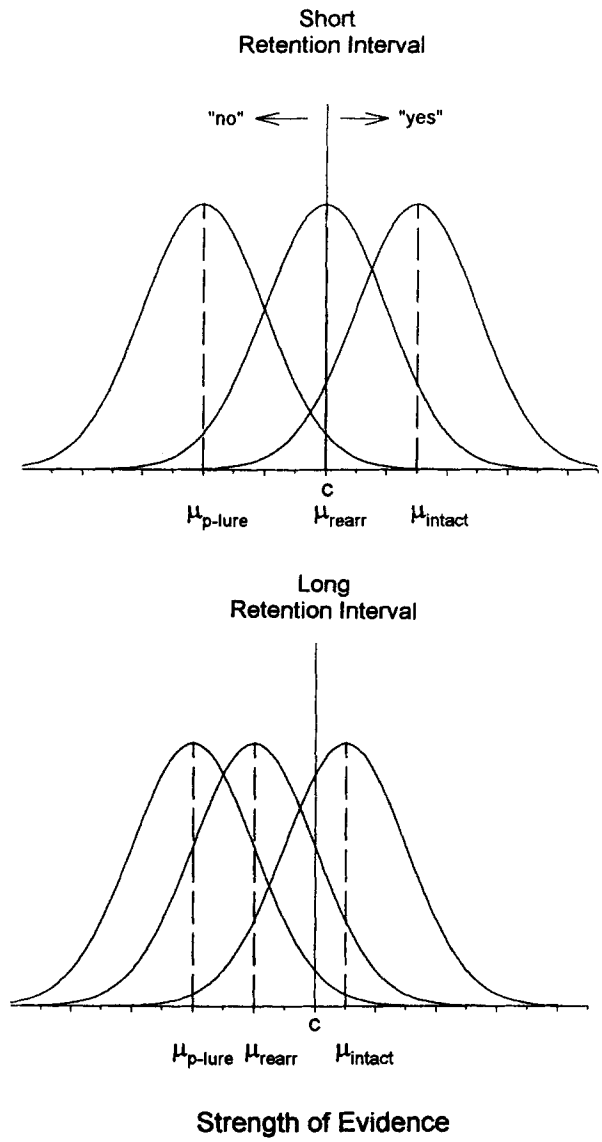

Figure 1. Hypothetical strength-of-evidence distributions for pair lures, rearranged pairs, and intact pairs following a short retention interval (upper panel) and long retention interval (lower panel).

The Role of Retrieval in Associative Recognition

Although the model illustrated in Figure 1 does a nice job of accounting for some findings in the associative-recognition literature, it does not easily accommodate other findings that suggest a role for retrieval processes in associative recognition. In accordance with this idea, one member of a word pair can aid in the decision process by serving as a retrieval cue for the other member of the pair (Humphreys, 1978; Mandler, 1980). Successful retrieval of this kind enables participants to confidently accept intact

[1] As discussed by Hockley (1992), one could instead assume that the evidence axis represents associative information only and that the change in hit rates and false-alarm rates reflects a criterion shift occasioned by the changing levels of item familiarity. Adopting this point of view would not change the essence of our arguments in any way, so we rely on the simpler account illustrated in Figure 1. Hockley (1992, Footnote 1) pointed out that the version of his model that we have adopted was actually suggested by Douglas Hintzman.

word pairs and to confidently reject rearranged word pairs. Thus, unlike the model illustrated in Figure 1, retrieval models assume that associative information plays a role for rearranged pairs as well as for intact pairs.

Experiments by Clark, Hori, and Callan (1993) and Clark and Hori (1995) provided compelling evidence that a recall-like mechanism is involved in associative-recognition decisions. Participants in these experiments studied a list of word pairs and then completed a forced-choice associative-recognition test. Each test trial on the recognition test involved an intact word pair (the correct choice) and two rearranged word pairs. The experimental manipulation of interest involved the degree to which the three pairs overlapped. In the overlap condition, one of the items was common to all three pairs (e.g., study AB, CD, EF; test AB–AD–AF). In the nonoverlap condition, individual items were unique to each pair (e.g., study AB, CD, EF; test AB–CF–DE).

If judgments are based entirely on the sum of item and associative familiarity (as the model depicted in Figure 1 assumes), then performance in the overlap condition should benefit from the correlation in familiarity between the three word pairs. This prediction can be derived in a formal way using familiarity-based models like the search of associative memory model (SAM; Gillund & Shiffrin, 1984) and the theory of distributed associative memory (TODAM; Murdock, 1982), but is perhaps most easily understood by imagining how easy the task would be if all of the individual items (A, B, C, D, E, and F) were always exactly equal in familiarity. In that case, the sum of item and associative information ($I + A$) would always be greatest for the intact pair (AB), and the correct answer could be given every time. If the familiarity values for all of the items were completely uncorrelated, however, I for a rearranged pair might sometimes exceed $I + A$ for an intact pair (in which case the wrong pair would be selected). The nonoverlap condition corresponds to the completely uncorrelated case, but the overlap condition is, by design, closer to the completely correlated case, and performance should be better as a result. Contrary to that prediction, performance was better in the nonoverlap condition. This result suggests that the greater number of recall cues available in the nonoverlap condition (which can be used to reject rearranged pairs) outweighs the disadvantage of higher variability in word-pair familiarity.

Clark and Hori (1995) noted that TODAM is the least challenged by these results because that model can at least predict equivalent performance in the overlap and nonoverlap conditions (rather than predicting an effect in the wrong direction) so long as one assumes that participants ignore item information. They noted, however, that "whether subjects actually can ignore this irrelevant information is not known" (p. 460). Evidence bearing on this particular issue is presented in the first three experiments reported in this article.

Additional compelling evidence suggesting that associative-recognition decisions are retrieval based (rather than familiarity based) was provided by an analysis of receiver operating characteristic (ROC) plots reported by Yonelinas (1997). If associative-recognition responses are based on a continuous strength-of-evidence variable (such as familiarity), then the shape of the ROC should be curvilinear, as ROC plots almost always are for item recognition. If, on the other hand, associative-recognition judgments are based on an all-or-none recall process, then the shape of the ROC should be linear. The data reported by Yonelinas (1997) were very nearly linear, thus supporting the all-or-none retrieval model (and contradicting the continuous strength-of-evidence model shown in Figure 1). Indeed, Yonelinas (1997) suggested that familiarity may not contribute at all to associative-recognition performance, and he proposed a retrieval-based, high-threshold account of associative-recognition performance as an alternative. According to this account, if participants recall that the two items of an intact pair were presented together (which occurs with probability Ro), then they respond "yes" with high confidence. If they do not recall that the two items of an intact pair were presented together (which occurs with probability $1 - Ro$), then they sometimes guess "yes" (with probability g) with varying degrees of confidence. Thus, the hit rate (H) is given by

$$H = Ro + (1 - Ro)g \quad (3)$$

(Ro represents the probability of successful retrieval given an "old" pair). Similarly, if participants recall that one of the items of a rearranged pair was presented as part of a different pair (which occurs with probability Rn), then they say "no" with high confidence. If they do not recall what either of the two items were originally paired with (which occurs with probability $1 - Rn$), then they sometimes guess "yes," with varying degrees of confidence (again with probability g). Thus,

$$FA = (1 - Rn)g \quad (4)$$

(Rn represents the probability of successful retrieval given a "new" pair). Solving Equation 4 for g and substituting the result into Equation 3 yields

$$H = Ro + [(1 - Rn)/(1 - Ro)]FA. \quad (5)$$

This equation, which is of the form $y = a + bx$, implies that the relationship between the hit rate and the false-alarm rate (i.e., the ROC plot) will be linear, and that is exactly what Yonelinas (1997) and, more recently, Rotello, Macmillan, and Van Tassel (2000) found.

Whereas prior research suggested that item and associative familiarity may combine to jointly influence associative-recognition decisions, the more recent research reviewed above suggests that associative retrieval may actually play the dominant role (cf. Westerman, 2001) and that associative information may be an all-or-none variable. The four experiments described next were designed to further investigate the role of item familiarity in associative-recognition decisions and to shed additional light on the nature and role of associative information in this task.

Experiment 1

The first three experiments are variations on the same simple theme. In Experiment 1, which is the least complicated of the three, participants studied a list of word pairs, some of which were presented once (the weak pairs), and some of which were repeated several times (the strong pairs). On the subsequent recognition test, participants were asked to distinguish intact pairs from rearranged pairs (and, in Experiments 2 and 3, from pair lures as well). All of the relevant models and simple intuition suggest that participants will exhibit a higher hit rate to strong intact word pairs relative to weak intact word pairs. That is, according to the familiarity-based model described earlier, repeating pairs should increase the familiarity values of the individual items (I) as well as the strength of their associative bond (A). Thus, according to Equation 1, the hit

rate for strong pairs should exceed that of weak pairs. The retrieval-based high-threshold model makes the same prediction. According to that model, repeating pairs should increase the probability that the intact pair will be retrieved (*Ro*). As a result, according to Equation 3, the hit rate should increase.

A question of interest in this experiment is how participants will respond to strong versus weak rearranged word pairs. A strong rearranged word pair consists of two items that did not appear together but instead appeared as part of two different strong word pairs. Thus, *I* (the familiarity value associated with the individual items) should be high, and, according to a familiarity-based model, the false-alarm rate to a strong rearranged pair should be correspondingly high (Equation 2). According to the retrieval-based model, by contrast, participants should be less likely to false alarm to strong rearranged word pairs. That is, because the original word pair was presented several times, each item of a rearranged pair should serve as a more effective retrieval cue for the other member of the original pair (i.e., *Rn* would increase). Thus, according to Equation 4, the false-alarm rate should be low in the strong condition.

In addition to testing these specific and contrasting predictions of the familiarity-based and retrieval-based accounts, Experiment 1 and the subsequent experiments also offer a rich array of ROC analyses that may shed additional light on the nature of associative information. The familiarity-based signal-detection model regards such information as continuous in nature, which leads to the prediction of a curvilinear ROC. The reason why a curvilinear ROC is predicted by this model is not easy to demonstrate mathematically (because it involves an unsolvable integral of the Gaussian distribution). However, the fact that a nonlinear ROC is predicted is easily appreciated by considering how the hit rate and false-alarm rate should change as the decision criterion moves from the far right to the far left in Figure 1. When the criterion is extremely far to the right, the participant will respond "no" to virtually every item, so both the hit rate and false-alarm rate will equal zero. This yields one point on the ROC, and it falls on the main diagonal in the lower left corner. When the criterion is placed between the two distributions, as shown in Figure 1, the hit rate will exceed the false-alarm rate, which yields the second point on the ROC (and it falls above the main diagonal). Finally, when the criterion is extremely far to the left, the participant will respond "yes" to virtually every item, so both the hit rate and the false-alarm rate will approach 1. This point again falls on the main diagonal, but at the upper right corner of the ROC. Thus, as the criterion sweeps from right to left, the points on the ROC trace out a path that begins and ends on the main diagonal (beginning at the lower left corner and ending at the upper right corner) but is some distance away from the main diagonal at the halfway mark (when the criterion is in the middle).

By contrast, and as indicated above, the retrieval-based high-threshold model considers associative information to be an all-or-none rather than a continuous variable, and, as shown by Equation 5, this model predicts a strictly linear ROC. A perennially confusing issue in ROC analysis is that the predictions of the two models differ depending on the details of how the ROC is constructed. Signal-detection theory predicts a curvilinear function, and high-threshold theory predicts a linear function when the ROC is constructed by plotting the hit rate versus the false-alarm rate. A common practice is to instead construct the ROC by plotting the *z*-transformed hit rate versus the *z*-transformed false-alarm rate (Macmillan & Creelman, 1991). When that is done, the signal-detection model now predicts a linear ROC, and the high-threshold model predicts a curvilinear ROC (the exact opposite of their predictions with respect to the untransformed ROC). Although there are good reasons for plotting the ROC in *z*-transformed coordinates (e.g., deviations from the predictions of signal-detection theory are easier to see), all of the ROCs reported here are presented in untransformed coordinates. We adopted that method because the most appropriate model fitting technique (discussed later) operates on untransformed data and because the use of untransformed hit rates and false-alarm rates makes it easy to understand what the numbers on the two axes of the ROC represent.

Method

Participants. The participants were 36 undergraduates from the University of California, San Diego, who met the criterion of English fluency by age 7. Their participation fulfilled a lower division psychology course requirement.

Materials. The word pool used in all three experiments consisted of 1,385 words drawn from The University of South Florida Word Association, Rhyme and Word Fragment Norms (Nelson, McEvoy, & Schreiber, 1998).

Design. Instructions and stimuli were displayed for the participant on a color monitor. The effect of repeated study presentations on item and associative information was investigated with the use of a yes–no recognition task. During the study phase of the experiment, participants were presented word pairs, half of which were presented once and the other half six times each. Except as noted below, word pairs were inserted randomly into the study list, and a different random order was used for every participant.

Word pairs were assigned two at a time to either the intact or the rearranged word-pair condition. If the two word pairs were designated as rearranged word pairs, then their rightmost words were switched at the time of testing, thereby maintaining the left–right word ordering of the pairs. The two rearranged word pairs were inserted into the study list two at a time, within a limited range of nine presentations. In this way, the two word pairs would occur with, at most, seven presentations between them. The specific range of nine presentations into which the pairs were inserted was determined randomly for each set of word pairs (e.g., it might be the 15th through the 23rd study presentations).

If the word pairs were to be repeatedly studied, then they were inserted such that every repetition was spaced, on average, about nine presentations away from the prior study of the same word pair. To accomplish this experimental design, we instructed the computer to randomly select a range of 54 presentations from within the study list (e.g., the 18th through the 134th presentations). This range was then divided into six blocks of nine continuous presentations, and the two word pairs were randomly inserted into each of the six adjacent blocks. Altogether, the study list consisted of 352 presentations. Inserted into the study list were 40 weak pairs (which appeared once each) and 40 strong pairs (which appeared six times each). In addition, there were 72 pairs that were not later tested (the first 12 pairs on the list and 60 filler pairs).

After the study list was presented, participants were given a recognition test. The recognition test included 40 intact pairs (half weak, half strong) and 40 rearranged pairs (half weak, half strong). The various test pairs were all randomly intermixed.

Procedure. After the participants signed a consent form and read instructions presented on the computer monitor, the study list was presented. Pairs on the study list were presented at a rate of one item or pair every 3 s. After the list was completed, the recognition test was presented.

For each word pair, participants were asked to decide (as quickly and accurately as possible) whether the two words had occurred together during the study list. For intact pairs, the correct answer was "yes," whereas for

rearranged pairs, the correct answer was "no." Participants responded to each recognition test item or pair by pressing one of six keys on the keyboard that were labeled —, –, -, +, ++, and +++. These keys represented "certain no," "no," "perhaps no," "perhaps yes," "yes," and "certain yes," respectively.

Results and Discussion

Hit and false-alarm rate analyses. The hit rates and false-alarm rates for each condition of Experiment 1 are reported in Table 1. The values in this table were collapsed across all "yes" responses, regardless of the level of confidence ("perhaps yes," "yes," or "certain yes"). Unless otherwise noted, all reported effects were significant at $p < .05$.

The data reveal that the strengthening manipulation had a very large effect on the hit rate for intact pairs but virtually no effect on the false-alarm rate for rearranged pairs. An analysis of variance (ANOVA) performed on the data shown in Table 1 revealed a main effect of strength (strong vs. weak), $F(1, 35) = 129.13$, $MSE = 0.01$; a main effect of pair type (intact vs. rearranged), $F(1, 35) = 173.88$, $MSE = 0.05$; and a significant interaction, $F(1, 35) = 120.62$, $MSE = 0.01$. The small difference in the false-alarm rate in the weak and strong conditions (.23 vs. .25) was in the direction predicted by the familiarity-based account, but the difference did not approach significance.

Whereas the familiarity-based account (Equations 1 and 2) predicted that the false-alarm rate to rearranged pairs would be higher in the strong condition, the retrieval-based account (Equations 3 and 4) made the opposite prediction. Why, then, were the false-alarm rates nearly equal, $F(1, 35) = 1.07$; $MSE = 0.01$? The simplest explanation for the present results, and the one that is the most compatible with earlier research, is that the enhanced familiarity associated with the individual items of a rearranged pair in the strong condition was counteracted by the increased efficiency of retrieval in that condition (which would allow participants to reject rearranged pairs on that basis). If the opposing forces (familiarity vs. recall) were approximately equal, then they would cancel each other out, thereby leaving the false-alarm rate to rearranged pairs essentially unchanged. For the intact pairs, by contrast, the two forces work in harmony, so the hit rate increased to a considerable degree.

Some evidence in favor of this idea is provided by the confidence ratings associated with responses made to rearranged pairs. Although the false-alarm rates in the weak and strong conditions were essentially equal, average confidence ratings were not. More specifically, participants gave higher average confidence ratings both to their correct "no" responses (i.e., to their correct rejections) and to their incorrect "yes" responses (i.e., to their false alarms) in the strong condition. For correct "no" responses, the mean confidence ratings, rated on a scale from 1 (*low*) to 3 (*high*), were 2.23 in the strong condition versus 1.93 in the weak condition, $t(35) = 5.00$. For incorrect "yes" responses, the mean ratings were 1.83 in the strong condition versus 1.58 in the weak condition, a difference that was marginally significant, $t(28) = 1.93$, $p = .064$. Participants were included in the latter t test only if their false-alarm rates were greater than 0 for both conditions (hence, the differing degrees of freedom for the two tests). A participant who yields a false-alarm rate of 0 in one condition provides no confidence ratings to false alarms that can be compared with the other condition.

Table 1
Proportion of "Yes" Responses to Each Pair Type in Experiment 1

Pair type	Weak	Strong
Intact	0.54	0.89
Rearranged	0.23	0.25

The higher confidence to correct "no" responses in the strong condition presumably reflects performance on those pairs for which retrieval was successful (which theoretically results in a high-confident rejection). Successful retrieval would presumably occur more often in the strong condition than in the weak condition. The higher confidence to incorrect "yes" responses in the strong condition may reflect performance on those pairs for which retrieval was unsuccessful but item familiarity was very high. High item familiarity is something that would also occur more often in the strong condition. A formal specification of this model is presented below as part of our discussion of the ROC data. Other explanations of the equivalent false-alarm rates in the weak and strong conditions are possible, and some of these are tested in the ensuing experiments. For the moment, however, we turn to the ROC analyses to see what new information they may provide.

ROC analyses. The confidence ratings that participants supplied for each recognition decision can be used to construct confidence-based ROC plots (e.g., Macmillan & Creelman, 1991; Stretch & Wixted, 1998; Yonelinas, 1997). Prior work by Yonelinas (1997) suggests that the ROC plots for associative recognition (intact vs. rearranged) will be linear in accordance with Equation 5. The conditions studied by Yonelinas most closely match the weak condition here, but the high-threshold model predicts a linear ROC in the strong condition as well. Presumably, both *Ro* and *Rn* (the probabilities of successful retrieval for "old" intact and "new" rearranged pairs, respectively) in Equation 5 will be higher in the strong condition than in the weak condition, but the form of the ROC function would not be expected to change.

Figure 2 presents the relevant ROC data for the weak and strong conditions. The ROC in the upper panel shows the weak intact pairs versus the weak rearranged pairs, whereas the lower panel shows the strong intact pairs versus the strong rearranged pairs. We created these plots by pooling confidence data over participants, and then fit the ROCs by the two-parameter high-threshold model (the parameters being *Ro* and *Rn*) and by the two-parameter signal-detection model (with one parameter representing the distance between the two distributions and the other representing their relative standard deviations).

The data were pooled over participants because individual participants did not yield enough data from each condition to produce a meaningful ROC. In previous research in which similar pooling was involved, conclusions based on pooled ROC analyses corresponded to conclusions based on individual participant ROC analysis (e.g., Ratcliff, Sheu, & Gronlund, 1992; Stretch & Wixted, 1998; Yonelinas, 1997). We address this issue empirically in Experiment 4 and arrive at the same conclusion, but for now we analyze pooled data.

We used the maximum likelihood estimation procedure described by Ogilvie and Creelman (1968) to fit the ROC data, except that for the signal-detection fits, we used a close approximation of the cumulative Gaussian (taken from Abramowitz & Stegun, 1970) instead of the somewhat less accurate logistic ap-

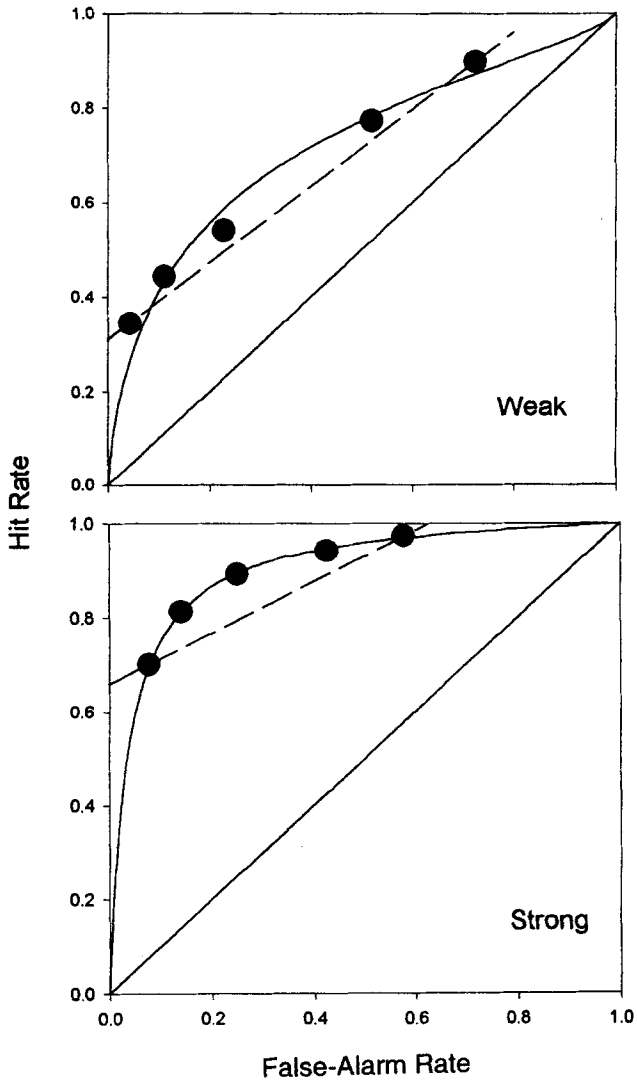

Figure 2. Confidence-based intact-versus-rearranged receiver operating characteristic plots for the weak and strong conditions of Experiment 1 (upper and lower panels, respectively). In both graphs, the dashed line indicates the maximum likelihood fit of the high-threshold model and the solid curve indicates the maximum likelihood fit of the signal-detection model. The high-threshold model does not provide a good visual fit in the strong condition because the data points are associated with different numbers of observations, and points with more observations exert a greater influence on the fit than those with fewer observations.

proximation.[2] Whereas the high-threshold fits involved estimating two retrieval parameters, *Ro* and *Rn*, the signal-detection fits involved estimating two distributional parameters, d_e and r. The parameter d_e is a measure analogous to d' that takes into account unequal variances, whereas r represents the standard deviation of the lure distribution divided by the standard deviation of the target distribution. If r is less than 1, as it usually is (e.g., Ratcliff et al., 1992), then the standard deviation of the signal (e.g., intact) distribution is greater than that of the noise (e.g., rearranged) distribution. For the intact and rearranged distributions shown in Figure 1, r would equal 1 because the two distributions have the same standard deviation.

As a technical aside, we note that if the models were fit to the empirical ROCs by the method of least squares, then, as just indicated, two parameters would be estimated for each model. Although a least-squares fit is usually adequate in practice, that approach is theoretically problematic because it assumes measurement error in the vertical direction only. In reality, both measures of an ROC (i.e., the hit rate and the false-alarm rate) are associated with error variance. For this reason, maximum likelihood estimation is a better option (Ogilvie & Creelman, 1968), and it is the option we used. This method actually involves estimating seven parameters for each fit, not two, because five confidence criteria (for the detection fits) or five confidence-specific guessing probabilities (for the high-threshold fits) are also estimated. Because the estimated values of the confidence and guessing parameters are not of theoretical interest and do not determine the shape of the ROC in any way (which is why a two-parameter least-squares fit of ROC data is possible), only the estimates of the two defining parameters of each model are reported.

Table 2 presents the best fitting parameters and chi-square goodness-of-fit statistics for both models. A significant chi-square indicates that the deviations are greater than would be expected on the basis of chance (i.e., the larger the chi-square, the poorer the fit).[3] In the weak condition, neither model provides an excellent fit, with both yielding a chi-square value of about 15. The expected chi-square value for 3 degrees of freedom is 3, so both models show significant deviations. Essentially, as is evident in Figure 2, the weak ROC is more linear than the detection model predicts and is more curvilinear than the high-threshold model predicts. Although it has been shown before by Yonelinas (1997), the fact that the high-threshold model rivals (and even slightly outperforms) the detection model in associative recognition is remarkable, considering that in item-recognition tasks, detection theory invariably enjoys a strong advantage (as we shall see again in the next two experiments). Thus, although the ROCs were not perfectly linear, the results from the weak condition largely replicate the results reported by Yonelinas.

As noted by Yonelinas (1997), the high-threshold model does not need to assume any role for familiarity. However, the equivalent false-alarm rates for weak and strong rearranged pairs shown in Table 1 raise the possibility that familiarity for the individual items is being taken into account and that it is being offset by retrieval (at least, that is the simplest explanation for this finding). How can this idea be reconciled with a linear ROC?

Whereas the retrieval-based high-threshold model that predicts a linear ROC must assume that continuously distributed associative

[2] The fits were also performed using the logistic approximation, and the differences were, for the most part, negligible.

[3] It is difficult to know what effect pooling data over participants has on the alpha level for the chi-square test, and that uncertainty should be kept in mind when considering the reported chi-square values. Simulations described later suggest that the impact is not extreme, but the issue was not formally investigated here. The most conservative approach would be to compare the relative chi-square values produced by the competing models without giving too much weight to whether deviations from a particular model are significant.

Table 2
Maximum Likelihood Parameter Estimates and Chi-Square Goodness-of-Fit Statistics for the Signal-Detection and High-Threshold Models Fit to the Intact Versus Rearranged ROC Data From Experiment 1

Condition	Model	p1	p2	χ^2	*df*
Weak	Detection	0.735	0.875	15.61*	3
	Threshold	0.310	0.152	14.42*	3
Strong	Detection	0.864	1.912	3.08	3
	Threshold	0.658	0.375	79.89*	3

Note. For the detection model, p1 = r and p2 = d_e. For the threshold model, p1 = *Ro* and p2 = *Rn*. For 3 degrees of freedom, a chi-square value of 7.81 is significant at the .05 level. N = 1440 for all chi-square tests. ROC = receiver operating characteristic.
* $p < .05$.

familiarity (A) plays no role (because associative familiarity would introduce curvilinearity), it need not assume that item familiarity (I) plays no role. In fact, if retrieval fails, participants may respond "yes" if the combined familiarity of the individual items exceeds a decision criterion (cf. Mandler, 1980). According to this idea, the hit rate and false-alarm rate would be as follows:

$$H = Ro + (1 - Ro)p(I > c) \text{ and} \quad (6)$$

$$FA = (1 - Rn)p(I > c), \quad (7)$$

where $p(I > c)$ represents the probability that item familiarity exceeds a decision criterion. Solving Equation 7 for $p(I > c)$ and substituting the result into Equation 6 yields Equation 5, so this model still predicts a linear ROC. However, unlike the original retrieval-based model, the new model assumes that participants do take item familiarity into account in deciding whether or not a pair was previously encountered (even though there is no reason why they should). What is regarded as "guessing" in the standard model is actually responding on the basis of item familiarity in this slightly revised model.

The simple model expressed in Equations 6 and 7 is consistent with all of the available data discussed so far. Equation 7, for example, predicts the offsetting effects for rearranged word pairs when conditions change from weak to strong. That is, as strength increases, *Rn* will increase such that the first term in Equation 7, $1 - Rn$, will decrease. However, the second term, $p(I > c)$, will increase, serving to offset the retrieval advantage. In addition, this model is compatible with (a) prior research by Hockley (1992) and Nairne (1983) that supported a role for item familiarity, (b) prior research by Clark and Hori (1995) that supported a role for retrieval, (c) the linear ROCs reported by Yonelinas (1997), and (d) the somewhat linear ROC obtained in the weak condition here. The slight deviations from linearity we observed are admittedly not consistent with this account.

If the high-threshold model represented by Equations 4 and 5 (or the elaborated version of it represented by Equations 6 and 7) is correct, then the linear ROC observed in the weak condition should be observed in the strong condition as well. Completely contrary to this prediction, however, the ROC for the strong condition shown in the lower panel of Figure 2 is clearly curvilinear. The goodness-of-fit measures shown in Table 2 reveal that the high-threshold model offers an extremely poor fit, which statistically confirms what is visually apparent. It is important to note that the line produced by the best-fitting high-threshold model does not pass through the ROC points in a way that intuition suggests that it should because the points are not given equal weight in the maximum likelihood fits (and the line tries to deviate less from points that are based on more observations). In contrast to the high-threshold model, the signal-detection model offers an excellent fit, and no systematic deviations are apparent (points with more observations were given more weight in these fits as well). Moreover, the obtained value of r in the strong condition increased considerably relative to the weak condition and is closer to 1 than is usually the case in recognition memory (cf. Ratcliff et al., 1992; Stretch & Wixted, 1998). In fact, a fit of the one-parameter detection model with r fixed at 1 still yielded a nonsignificant chi-square, $\chi^2(4, N = 1440) = 6.74$, and the fit was not significantly improved by allowing r to take on a value less than 1. By way of comparison, forcing r to assume a value of 1 in the weak condition yielded a significantly worse fit by far compared with when r was free to vary.

What this result suggests is that there may be something wrong with even the modified model represented by Equations 6 and 7. What could explain the apparent transformation of a nearly linear ROC into a clearly curvilinear ROC when strength is increased? A possible solution to this puzzle, which we simply mention briefly here and then work out in more detail later, is that associative information is not all or none (as the high-threshold account assumes) but is instead better regarded as some or none. That is, in agreement with the high-threshold view, a participant may fail to retrieve any associative information for a given pair (none). However, in contrast to the high-threshold view, participants may retrieve associative information to varying degrees (some). According to this idea, associative information, when it is available, is a continuous variable like item familiarity is. Unlike item familiarity, associative information is not always available. How this idea can simultaneously explain a (nearly) linear ROC in the weak condition that becomes a symmetric curvilinear ROC in the strong condition is discussed in more detail after presenting additional relevant information provided by the next two experiments.

Experiment 2

Experiment 2 was similar to Experiment 1 in that participants studied both weak and strong word pairs and were later presented with a recognition test in which they were asked to discriminate between pairs that were intact and pairs that were rearranged. Thus, part of the purpose of Experiment 2 was to replicate the equivalent false-alarm rates in the weak and strong conditions as well as the curious ROC results in those two conditions. In addition, the recognition test now included pair lures (i.e., pairs comprised of items not previously encountered during the experimental session) as well. Unlike rearranged pairs, which may elicit the retrieval of associative information, pair lures must be judged without the benefit of any associative information.

The inclusion of pair lures offers several useful pieces of information. For example, one possible explanation for the equivalent false-alarm rates observed in Experiment 1 is that participants simply ignore item familiarity (so there is no upward pressure on the false-alarm rate in the strong condition) and do not rely on recollection to reject rearranged pairs (so there is no downward pressure either). Instead, participants might respond entirely on the

Table 3
Proportion of "Yes" Responses to Each Pair Type and Item Type in Experiment 2

Pair and item type	Weak	Strong
	Pairs	
Intact	0.56	0.84
Rearranged	0.23	0.26
Pair lure	0.07	
	Items	
Item target	0.61	0.86
Item lure	0.31	

basis of the associative familiarity of a word pair. If so, then one might expect the false-alarm rate to rearranged pairs to equal that of pair lures. In both cases, associative information is absent (except, perhaps, for random-associative noise) and, if item information is ignored as well, there is no reason to expect differing false-alarm rates. Also, as described in more detail later, the inclusion of pair lures allows one to test the variability of associative information (if that information is indeed continuously distributed) relative to the variability of item information. This can be done, for example, by constructing an ROC consisting of intact pairs versus pair lures. Finally, for comparative purposes, Experiment 2 also included individual items on the study list as well as individual item lures on the recognition test.

Method

Participants. The participants were 30 undergraduates from the University of California, San Diego, who met the criterion of English fluency by age 7. Their participation fulfilled a lower division psychology course requirement.

Materials. The word stimuli were the same as those used in Experiment 1.

Design and procedure. The design and procedure were similar to Experiment 1, except that individual items (both weak and strong) also appeared on the study list, and the recognition test involved intact and rearranged pairs (as before) as well as item targets, item lures, and pair lures. Strengthened items were randomly inserted into the study list in a manner similar to the strengthened pairs (i.e., six insertions spaced an average of nine study presentations apart).

The study list consisted of 380 presentations. Inserted into the study list were 32 strong pairs, 16 strong items, 32 weak pairs, and 16 weak items. An additional 32 study pairs served as fillers and were not later tested (the first 12 pairs of the list were also not tested).

After the study list was presented, participants were given a recognition test. The recognition test included 32 intact pairs (half weak, half strong), 32 rearranged pairs (half weak, half strong), and 32 target items (half weak, half strong). Also included were 64 pair lures (neither word previously studied) and 32 item lures (nonpreviously studied words). The various test items were all randomly intermixed. On the recognition test, participants were instructed to say "yes" both to pairs they had seen appear together on the list and to items that had appeared on the list.

Results and Discussion

Hit and false-alarm rate analyses. The hit rates and false-alarm rates for each condition of Experiment 2 are reported in Table 3. The values in this table were again collapsed across all "yes" responses regardless of the level of confidence ("perhaps yes," "yes," or "certain yes"). The data once again reveal that the strengthening manipulation had a very large effect on the hit rate for intact pairs but virtually no effect on the false-alarm rate for rearranged pairs. An ANOVA performed on the data shown in Table 3 revealed a main effect of strength (strong vs. weak), $F(1, 29) = 51.11$, $MSE = 0.01$, a main effect of pair type (intact vs. rearranged), $F(1, 29) = 93.45$, $MSE = 0.07$, and a significant interaction $F(1, 29) = 27.61$, $MSE = 0.01$. As before, the small difference in the false-alarm rate in the weak and strong conditions (.23 vs. .26) was in the direction predicted by the familiarity-based account, but the difference did not approach significance $F(1, 29) < 1$, $MSE = 0.01$. For the items, the hit rate to the strong items was, not surprising, significantly greater than the hit rate to the weak items, $F(1, 29) = 70.15$, $MSE = 0.01$.

As indicated earlier, one possible explanation for the equivalent false-alarm rates to weak and strong rearranged pairs is that item familiarity was ignored altogether and responding was based entirely on the presence or absence of associative familiarity (i.e., the familiarity of the word pair). Because a rearranged word pair is equally lacking in associative familiarity whether items were seen once or many times, the false-alarm rate would be unaffected. However, some evidence against this possibility is provided by a comparison of the false-alarm rates to rearranged word pairs and to pair lures in Table 3. Specifically, the false-alarm rate to weak rearranged word pairs was significantly higher than the false-alarm rate to pair lures, $F(1, 29) = 59.35$, $MSE = 0.01$, even though associative information is lacking for both. Thus, the difference in false-alarm rate here may indicate that participants responded differentially to these pairs on the basis of differences in item familiarity (which would be higher for rearranged pairs).

An alternative possibility is that participants attended to item familiarity (thereby accounting for the false-alarm rate difference between rearranged pairs and pair lures) but that item familiarity, which increases when a word pair is first presented, did not further increase when the pairs were repeated in the strong condition. If participants were somehow able to focus their efforts entirely on strengthening the associative bond (A) when a pair was seen for the second or third time without further strengthening item information (i.e., without increasing I), then the familiarity-based model predicts equivalent false-alarm rates. This possibility is tested in Experiment 3, which, unlike Experiment 2, tested item recognition for items that originally appeared as a member of a weak or strong pair (in Experiment 2, the individual test items appeared as individual items on the study list). While this is a possibility that needs to be tested, the simplest explanation for the present results is, once again, that the enhanced familiarity associated with the individual items of a rearranged pair in the strong condition was counteracted by the increased efficiency of retrieval in that condition. If the opposing forces were approximately equal, then they would cancel each other, thereby leaving the false-alarm rate to rearranged pairs essentially unchanged.

The confidence ratings to responses made to rearranged pairs again support this view. Even though the false-alarm rates to weak and strong rearranged pairs were the same, participants were again more confident in both their correct "no" and incorrect "yes" responses to strong rearranged pairs. The mean confidence rating to correct "no" responses was 2.36 in the strong condition versus 2.19 in the weak condition, $t(29) = 3.48$, whereas the mean confidence rating to incorrect "yes" responses was 2.17 in the

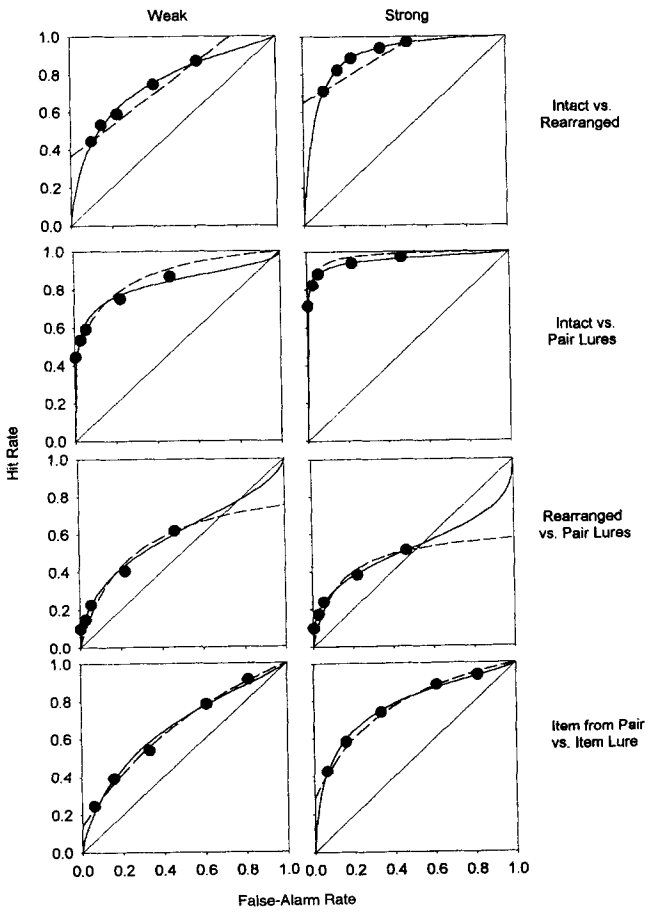

Figure 3. Confidence-based receiver operating characteristic plots for the weak (left panel) and strong (right panel) conditions of Experiment 2. The upper panel plots performance for intact versus rearranged pairs, the second panel from the top plots performance for intact pairs versus pair lures, the third panel plots performance for intact versus rearranged pairs, and the bottom panel plots performance for items versus item lures. In each graph, the dashed line indicates the maximum likelihood fit of the high-threshold model and the solid curve indicates the maximum likelihood fit of the signal-detection model.

strong condition versus 1.89 in the weak condition, $t(25) = 2.37$. As indicated earlier, the higher confidence to correct "no" responses presumably reflects the influence of those pairs for which retrieval was successful, which is more likely to occur in the strong condition. The higher confidence to incorrect "yes" responses in the strong condition may reflect the influence of those pairs for which retrieval was unsuccessful but item familiarity was very high (because the items had been seen repeatedly on the study list).

ROC analyses. Figure 3 presents the relevant ROC data for the weak and strong conditions, respectively, and Table 4 presents the maximum likelihood parameter estimates and chi-square goodness-of-fit statistics for the two-parameter high-threshold model (based on Equations 4 and 5) and for the two-parameter signal-detection model. These ROCs were constructed and analyzed in the same manner as those in Experiment 1.

The upper panel of Figure 3 shows the ROC results for intact versus rearranged pairs from the weak (upper left) and strong (upper right) conditions. The plot for the weak condition, which corresponds most closely to the conditions studied by Yonelinas (1997), is again equally well fit by both the signal-detection model and the high-threshold model. Although the data do not deviate significantly from either model, the results are probably best described as falling between the predictions of the two models. That is, deviations from the best fitting high-threshold model are in the direction predicted by signal-detection theory and vice versa.

Once again, however, the ROC for the strong condition is undeniably curvilinear. The results shown in Table 4 confirm what is visually apparent, namely, that the linear high-threshold model provides an extremely poor fit. On the other hand, the detection model again provides an excellent fit, and, again, the estimated value of r is sufficiently close to 1 that an equal-variance model may apply. Indeed, the fit of a one-parameter detection model (with r fixed at 1) to the strong intact-versus-rearranged ROC yields an overall chi-square with 4 degrees of freedom that is still not significant, $\chi^2(4, N = 1440) = 2.37$, so, obviously, the fit is not significantly improved by allowing r to vary. It is curious, and presumably theoretically significant, that the mere act of strengthening pairs transforms a nearly linear ROC (something not often seen in recognition tasks) into a curvilinear ROC that is adequately described by an equal-variance detection model (something that is also not often seen in recognition). As indicated earlier, recognition memory ROCs are typically curvilinear and are invariably best fit by an unequal variance detection model (e.g., Ratcliff et al., 1992).

The ROC plots shown in the second panel from the top in Figure 3 are similar to those shown in the top panel in that the hit rates are based on responses to intact weak and strong pairs (as before). The difference is that the false-alarm rates are now based on responses to lures that consisted of a pair of new items (i.e., the pair lures). If associative-recognition decisions are based on retrieval or, in the absence of retrieval, guessing, then it could be argued that these ROC plots ought to be linear. That is, as indicated earlier, Equation 3 states that the hit rate for intact pairs is equal to $Ro + (1 - Ro)g$. What is the corresponding false-alarm rate equation for pair lures? Whereas the false-alarm rate to rearranged pairs is, according to Equation 4, $(1 - Rn)g$, the false-alarm rate for pair lures is, perhaps, simply equal to g (i.e., no retrieval information is available for these pairs, so participants must always guess). Thus, if $H = Ro + (1 - Ro)g$ and $FA = g$, then the predicted relationship between H and FA (i.e., the ROC plot) is $H = Ro + (1 - Ro)FA$, which is a linear function.

On the other hand, one need not assume that the probability of guessing "yes" to a pair lure is the same as the probability of guessing "yes" to an intact or rearranged pair for which no retrieval information is available. Indeed, the fact that the false-alarm rates to pair lures are much lower than that of the rearranged pairs already shows that the guessing rates must differ. If they did not, then the false-alarm rate pattern would actually be reversed because $(1 - Rn)g$, the predicted false-alarm rate to rearranged pairs, is less than g, the predicted false-alarm rate to pair lures. It is perhaps not surprising, then, that fits of this version of the linear high-threshold model to the obviously nonlinear pair-lure ROCs were extremely poor (and are not discussed further).

Within the high-threshold framework, the problem of differential guessing rates is easily resolved by assuming that guessing rates are determined by item familiarity, as in Equations 6 and 7. Because the items of an intact or rearranged pair are more familiar than those of a pair lure, a higher rate of guessing for those pairs

Table 4
Maximum Likelihood Parameter Estimates and Chi-Square Goodness-of-Fit Statistics for the Signal-Detection and High-Threshold Models Fit to the ROC Data From Experiment 2

Condition	Model	p1	p2	χ^2	*df*
Intact vs. rearranged[a]					
Weak	Detection	0.672	0.847	5.18	3
	Threshold	0.358	0.149	6.59	3
Strong	Detection	0.867	1.690	0.45	3
	Threshold	0.656	0.367	15.45*	3
Intact vs. pair lure[b]					
Weak	Detection	0.511	1.166	4.84	3
	Threshold	0.379	0.651	17.44*	3
Strong	Detection	0.484	2.231	1.27	3
	Threshold	0.666	1.277	12.66*	3
Rearranged vs. pair lure[b]					
Weak	Detection	0.750	0.285	11.01*	3
	Threshold	0.233	0.788	12.75*	3
Strong	Detection	0.547	0.096	5.07	3
	Threshold	0.381	1.132	17.63*	3
Item target vs. item lure[c]					
Weak	Detection	0.752	0.793	4.86	3
	Threshold	0.206	0.495	7.70	3
Strong	Detection	0.678	1.604	4.62	3
	Threshold	0.434	1.053	12.70*	3

Note. For all fits of the detection model, p1 = r and p2 = d_e. For the threshold model, p1 = *Ro* and p2 = *Rn* for the intact vs. rearranged fits; p1 = *Ro* and p2 = d' for the intact vs. pair lure and item vs. item lure fits, and p1 = *Rn* and p2 = d' for the rearranged vs. pair lure fits. For 3 degrees of freedom, a chi-square value of 7.81 is significant at the .05 level. ROC = receiver operating characteristic.
[a] $N = 960$. [b] $N = 2400$. [c] $N = 1440$.
* $p < .05$.

would not be surprising. Thus, when fitting the high-threshold model to the intact versus pair-lure ROCs, we assumed that the hit rate was governed by Equation 6 (rather than Equation 4), and that the false-alarm rate was equal to the probability that item familiarity exceeded the decision criterion, $p(I_{PL} > c)$, where I_{PL} represents the familiarity of the items of a pair lure. This predicted false-alarm rate corresponds to Equation 7, except that retrieval is not assumed to play a role, so *Rn* is set to 0.

On any trial in which retrieval played no role (i.e., on all trials involving pair lures and on a subset of trials involving intact pairs), item familiarity was assumed to determine the level of confidence in accordance with a standard signal-detection model. As usual, the relevant item familiarity values were assumed to be normally distributed. For the pair lures, the item familiarity distribution was arbitrarily set to a mean of 0 and a standard deviation of 1. For the intact pairs, the item familiarity distribution had a mean of d' (reflecting higher mean familiarity for these items) and a standard deviation of 1, where d' was a parameter estimated from the data. It is important to note that for these fits, this parameter actually is d', not d_e, because an equal-variance model is assumed (allowing for unequal variances here is possible but would introduce a third free parameter).

As an aside, it should be noted that the linear high-threshold model that was fit to the intact versus rearranged ROCs could be derived either from Equations 4 and 5 or from Equations 6 and 7. Because the item familiarities are presumably equal for both intact and rearranged pairs, the term $p(I > c)$ in Equations 6 and 7 are equal, so this aspect of the equation simply divides out (such that the predicted ROC is strictly linear, as shown by Equation 5). For the ROCs involving pair lures, on the other hand, the version of the high-threshold model that allows for differing levels of item familiarity predicts a curvilinear ROC with a y-intercept greater than 0 (whereas the curvilinear path predicted by signal-detection theory has a y-intercept of 0).

The two parameters of interest in the intact versus pair-lure ROC fits were *Ro* and d' (although, as in all of these maximum likelihood fits, seven parameters were actually estimated, including the five confidence criteria). As is evident from Figure 3, this version of the high-threshold model offered a reasonably accurate fit, and it is difficult to distinguish the threshold and detection models on these grounds alone. As shown in Table 4, the fit of the high-threshold model was statistically somewhat less accurate than that provided by the detection model. Whereas the deviations from the best-fitting detection model were not significant in either condition, the deviations from the threshold model were significant in both conditions. In spite of the significant deviations from the high-threshold model, the estimates of the parameter, *Ro*, in both strength conditions are reasonably close to the corresponding estimates of *Ro* obtained from the intact versus rearranged fits. This is as it should be, because they are redundant estimates of how often retrieval occurred for the intact pairs.

With regard to the signal-detection parameter estimates, the most important finding may be that the ratio of noise-to-signal distribution standard deviations (i.e., r) was approximately 0.5. Interpreted in terms of signal-detection theory, this result suggests that the standard deviation of the intact-pair distribution is twice that of the pair-lure distribution. This represents yet another ROC result rarely seen in recognition memory, for which r values are almost always in the .70 to .80 range (Ratcliff et al., 1992).

The next panel of ROC plots is similar, except that now the rearranged pairs represent the "signal" distribution, and the pair lures again represent the "noise" distribution. Although the correct answer is "no" to both kinds of pairs, the false-alarm rate was higher for rearranged pairs, so these pairs can be regarded as the signal distribution for analytic purposes. The high-threshold fits were performed in the same way they were for the intact versus pair-lure ROCs, except that the hit rate was given by Equation 7 rather than Equation 6. Thus, the two parameters of interest are *Rn* and *d'*. The chi-square goodness-of-fit statistics shown in Table 4 reveal that the models provided comparable fits in the weak condition (with deviations from the best-fitting models significant in both cases), but the detection model provided a better fit than the high-threshold model in the strong condition. Visually, the fits appear to be comparable, although systematic deviations from the best fitting threshold model are discernable in the strong condition. Even so, the estimates of the high-threshold parameter, *Rn*, in both strength conditions (Parameter p1 in Table 4) are reasonably close to the corresponding estimates of *Rn* obtained from the intact versus rearranged fits (Parameter p2 in Table 4). This would be expected because, from the high-threshold theory point of view, these are redundant estimates (both estimating how often retrieval succeeded for rearranged pairs). The estimates of *d'* from the fits of the high-threshold model to the intact versus pair-lure ROCs (Parameter p2 in Table 4) should also roughly equal the corresponding estimates of *d'* from the fits of the high-threshold model to the rearranged versus pair-lure ROCs (also Parameter p2 in Table 4). In both cases, this parameter theoretically captures the familiarity of the items, which should be equal for intact and rearranged pairs. As shown in Table 4, the estimates are reasonably similar.

The best fitting detection model again implies that the variance of the rearranged distribution is much greater than that of the pair-lure distribution, especially in the strong condition. Thus, from the point of view of signal-detection theory, this result suggests that a distribution of evidence values that involves associative information (the intact distribution or the rearranged distribution) is much more variable than a distribution of evidence values that does not involve associative information (the pair-lure distribution).

The lower panel of Figure 3 presents the data for the weak (lower left) and strong (lower right) items versus item lures. These were included as a check to see whether our participants yielded typical data with respect to individual items, and, indeed, they did. The items exhibit the expected curvilinearity in both the weak and strong conditions. A linear high-threshold model would obviously not fit these data well, and no one advocates such a model for item-recognition memory anymore. A more viable version of the high-threshold account relies on the exact same logic that was just applied to the intact versus pair lure ROCs. That is, as argued by Yonelinas (1997), a response to an individual target item may be based on retrieval or, in the absence of retrieval, familiarity (which helps to distinguish targets from lures, because targets are associated with higher levels of familiarity). For lures, by contrast, responding would be based entirely on familiarity, as retrieval plays no role. If one assumes an equal-variance signal-detection model for the subset of trials based on familiarity, which is what Yonelinas assumed for item ROCs, then the model we fit to the intact versus pair-lure ROCs is the same high-threshold model that applies here. As shown in Table 4, this version of the high-threshold model provides a fit that is comparable with that of the detection model in the weak condition (and obtained deviations are not significant in either case) and is somewhat inferior to that of the detection model in the strong condition. Visually, the fit of the high-threshold model is quite good in both conditions. However, small but systematic deviations are apparent in the strong condition.

Finally, for the item ROCs, the best-fitting detection model implies that the variance of the target distribution is somewhat greater than that of the lure distribution, which is the usual finding (i.e., *r* is usually in the .7 to .8 range). Unlike the strong intact-versus-rearranged ROC, forcing the value of *r* to equal 1 for the item ROC fits yielded very poor (and significantly worse) fits compared with when *r* was free to assume a value less than 1. The goodness-of-fit statistics for the fits with *r* fixed at 1 were $\chi^2(4, N = 1440) = 30.78$ and $\chi^2(4, N = 1440) = 44.68$ (both highly significant) for the weak and strong fits, respectively. The fact that an equal-variance detection model can be rejected for item recognition is well known (Ratcliff et al., 1992).

Experiment 3

Experiment 3 was designed to replicate all of the major effects of Experiments 1 and 2 (especially the ROC results just discussed) and to evaluate one possible alternative explanation for the equivalent false-alarm rates to the weak and strong rearranged pairs. More specifically, this experiment provided a more direct test of the idea that repeating pairs not only strengthens memory for the association but also increases the familiarity of the two items comprising the pair.

In Experiment 2, the strong and weak items that appeared on the recognition test had appeared as individual items on the study list. Thus, the individual items from strong and weak pairs were never tested individually to ensure that item familiarity was higher in strong pairs than in weak pairs. If item familiarity did not increase as a result of repeating pairs (e.g., if participants were able to concentrate mainly on strengthening the associative connection between items when the pair was repeated), then that might explain why the false-alarm rates for weak and strong rearranged pairs were equivalent in Experiments 1 and 2.

If the familiarity of individual items does increase when pairs are repeated (as one would naturally expect), then the hit rate for items drawn from strong pairs should exceed the hit rate of items drawn from weak pairs. On the other hand, the mere fact that the hit rate for strong items exceeds that of weak items does not prove that increased item familiarity is responsible. It could be increased by all-or-none retrieval (such that participants would respond "yes" if they retrieved the member with which the item was originally paired). An analysis of the relevant ROC data may help to shed light on this issue. If all-or-none retrieval operates to a greater extent for items that were studied as part of a pair than for items that were studied individually, then the fit of the high-threshold model to the ROC data from this condition might be expected to yield higher estimates of *Ro* compared with the fit for items that were studied individually. This is not a necessary prediction of the high-threshold account (i.e., all-or-none retrieval may be equally involved whether items are presented as part of a

pair or individually), but such a result would seem to fit naturally with that account.

Method

Participants. The participants were 30 undergraduates from the University of California, San Diego, who met the criterion of English fluency by age 7. Their participation fulfilled a lower division psychology course requirement.

Design and procedure. The design was identical to that of Experiment 1, except that the precise mix of item types differed slightly (as noted below), and half of the tested individual words were drawn from word-pair study presentations. The other half were from study presentations of individual words. The study list consisted of 312 presentations, followed by the test list, which consisted of 160 presentations. The recognition test included 20 intact word pairs (half weak, half strong), 20 rearranged word pairs (half strong, half weak), 20 items drawn from word pairs (half strong, half weak), and 20 items studied as items (half strong, half weak). Tests of items drawn from word-pair study presentations were equally likely to be either the left or the right word (the other word in the word pair was not tested). The recognition test also included 40 pair lures and 40 item lures. The first 12 study presentations were not later tested. On the recognition test, participants were instructed to say "yes" to intact pairs and to items they recognized as having been seen on the list (they were not specifically informed that some items from pairs would be tested as items on the recognition test).

Results and Discussion

Hit and false-alarm rate analyses. The hit rate and false-alarm rate for each condition of Experiment 3 are reported in Table 5. The values in this table were once again collapsed across all "yes" responses regardless of the level of confidence ("perhaps yes," "yes," or "certain yes"). The results presented in Table 5 show yet again that the strengthening manipulation had a very large effect on the hit rate for intact pairs but almost no effect on the false-alarm rate for rearranged pairs. An ANOVA performed on the data presented in Table 5 revealed a main effect of strength (strong vs. weak), $F(1, 29) = 36.99$, $MSE = 0.02$, a main effect of pair type (intact vs. rearranged), $F(1, 29) = 125.39$, $MSE = 0.06$, and a significant interaction, $F(1, 29) = 27.86$, $MSE = 0.02$. As in Experiments 1 and 2, the small difference in false-alarm rates to weak and strong rearranged pairs (.23 vs. .25) was in the direction predicted by the familiarity-based model, but the difference did not approach significance, $F(1, 29) < 1$, $MSE = 0.02$.

Table 5
Proportion of "Yes" Responses to Each Pair Type and Item Type in Experiment 3

Pair and item type	Weak	Strong
Pairs		
Intact	0.58	0.88
Rearranged	0.23	0.25
Pair lure	0.06	
Items		
Item target	0.63	0.91
Item from pair	0.53	0.73
Item lure	0.34	

The confidence ratings to responses made to rearranged pairs are again consistent with the idea that the enhanced familiarity associated with the individual items of a rearranged pair in the strong condition was counteracted by the increased efficiency of retrieval in that condition. Although the false-alarm rates were equal in the weak and strong conditions, participants were again more confident in their correct "no" responses to strong rearranged pairs (2.39 in the strong condition vs. 2.24 in the weak condition), and they were also more confident in their incorrect "yes" responses to strong rearranged pairs (2.15 in the strong condition vs. 1.94 in the weak condition). The effects in this experiment were not significant, $t(29) = 1.87$, $p = .071$, and $t(25) = 1.28$, $p = .215$, for "yes" and "no" responses, respectively, but that may not be surprising, given that only 10 observations were taken from each participant to weak and strong rearranged pairs because of the many conditions involved (whereas 20 and 16 observations were taken from each participant in the corresponding conditions of Experiments 2 and 3, respectively).

Item information again significantly influenced associative-recognition judgments. The false-alarm rate to weak rearranged pairs was higher than the false-alarm rate to pair lures, $F(1, 30) = 60.33$, $MSE = 0.01$. This finding again suggests that presenting items once as part of a pair increases their familiarity above that of nonpresented items and that participants were taking into account item information when making associative-recognition decisions. In addition, pairs that were presented more than once apparently increased item familiarity still further. Specifically, items that were drawn from strong pairs were associated with a much higher hit rate than were items that were drawn from weak pairs, $F(1, 30) = 24.93$, $MSE = 0.02$. As indicated earlier, this higher hit rate might reflect increased familiarity for those items, or it might reflect an enhanced retrieval process (i.e., participants may have responded "yes" to the item if they could retrieve the item with which it was paired during study). Although all-or-none retrieval may also play an important role for the items that appeared individually on the study list, it seems reasonable to assume that if retrieval does play an important role in item recognition, it would be more pronounced for items that originally appeared as part of a pair. The ROC analyses presented below bear on this issue.

ROC analyses. Figure 4 presents the relevant ROC data for the weak and strong conditions, respectively, and Table 6 presents the maximum likelihood parameter estimates and chi-square goodness-of-fit statistics for the two-parameter high-threshold model and for the two-parameter signal-detection model.

The upper panel of Figure 4 shows the ROC results for intact versus rearranged pairs from the weak (left) and strong (right) conditions. These results differ somewhat from those of Experiments 1 and 2. The plot for the weak condition, which was slightly better fit by the threshold model in Experiments 1 and 2, is now better fit by the detection model. However, in agreement with the results of the first two experiments, the ROC is still reasonably well fit by the linear high-threshold account (in contrast to item ROCs, which are generally very poorly fit by linear high-threshold models).

As before, the ROC for the strong condition is undeniably curvilinear. The results shown in Table 4 confirm that the linear high-threshold model provides a rather poor fit in this case and that the detection model again provides an excellent fit. Yet again, an equal-variance model appears to apply. Indeed, the fit of a one-

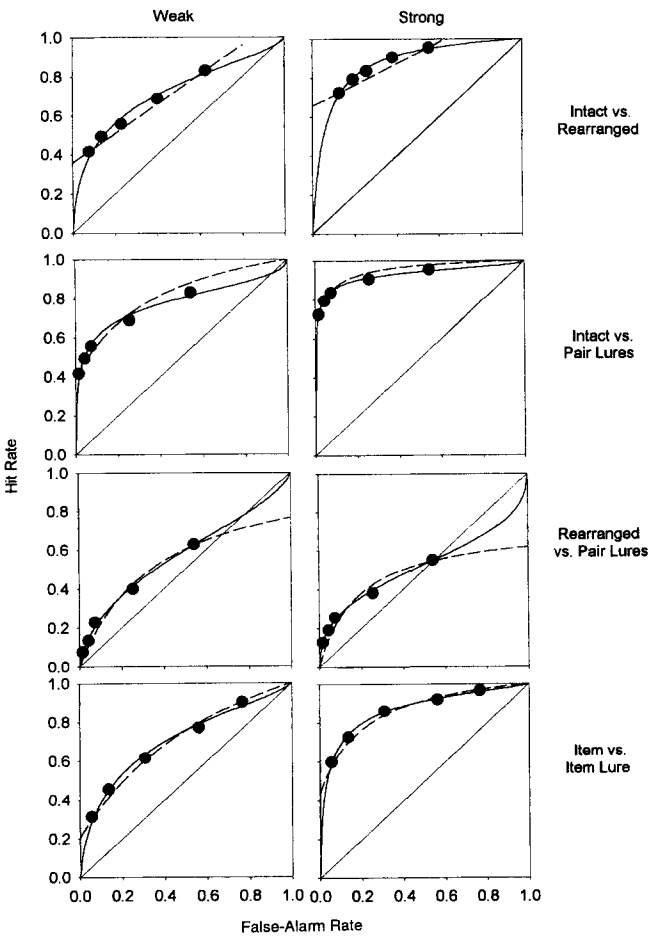

Figure 4. Confidence-based receiver operating characteristic plots for the weak (left panel) and strong (right panel) conditions of Experiment 3. The upper panel plots performance for intact versus rearranged pairs, the second panel from the top plots performance for intact pairs versus pair lures, the third panel plots performance for intact versus rearranged pairs, and the bottom panel plots performance for items-from-pairs versus item lures. In each graph, the dashed line indicates the maximum likelihood fit of the high-threshold model and the solid curve indicates the maximum likelihood fit of the signal-detection model.

parameter detection model (with r fixed at 1) is not significantly improved by allowing r to vary.

In the second panel from the top in Figure 4, weak and strong targets are plotted against pair lures instead of against rearranged pairs. As is evident from Table 6, the high-threshold model again offered a reasonable fit, one that was slightly better than that provided by the detection model in the weak condition and slightly worse than that provided by the detection model in the strong condition. The deviations from the best-fitting threshold model are significant in the strong condition, whereas the deviations from the best-fitting signal-detection model are not significant in either condition. In spite of significant deviations, the high-threshold parameter estimates are again sensible in that the estimates of *Ro* in the weak and strong conditions are similar to those obtained from the corresponding intact versus rearranged ROC analyses. The parameters of the signal-detection analysis suggest that, from that theory's point of view, the ratio of noise-to-signal variance (i.e., r) is approximately 0.5, again suggesting that the standard deviation of the intact pair distribution is about twice that of the pair-lure distribution.

The third panel of ROC plots is similar, except that now the rearranged pair lures represent the "signal" distribution and the pair lures again represent the "noise" distribution. The chi-square goodness-of-fit statistics shown in Table 6 reveal that the detection model provided a better fit than the high-threshold model in both conditions. Deviations from the best-fitting detection model are significant in the weak condition only (whereas deviations from the best-fitting high-threshold model are significant in both conditions). As in Experiment 2, deviations from the threshold model are small but systematic, as shown in Figure 4. The estimates of the high-threshold parameter, *Rn,* (Parameter p1 in Table 6) are quite close to the corresponding estimates that were obtained from the intact versus rearranged ROC analyses (Parameter p2 in those fits). The estimates of d' from the high-threshold fits are also similar for the weak intact versus pair-lure and weak rearranged versus pair-lure fits (as they theoretically should be). For the corresponding strong conditions, however, the d' estimates appear to diverge. With regard to the signal-detection parameter estimates, the best-fitting model again implies that the variance of the rearranged distribution is much greater than that of the pair-lure distribution, especially in the strong condition (where r again approached 0.5).

The lower panel of Figure 4 presents the data for the weak (left) and strong (right) items that were originally presented as part of a pair versus item lure (the ROCs for items presented as items on the list are not shown because the results were so typical). The maximum likelihood parameter estimates for the two types of item ROC fits (i.e., item vs. item lure, and item-from-pair vs. item lure) are presented in Table 6. For the most part, the fits of the two models were good and were comparable with each other, except that deviations from the threshold model were significant in the strong condition for ROCs involving items from pairs. This actually corresponds to what appears to be a trend over the first three experiments. Specifically, the performance of the threshold model deteriorates relative to the detection model as strength increases. This is especially true for the intact versus rearranged ROCs, but it occurs with some regularity for the other ROCs as well.

One should note that, with regard to the high-threshold fits, estimates of the retrieval parameter (p1 in Table 6, which corresponds to *Ro*) were not higher when items were taken from pairs compared with when items appeared on the list as items. In fact, they were lower. One might have expected to see a higher estimate of retrieval for items that had appeared as part of pairs because the associated item provides an extra retrieval opportunity not shared by items that appeared alone on the study list. The fact that the retrieval estimates are in the opposite direction is, from the high-threshold perspective, somewhat surprising.

With regard to the signal-detection fits, the ROCs for items that appeared in pairs on the study list are similar to those based on items that appeared alone on the study list. Indeed, the obtained r values shown in Table 6 are in the .7 to .8 range, just as item ROCs typically are. For all of the item fits, fixing r at 1 yielded a significantly worse fit, which is the expected result, as r is known to be less than 1 for item recognition memory (Ratcliff et al., 1992).

As indicated earlier, we did not report a fit of the linear version of the high-threshold model to the item ROCs in either Experiment 2 or Experiment 3 because that model invariably provides a

Table 6
Maximum Likelihood Parameter Estimates and Chi-Square Goodness-of-Fit Statistics for the Signal-Detection and High-Threshold Models Fit to the ROC Data From Experiment 3

Condition	Model	p1	p2	χ^2	*df*
Intact vs. rearranged[a]					
Weak	Detection	0.794	0.970	2.67	3
	Threshold	0.363	0.222	10.01*	3
Strong	Detection	1.015	1.860	0.57	3
	Threshold	0.651	0.441	23.66*	3
Intact vs. pair lure[b]					
Weak	Detection	0.517	1.465	6.49	3
	Threshold	0.423	0.926	2.79	3
Strong	Detection	0.528	2.571	4.25	3
	Threshold	0.650	1.809	9.70*	3
Rearranged vs. pair lure[b]					
Weak	Detection	0.678	0.394	9.35*	3
	Threshold	0.249	1.005	18.74*	3
Strong	Detection	0.515	0.106	2.89	3
	Threshold	0.426	1.333	9.80*	3
Item target vs. item lure[b]					
Weak	Detection	0.769	0.865	4.02	3
	Threshold	0.226	0.540	3.06	4
Strong	Detection	0.814	1.858	1.76	3
	Threshold	0.388	1.447	1.25	3
Item from pair vs. item lure[b]					
Weak	Detection	0.836	0.604	4.61	3
	Threshold	0.136	0.397	2.64	3
Strong	Detection	0.724	1.088	2.01	3
	Threshold	0.278	0.721	8.40*	3

Note. For all fits of the detection model, p1 = r and p2 = d_e. For the threshold model, p1 = *Ro* and p2 = *Rn* for the intact vs. rearranged fits; p1 = *Ro* and p2 = d' for the intact vs. pair lure and item vs. item lure fits, and p1 = *Rn* and p2 = d' for the rearranged vs. pair lure fits. For 3 degrees of freedom, a chi-square value of 7.81 is significant at the .05 level. ROC = receiver operating characteristic.
[a] $N = 620$. [b] $N = 1550$.
* $p < .05$.

poor fit and is no longer regarded as viable. We did actually perform those item fits, however, and we briefly mention the results here to underscore the point that the weak intact versus rearranged ROCs really are more linear than their item ROC counterparts are. The results already presented in Tables 4 and 6 show that the chi-square values obtained from fitting the linear high-threshold model to the weak intact versus rearranged ROCs in Experiments 2 and 3 were quite low (6.59 and 10.01, respectively). The corresponding values from a fit of the linear high-threshold model to the weak item ROCs from Experiments 1 and 2 were 31.22 and 23.35, respectively. These poor fits occurred even though the overall levels of performance were about the same for the weak associative-recognition and weak item-recognition conditions. Thus, the weak intact versus rearranged ROCs (i.e., the associative-recognition ROCs) are noticeably more linear than those obtained on item recognition tasks even when overall levels of performance are comparable.

Experiment 4

All of the preceding ROC analyses were performed on data that were pooled over participants because too few observations were obtained from each participant to permit their individual ROCs to be analyzed. Whenever data are pooled in this way, a natural question to ask is whether the results are representative of individual participant performance. This question seems especially relevant to the strong condition of each of the preceding experiments because that condition consistently yielded the most unexpected outcome (namely curvilinear intact vs. rearranged ROC). In the previous three experiments, the strong condition was only one of several conditions presented during the course of a single session. In Experiment 4, by contrast, participants were exposed solely to the strong condition for two complete sessions in hopes of collecting enough data from each participant to allow their individual ROCs to be analyzed. Of particular interest was whether the individual ROCs would be better fit by signal-detection theory than high-threshold theory and whether pooling the data over participants would introduce any systematic distortions in the shape of the ROC.

In addition to these individual ROC analyses, we performed simulations to determine what the ROC plot would look like if data were pooled over hypothetical high-threshold participants who differed with respect to their individual retrieval and guessing parameters. Would pooling data over participants known to be responding in accordance with high-threshold theory (each of whom would yield a linear individual ROC) result in a curvilinear group ROC when overall accuracy was high (as it was in the strong conditions of the preceding experiments)?

Method

Participants. The participants were 26 undergraduates from the University of California, San Diego, who met the criterion of English fluency

by age 7. Their participation fulfilled a lower division psychology course requirement.

Design and procedure. Participants were presented with 80 study pairs in each of two sessions (with no overlap between sessions). In each session, the word pairs were presented five times each, for a total of 400 presentations. An additional 112 study presentations were study fillers and were not later tested. Forty of the strengthened study pairs were later tested as strengthened intact pairs, and the other 40 strengthened pairs were tested as strengthened rearranged word pairs. When building the study lists, we filled the first 12 presentations with study fillers. The remaining 500 study presentations were then used for inserting the study pairs. The strengthened study pairs were inserted into the list in sets of two pairs. When inserting each set, we used a randomly selected range of 45 study presentations. The selected range was segmented into five equal segments of nine presentations each. The two study pairs were then inserted into the study list, such that the first presentation of both pairs occurred in the first segment, the second presentation of both pairs occurred in the second segment, and so forth until all five segments were filled. If the word pairs were later to be tested as rearranged word pairs, the members making up the two pairs were switched at test (study A–B, C–D; test A–D, B–C). After inserting the 80 study pairs, we filled the remaining 100 presentations with study fillers, which were not later tested. The test list consisted of 80 test pairs, of which 40 were the strengthened intact pairs and 40 were the strengthened rearranged pairs.

Participants were asked to study each study pair presented and to form an association between the two words. They were informed that the test would consist of intact and rearranged word-pairs. Examples of an intact and rearranged word-pair were given in the instructions. No particular memory strategy was suggested. During the study list presentation, the word pairs were presented for 2 s each, with an interpresentation interval of half a second. During the test session, half of the strengthened word pairs were tested as strengthened intact pairs, and the other half as strengthened rearranged pairs. Each test presentation was presented with the question "Did you previously study these two words together?" Participants were asked to respond by indicating their confidence that the two words were previously studied together. Participants used the same 6-point confidence scale that was used in Experiments 1–3. Following the second test session, participants were then debriefed and provided with a credit slip for their experimental participation.

Results and Discussion

Even after two sessions of exposure to the strong condition, some participants still did not provide a sufficient range of confidence ratings needed to perform an ROC analysis. Of the 26 participants tested, 18 yielded at least three points on the ROC plot, which is the minimum requirement needed to fit the high-threshold and signal-detection models. Of the 8 participants who did not meet the requirement, 1 supplied high-confident "yes" responses to every item or to nearly every item (both intact and rearranged). Whether this participant understood the instructions or not is unclear. The remaining 7 participants yielded only one or two points on the ROC because their accuracy was very high, and most of their responses were correct high-confident "yes" or "no" responses. For the 18 participants who yielded at least three points on the ROC, their data were analyzed in the same manner that the group ROCs were analyzed in the preceding experiments. Before presenting those analyses, we first present an analysis of the pooled ROC to see whether or not the findings from the first three experiments were replicated at that level of analysis.

Pooled ROC analysis. Figure 5 shows the pooled ROC (with data pooled over the 18 participants for whom individual ROCs were available to be fit) along with the best-fitting signal-detection and high-threshold models. When the data were fit by the two competing models using maximum likelihood estimation (as before), neither model fit extremely well, but signal-detection theory did provide a much better fit than high-threshold theory (as is visually apparent). The goodness-of-fit statistics were $\chi^2(3, N = 2880) = 57.80$ and $\chi^2(3, N = 2880) = 328.0$ for the detection and threshold models, respectively, and the maximum likelihood estimates of d_e and r were 1.83 and 0.77, respectively. When the data from the 6 participants who yielded only one or two points on the ROC were included in the pooled data, the results were hardly affected except that the estimate of d_e increased to 2.13. All other parameters and the chi-square goodness-of-fit statistics changed minimally.

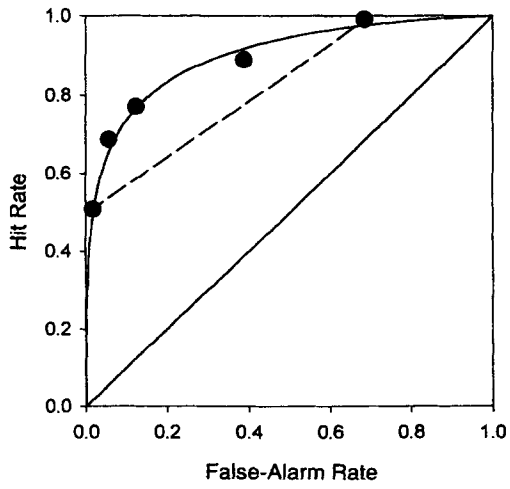

Figure 5. Confidence-based intact-versus-rearranged pooled receiver operating characteristic plot from Experiment 4. The dashed line indicates the maximum likelihood fit of the high-threshold model, and the solid curve indicates the maximum likelihood fit of the signal-detection model.

Given the results of the first three experiments, the fact that signal-detection theory provided the better fit to the pooled ROC in Experiment 4 was not surprising. What was surprising was the fact that the detection model did not provide the nearly perfect fit seen in the strong condition of the previous experiments. Instead, the results are best described as being somewhat mixed, with the clear advantage going to signal-detection theory. Also surprising was the fact that the estimated value of r, which was expected to be close to 1.0 on the basis of the preceding experiments, was closer to .80, the value typically seen in item-recognition experiments. Unlike the preceding experiments, forcing r to assume a value of 1 here resulted in a significantly worse fit, $\chi^2(4, N = 1440) = 91.33$. Why the deviations from the signal-detection model were more significant in Experiment 4 compared with the previous three experiments and why the value of r was less than 1 is not clear. Exposing participants only to the strong condition over the course of two sessions does seem to introduce some differences, possibly strategic in nature, even though the main result (namely, that the detection model provided a better fit than the threshold model) was replicated. The question we address next is whether the individual participant ROC analyses yield the same conclusions as the pooled ROC analysis.

Individual participant ROC analyses. The 18 individual participant ROCs with at least three points were analyzed in the same

manner as the pooled ROCs were, and the results are shown in Table 7. The table shows the chi-square goodness-of-fit statistic for both the high-threshold and signal-detection fits as well as the maximum likelihood estimates of the d_e and r parameters from the detection fits. Note that these chi-square values cannot be used for making statistical inferences (e.g., are the observed deviations greater than would be expected on the basis of chance?) because in most cases the number of cells with more than five expected observations was equal to the number of free parameters (so degrees of freedom were equal to 0). Still, the measure serves to indicate which model provided the better fit.

According to the chi-square goodness-of-fit statistic, the detection model provided the better fit for 14 of the 18 participants, although in some cases the advantage was slight. For 12 of the participants, one model seemed to clearly outperform the other. Of these, the high-threshold model provided a noticeably better fit for 3 participants (the first 3 participants in the table), and the signal-detection model provided a noticeably better fit for 9 participants (the last 9 participants in the table). The two models provided essentially equivalent fits for 6 participants (the middle 6 participants in the table), although the detection model provided a slightly better fit for 5 of these. As might be expected, visual inspection of the individual participant ROCs revealed the same information as the relative chi-square values. Summed over participants, the chi-square for the signal-detection fits (36.68) was much lower than that for the high-threshold fits (123.62). With regard to the signal detection parameter estimates, the average value of d_e was 1.90, and the average value of r was 0.71.

These results suggest that pooling data over participants did not introduce any significant bias in favor of one model over the other or alter conclusions in any significant way. That is, as was true of the pooled ROC analysis, the signal-detection model fit much better than the high-threshold model at the individual level (although the average quality of the signal-detection fit was not especially impressive), and the estimated value of r was less than 1.

Table 7
Chi-Square Goodness-of-Fit Statistics for the High-Threshold and Signal-Detection Models and Maximum Likelihood Signal-Detection Parameter Estimates Based on Fits to Individual ROC Data From Experiment 4

Participant	HTT χ^2	SDT χ^2	r	d_e
15.00	0.23	8.13	0.85	0.88
20.00	1.07	9.83	0.60	1.04
22.00	0.02	0.93	1.56	2.40
2.00	1.89	1.54	0.74	2.23
4.00	0.18	0.07	0.59	2.90
7.00	1.21	0.39	0.63	1.51
9.00	5.19	3.50	0.79	1.23
10.00	0.18	0.20	0.47	3.06
16.00	4.44	2.66	0.72	1.57
1.00	28.64	4.38	0.55	1.82
5.00	12.72	0.19	0.92	2.63
8.00	9.73	0.67	0.77	0.87
14.00	24.26	0.13	0.81	1.89
17.00	11.01	2.95	0.49	2.65
18.00	4.69	0.02	0.53	1.13
19.00	5.14	0.14	0.78	2.51
26.00	12.09	0.96	0.65	1.48
24.00	0.94	0.00	0.44	2.35
Summary	123.63[a]	36.68[a]	0.72[b]	1.90[b]

Note. N = 160 for all chi-square tests. ROC = receiver operating characteristic; HTT = high-threshold; SDT = signal-detection.
[a] Summed value. [b] Mean value.

It might be worth noting that 2 of the 3 participants whose data were better fit by the high-threshold model were among the 3 participants with the lowest overall d_e scores. In fact, the participant with the second lowest d_e of all (Participant 15) and the participant with the third lowest d_e (Participant 20) both yielded the most convincing linear ROCs. In some ways, this result corresponds with what was found at the group level in Experiments 1 through 3, namely, that high-threshold theory is likely to provide the better fit when memory is relatively weak, whereas signal-detection theory is likely to provide the better fit when memory is strong.

High-threshold simulations. Although the ROCs generated by the individual participants in Experiment 4 were generally better fit by a detection model than by a threshold model, we also investigated what the effect of pooling data over participants would be if all participants were known to be high-threshold responders. To determine what the effect would be, we performed simulations in which 18 simulated participants responded to 80 intact and 80 rearranged pairs in accordance with high-threshold theory. Each participant was assigned different retrieval and guessing parameters in an effort to represent the kind of variability that would be present in a real experiment. The beta distribution (described in the Appendix) was chosen to represent error variance because all of the parameters in the high-threshold model are probabilities that range from 0 to 1, and the beta distribution covers that range (unlike, for example, the Gaussian distribution, which ranges from minus infinity to plus infinity). Of particular interest in these simulations was what the pooled ROC would look like when retrieval was fairly high (as it was in the strong condition of all four experiments), such that the ROC data fell mainly in the upper left corner of the plot.

The details of the simulation are presented in the Appendix. The main result was that pooling ROC data over 18 high-threshold responders who differed considerably from each other in terms of *Ro, Rn,* and the five confidence-specific guessing parameters invariably yielded an almost perfectly linear ROC. In a representative run of this simulation, a fit of the high-threshold model to the pooled ROC data yielded a small and nonsignificant chi-square, $\chi^2(3, N = 1440) = 1.84$, whereas a fit of the signal detection model to the same data yielded a highly significant chi-square, $\chi^2(3, N = 1440) = 122.7$. Thus, at least to the extent that these simulations accurately represented error variance in the high-threshold parameters, there appears to be no evidence that averaging artifacts distorts the shape of the ROC if participants respond in accordance with high-threshold theory. The overall conclusion from Experiment 4, then, is that the results of the pooled ROC analyses in the first three experiments are probably representative of the individual participants.

General Discussion

The experiments reported here contribute to the understanding of associative recognition in several ways. First, the results provide fairly compelling evidence that associative-recognition decisions are based on both associative retrieval and item familiarity. When

participants retrieve the associate of one member of the pair, they accept (if intact) or reject (if rearranged) the pair. Further, when retrieval fails and the items of a pair are familiar, participants take that as evidence that the pair was previously studied (even though item familiarity is not diagnostic of this). The joint effects of associative retrieval and item familiarity readily explain why repeating word pairs on a list increases the hit rate for those pairs (because retrieval and familiarity work together) without affecting the false-alarm rate to rearranged word-pairs (because retrieval and familiarity oppose each other).

On the other hand, it does seem rather coincidental that the two opposing forces would be equal in magnitude, and one would like to see false-alarm rates to rearranged pairs change in the expected direction when item or associative information is selectively manipulated. The effects of maintenance rehearsal (Glenberg & Bradley, 1979; Nairne, 1983) and retention interval (Hockley, 1991, 1992) discussed earlier can already be viewed as selectively influencing item information (and false-alarm rates do increase as item information theoretically selectively increases). Also, using a remember–know paradigm, Hockley and Consoli (1999) recently found that associative-recognition "remember" responses were very accurate, but the corresponding "know" responses were essentially at chance, suggesting that "know" responses are based on item familiarity when retrieval fails. If so, "know" responses would yield chance performance, because the items are equally familiar for intact and rearranged pairs. In spite of this converging evidence, our interpretation of the equivalent false-alarm rates observed here in terms of opposing forces would be solidified by future research showing that a selective increase in associative information decreases the false-alarm rate to rearranged pairs.

A second contribution of the present research is that the shape of the ROC changes in a surprising way as a function of strength. Previous research showed that intact versus rearranged ROCs were essentially linear, suggesting a high-threshold process (Yonelinas, 1997; Yonelinas, Kroll, Dobbins, & Soltani, 1999). The intact versus rearranged ROCs reported here were not unambiguously linear in the weak condition (the condition most similar to prior research), but they were certainly more linear than is typical of item-recognition memory ROCs. In two of the experiments reported here, the linear high-threshold model provided a fit that rivaled that of the curvilinear signal-detection model. Summed across the first three experiments, the chi-square goodness-of-fit value for the threshold model (31.02) was comparable with that of the detection model (23.46). Thus, the weak intact-versus-rearranged ROCs are probably best described as falling between the predictions of the two models. The deviations from linearity evident in the data reported by Yonelinas (1997), although not significant, were also usually in the direction consistent with curvilinearity. Thus, although our ROCs were not quite as linear as those observed by Yonelinas, the present results are largely consistent with prior ROC analyses of associative-recognition memory.

The picture changes dramatically when pairs are strengthened. In all four experiments, the detection model provided a much better fit than the threshold model in the strong condition. Summed across the first three experiments, the chi-square goodness-of-fit value for the threshold model (119.00) was much higher than that of the detection model (4.10). Moreover, in these three experiments, an equal-variance detection model was suggested. The value of r (the estimated ratio of the standard deviations of the rearranged and intact distributions) averaged across the first three experiments was .91, and in none of the three experiments was the fit significantly improved by allowing r to vary (compared with fixing it at 1.0). In Experiment 4, which differed procedurally from the first three experiments in that participants were exposed exclusively to the strong condition over the course of two sessions, the value of r was less than 1.0 for reasons that are not clear. Because a weak condition was not included in that experiment, it is impossible to say whether or not the observed value of r was greater than what would have been obtained in the weak condition.

Although the results of Experiment 4 were not completely in line with the other three experiments, they do correspond to our central finding that a nearly linear ROC in the weak condition becomes curvilinear in the strong condition. What could explain such a dramatic transformation of the ROC based merely on strengthening the pairs? A clue is provided by the third contribution of the present research, namely, that associative information may be continuously distributed after all (rather than being all or none) and may be much more variable than item information. The evidence for this suggestion is provided by the intact versus pair-lure ROCs and the rearranged versus pair-lure ROCs. Associative information presumably comes into play for both intact and rearranged pairs. For pair lures, by contrast, associative information must be negligible because the two items of the pair were not previously encountered in the experimental session. Thus, these ROCs consist of one pair involving associative information (in addition to item information) and one pair involving item information only. In general, the detection model offered a better fit than the high-threshold model to the ROCs involving pair lures, and the estimated standard deviation of the associative distribution (intact or rearranged) was generally twice that of the pair-lure distribution. In the next section, we suggest a model that relies on this information to account for our curious pattern of results.

A Some-or-None Model of Associative Recognition

The two models discussed so far assume that associative information is a continuously distributed evidence variable (like the signal-detection model illustrated in Figure 1) or an all-or-none retrieval variable (i.e., the high-threshold model represented by Equations 3 and 4). Neither one of these models by itself can explain the pattern of results we observed. However, a "some-or-none" model like one proposed long ago by McFadden and Greeno (1968) seems to provide a relatively parsimonious account. We propose a specific some-or-none model below and then present the results of simulations that serve to best illustrate how the model behaves.

Assumptions. The assumptions underlying the particular some-or-none model we propose here are as follows:

> *Assumption 1.* Both item information (I) and associative information (A) are continuously distributed random variables that sum to yield a strength-of-evidence variable that is used to decide whether or not a pair was previously encountered. The mean and standard deviation of I are denoted μ_{item} and σ_{item}, respectively, and the mean and standard deviation of A are denoted μ_{assoc} and σ_{assoc}, respectively.
>
> *Assumption 2.* Associative information is much more variable than item information ($\sigma_{assoc} > \sigma_{item}$). This was not something we predicted in advance on the basis of theoretical considerations—the assumption is based on the ROC results involving pair lures. Later, we

offer some thoughts as to why this might be true, but for the moment it is simply an assumption driven by our unexpected results.

Assumption 3. Retrieval failure can occur. That is, like the high-threshold model (Yonelinas, 1997), we assume that for some proportion of intact and rearranged pairs, no associative information is retrieved (leaving, we assume, only item information on which to base an associative-recognition decision). Unlike the high-threshold model, we assume that when retrieval does not fail, associative information can be retrieved to varying degrees. As in the equations for the high-threshold model, the probability of retrieving associative information is denoted *Ro* and *Rn* for intact and rearranged pairs, respectively.

More formally, the hit and false-alarm rate equations for the new model proposed here are:

$$H = Ro \times [p(I + A) > c] + (1 - Ro) \times [p(I) > c] \text{ and} \quad (8)$$

$$FA = Rn \times [p(I - A) > c] + (1 - Rn) \times [p(I) > c]. \quad (9)$$

Equation 8 states that with probability *Ro,* the decision for an intact pair is based on the sum of item and associative information. If that sum exceeds a decision criterion, the participant responds "yes," otherwise the response is "no." With probability $1 - Ro$, no associative information is retrieved, and the decision is based on item information only. Thus, in this case, the participant responds "yes" if item familiarity alone exceeds the decision criterion. Equation 9 states that with probability *Rn,* the decision for a rearranged pair is based on the difference between item and associative information. Only if that difference exceeds a decision criterion does the participant mistakenly respond "yes." With probability $1 - Rn$, the decision is based on item information only, such that the participant responds "yes" if item familiarity alone exceeds the decision criterion.

The basic structure of these two equations is consistent with the hit rate and false-alarm rate to intact and rearranged pairs in the weak and strong conditions across the first three experiments. That is, in the strong condition, *I* increases, and so do *Ro* and *Rn.* Thus, the hit rate should increase considerably, but the false-alarm rate should be much less affected. Similar ideas were represented in the context of a high-threshold model described earlier in Equations 6 and 7. In that sense, the new model offered here is not a radical departure from the high-threshold model proposed by Yonelinas (1997).

The mean strength-of-evidence values for intact and rearranged pairs, according to the new model, are

$$\mu_{intact} = \mu_{item} + Ro \times \mu_{assoc}$$

$$\mu_{rearr} = \mu_{item} - Rn \times \mu_{assoc}.$$

This model is a lot like the one proposed by Hockley (1992), except that (a) the variance of the associative-information distribution is assumed to be large, (b) associative information is not always added to item information for intact pairs, and (c) associative information is sometimes subtracted from item information for rearranged pairs.

Adding or subtracting random variables like *I* and *A* increases the variance of the resulting value beyond that of the two constituents. If *A* is added to *I* with a probability of *Ro* for intact pairs and *A* is subtracted from *I* with probability *Rn* for rearranged pairs, then the variances of the resulting intact and rearranged distributions (σ^2_{intact} and σ^2_{rearr}, respectively) would be

$$\sigma^2_{intact} = \sigma^2_{item} + Ro \times \sigma^2_{assoc} \quad \text{and}$$

$$\sigma^2_{rearr} = \sigma^2_{item} + Rn \times \sigma^2_{assoc}.$$

A key assumption of our account is that *Ro* and *Rn* are less than 1.0 in the weak condition (i.e., retrieval failure can occur in that condition) but that they both approach 1.0 in the strong condition (i.e., some associative information may be retrieved for all strong intact and rearranged pairs). Thus, in the strong condition,

$$\sigma^2_{intact} = \sigma^2_{item} + \sigma^2_{assoc} \quad \text{and}$$

$$\sigma^2_{rearr} = \sigma^2_{item} + \sigma^2_{assoc},$$

which is to say that an equal variance ROC should result (i.e., $\sigma^2_{intact} = \sigma^2_{rearr}$). These equations assume that item and associative information are uncorrelated. If they were instead positively correlated (i.e., if the more familiar items were also the ones most likely to occasion successful retrieval), then the variance of the rearranged distribution would be much less than that of the intact distribution,[4] which would be contrary to the equal-variance intact-versus-rearranged ROCs observed in the strong conditions of Experiments 1 through 3. If this assumption turns out to be incorrect then, obviously, the model would need to be revised.

In the weak condition, *Ro* and *Rn* are theoretically less than 1.0. Thus, in this case, the intact distribution is produced by one random variable (*I*) that is only sometimes added to another random variable with much greater variance (*A*). Similarly, the rearranged distribution is produced by one random variable (*I*) from which another random variable with much greater variance (*A*) is only sometimes subtracted. The resulting distributions of evidence values for intact and rearranged pairs would not be Gaussian in form even if *I* and *A* are themselves normally distributed, and the resulting implications for an empirical intact-versus-rearranged ROC are not easy to grasp intuitively. As we show next by means of simulation, if *A* is added to or subtracted from *I* on only some of the trials (Assumption 3), the expected ROC can be nearly linear in form even if *A* is a continuously distributed random variable.

Note that this model also generally predicts that ROC analyses for a weak condition involving pair lures (weak intact vs. pair lure, or weak rearranged vs. pair lure) will be curvilinear but perhaps not perfectly fit by the signal-detection model because the strength-of-evidence distribution (intact or rearranged) will not be Gaussian. For example, the strength of evidence for weak intact pairs consists of item information that is sometimes added to associative information (a non-Gaussian distribution), whereas the strength of evidence for pair lures consists of item information only (which is theoretically Gaussian in form). Thus, the weak intact versus pair-lure ROC pits one non-Gaussian distribution against another Gaussian distribution, so the fit of the detection model should not be perfect. In the strong condition, by contrast, the strength-of-evidence distribution for intact pairs becomes

[4] The reason is that the variance of the sum of two correlated random variables like *I* and *A* is equal to $\sigma^2_{item} + \sigma^2_{assoc} + 2\rho(\sigma_{item}\,\sigma_{assoc})$, where ρ is the correlation between the two variables. By contrast, the variance of the difference between two correlated random variables (i.e., $I - A$) is equal to $\sigma^2_{item} + \sigma^2_{assoc} - 2\rho(\sigma_{item}\,\sigma_{assoc})$. Only when ρ equals 0 are the variances the same whether the random variables are added or subtracted.

Gaussian, so the fit of the detection model in that case should improve. A tendency for the detection model to perform particularly well in the strong conditions was evident in all four experiments.

Simulations. The purpose of the simulations was (first and foremost) to determine what the impact of a some-or-none process would be on the shape of the ROC. Would it ever yield a linear ROC? A second purpose was to show that a set of parameters could be found that faithfully reproduced the observed array of findings. That would not prove that the some-or-none model is correct (especially because the model was developed to explain these findings), but failure to account for any important quantitative trends would be grounds for doubting the model's validity.

The simulations simply involved generating item information and associative information values from different normal distributions and summing them to yield strength-of-evidence values in accordance with Equations 8 and 9. The parameters of the relevant distributions in the simulation were selected on the basis of trial and error, but, as is probably apparent from the values chosen, an exhaustive search was not performed. The mean of the item distribution was set to 1.0 for a pair of new items, to 2.0 for a pair of items seen once (weak), and to 4.0 for a pair of items seen six times (strong). For the sake of simplicity, the item information distribution was always assumed to have a standard deviation of 1.0, regardless of its mean.

The mean of the associative-information distribution (μ_{assoc}) was set to 3.0 for both the weak and strong conditions, and its standard deviation (σ_{assoc}) was set to 3.0 in the weak condition and to 2.0 in the strong condition. That is, associative information was assumed to be much more variable than item information, especially in the weak condition. If we did not assume that associative information was more variable than item information, then it would not be possible to account for the remarkably low values of r observed in the ROCs involving pair lures.

Holding the mean value of associative information (i.e., μ_{assoc}) constant as a function of strength while decreasing the standard deviation may seem somewhat odd. It might seem more natural to increase μ_{assoc} while leaving σ_{assoc} fixed (as we did for items). However, it should be noted that decreasing σ_{assoc} increases the associative d' as effectively as increasing μ_{assoc} would. To appreciate this point, assume for the moment that all decisions were based on associative information without item information being taken into account in any way (i.e., $I = 0$ in this hypothetical scenario, and *Ro* and *Rn* both equal 1.0). Under such conditions, d' would be equal to the mean associative strength for intact pairs (i.e., $I + A = 3.0$) minus the mean associative strength for rearranged pairs (i.e., $I - A = -3.0$) divided by the standard deviation of the associative distributions (3.0). That is, d' for the weak condition would be 6.0/3.0, or 2.0. For the strong condition, the means are assumed to be the same, but σ_{assoc} decreases to 2.0 such that d' would increase to 6.0/2.0, or 3.0. Although associative d' increases in the strong condition either by increasing the mean or by decreasing the standard deviation, we found that the latter option provided a better fit to the data when item and associative were combined.

In the simulations, the values of *Ro* and *Rn* were both set to 1.0 in the strong condition and to 0.60 and 0.40, respectively, in the weak condition. The values could have both been set to, say, 0.50 in the weak condition without changing anything essential, but the nearly linear ROCs that are observed in the weak condition have a slope less than 1.0 (e.g., Yonelinas, 1997). That is usually explained by assuming that the probability of retrieval for an intact pair exceeds that of a rearranged pair. The same must be assumed here to yield an ROC with the appropriate slope. Table 8 presents a summary of the parameter values used in the simulations.

To produce ROC data, we ran 2,000 trials for each of several settings of c, the location of the decision criterion (the different locations reflect different levels of confidence). On each trial, a value was randomly selected from the item distribution, then a value was randomly selected from the associative-information distribution. On a random *Ro* proportion of the trials, the item and associative-information values were summed to yield a strength-of-evidence value. On the remaining $1 - Ro$ proportion of the trials, strength of evidence was determined by the item information only. If the evidence value exceeded the preset value of c, the hit rate counter was incremented (otherwise the counter did not increment). A similar sequence of events then unfolded to yield an evidence value for rearranged pairs. That is, a value was randomly selected from the item distribution, and another value was randomly selected from the associative-information distribution. On a random *Rn* proportion of the trials, the associative information value was subtracted from the item information value to yield a strength value. On the remaining $1 - Rn$ proportion of the trials, strength was determined by the item information only. If the evidence value exceeded the preset value of c, the false-alarm rate counter was incremented (otherwise the counter did not increment). The 2,000 trials yielded a hit rate and a false-alarm rate for that value of c (i.e., for a particular criterion setting). The value of c was then changed and the process was repeated to yield another pair of hit rates and false-alarm rates. A full ROC was constructed by conducting simulations with c set to 0.90, 1.55, 2.20, 2.85, and 3.50 (these were chosen because they yielded confidence-specific hit and false-alarm rates roughly comparable with what was observed in the data). The ROCs were then fit by maximum likelihood estimation.

The predicted hit rates and false-alarm rates, which are shown in Table 9, were those associated with a c of 2.85 simply because those values were in the range observed in Experiments 1 through 3. It is clear that the hit rates changed considerably as a function of strength, but the false-alarm rates remained nearly constant. The maximum likelihood parameter estimates and goodness-of-fit statistics for both the detection and threshold models are shown in Table 10. Of most interest is the fact that the model yields a linear intact-versus-rearranged ROC with a slope

Table 8
Parameter Settings Used to Simulate the Some-or-None Model

Parameter	Condition		
	Lure	Weak	Strong
Item information			
μ_{item}	1.0	2.0	4.0
σ_{item}	1.0	1.0	1.0
Associative information			
μ_{Assoc}	0.0	3.0	3.0
σ_{Assoc}	0.0	3.0	2.0
Retrieval probabilities			
Ro	0.0	0.6	1.0
Rn	0.0	0.4	1.0

Table 9
Proportion of Simulated "Yes" Responses to Each Pair Type

Pair and item type	Weak	Strong
Intact	0.52	0.97
Rearranged	0.17	0.21
Pair lure	0.04	

less than 1 in the weak condition (one that is better fit by the threshold model than the detection model) and a curvilinear intact-versus-rearranged ROC that is symmetric about the main diagonal in the strong condition (one that is better fit by the detection model than the threshold model). Although several parameters were changed as a function of strength to reproduce the full range of data we observed, the transformation in the shape of the ROC from linear to curvilinear occurs because of the increase in *Ro* and *Rn* to 1.0.

The fact that occasionally combining variable associative information with item information yields a nearly linear ROC is the most important revelation provided by these simulations. Note that deviations from this nearly linear ROC will always be in the direction of curvilinearity (as is almost always true in real data). A less pronounced advantage for the high-threshold model in the weak condition (which is what we found empirically) could have been easily arranged by increasing the values of *Ro* and *Rn* for that condition somewhat and adjusting the other parameters accordingly.

The other quantitative details in Table 10 are also consistent with observed trends. For example, the ROCs involving pair lures are generally better fit by the detection model than by the threshold model, especially in the strong conditions. Also, the *r* parameter for the intact versus pair-lure ROCs is close to 0.50 (as it should be), although it does decrease moderately as a function of strength, contrary to what we actually observed in Experiments 2 and 3. The *r* parameter for the rearranged versus pair-lure ROCs is also in the appropriate range, and it decreases substantially as a function of strength, which was true of fits to actual data.

It should be pointed out that with the parameter settings shown in Table 8, item and associative-information values were not always positive values. Thus, for example, the associative distribution in the weak condition had a mean of 3.0 and a standard deviation of 3.0, so some of the associative values were actually negative (namely, the 16% of values that were less than one standard deviation below the mean). A negative value that might occasionally be generated by an intact pair means that, had the decision been made exclusively on the basis of associative information, that pair would have mistakenly been judged to be a rearranged pair. That is, if I is set to zero (for purposes of illustration), then $\mu_{intact} = 3.0$ and $\mu_{rearr} = -3.0$. Assuming unbiased responding (i.e., assuming that the decision criterion was placed at 0), a value greater than 0 would be taken as evidence that the pair was intact, whereas values less than 0 would be taken as evidence that the pair was rearranged. Under these conditions, even intact pairs would sometimes generate negative values that would be taken as evidence that the pair was rearranged (and, of course, rearranged pairs would sometimes generate positive values that would be taken as evidence that the pair was intact). In the simulations, associative evidence values were added to item information in a way that preserved this diagnosticity (i.e., some values from the intact distribution were negative, and some from the rearranged distribution were positive).

A question about the some-or-none model that has yet to be asked is why associative information might be more variable than item information. Although we have no formal account of this aspect of the model, the increased variability does not seem surprising in retrospect if associative information is indeed a continuously distributed variable. First, if the two types of information are different (as item familiarity and associative retrieval presumably are), it would seem highly coincidental if they turned out to

Table 10
Maximum Likelihood Parameter Estimates and Chi-Square Goodness-of-Fit Statistics for the Signal-Detection and High-Threshold Models Fit to the Simulated ROC Data

Condition	Model	p1	p2	χ^2	*df*
Intact vs. rearranged					
Weak	Detection	0.795	0.968	13.31*	3
	Threshold	0.379	0.258	5.81	3
Strong	Detection	1.012	2.657	2.12	3
	Threshold	0.939	0.483	21.37*	3
Intact vs. pair lure					
Weak	Detection	0.553	1.443	7.94*	3
	Threshold	0.424	0.875	14.98*	3
Strong	Detection	0.430	3.689	5.78	3
	Threshold	0.927	1.797	20.92*	3
Rearranged vs. pair lure					
Weak	Detection	0.665	0.362	7.47	3
	Threshold	0.308	1.141	3.82	3
Strong	Detection	0.470	−0.001	1.73	3
	Threshold	0.465	1.381	27.69*	3

Note. For all fits of the detection model, p1 = *r* and p2 = d_e. For the threshold model, p1 = *Ro* and p2 = *Rn* for the intact vs. rearranged fits; p1 = *Ro* and p2 = d' for the intact vs. pair lure fits, and p1 = *Rn* and p2 = d' for the rearranged vs. pair lure fits. For 3 degrees of freedom, a chi-square value of 7.81 is significant at the .05 level. $N = 20{,}000$ for all chi-square tests. ROC = receiver operating characteristic.
* $p < .05$.

have the same variance. Thus, the fact that they differ in this respect should not be surprising. Second, low item familiarity and a small amount of associative information may both contribute very little to the strength of evidence that a pair was seen before. Thus, at the low end of the scale, both kinds of information may be more or less equivalent. At the high end of the scale, by contrast, item familiarity is probably vastly exceeded by associative information. That is, no matter how familiar the two items of a pair happen to be, it would still provide much less evidence that the pair was previously seen than would a clear recollection of that fact. According to this notion, whereas item familiarity ranges from low to high, associative information ranges from low to much higher than that (i.e., associative information is more variable than item information is).

Implications for Other Models

The present results are not easily reconciled with a standard high-threshold account of associative-recognition memory. That account predicts a linear ROC for intact versus rearranged pairs regardless of strength. The results reported here suggest that deviations from the high-threshold model increase (and deviations from the signal-detection model decrease) as strength increases. Although these results seem to suggest a role for signal-detection theory in accounting for associative-recognition memory, it is also possible that other versions of the high-threshold account could accommodate our findings. For example, a double high-threshold account (Macmillan & Creelman, 1991), or an account that allows for gradations of confidence for above-threshold evidence values, may be able to account for the present results. The simplest high-threshold account, though, may be unsustainable.

Finally, the present results add to prior research suggesting that global matching models, which typically do not include a specific role for retrieval in associative recognition, may need to do so (cf. Clark & Gronlund, 1996). The results also suggest that, in addition to including a role for retrieval, models like SAM (Gillund & Shiffrin, 1984) and MINERVA 2 (Hintzman, 1984) may need to modify their fundamental assumptions about how associative recognition works. In both models, item and associative information are inseparable. Therefore, if item familiarity is contributing to the decision-making process, associative familiarity with the same variance should be contributing as well. If that were the case, though, ROC plots should never be linear, and an extremely unequal variance detection model would not be expected to apply to intact versus pair-lure ROCs. Thus, these models need to specify how item familiarity can contribute independently of associative familiarity and why associative information is so variable. One model that can already do that is TODAM (Murdock, 1982, 1997). As noted by Clark et al. (1993), TODAM allows for associative information to be represented separately from item information because the two items of a pair can be convolved and then stored with the other items and pairs in a common, distributed, memory representation. A recent revision of TODAM (Murdock, 1997) also accommodates separate effects on item information versus associative information, such as the differential forgetting rates they engender. Conceivably, that model could be brought to bear on the current findings as well. For the moment, the simple (but admittedly incomplete) some-or-none model proposed here provides a reasonably parsimonious account of many associative-recognition phenomena.

References

Abramowitz, M., & Stegun, I. A. (1970). *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (9th printing). New York: Dover.

Bradley, M. M., & Glenberg, A. M. (1983). Strengthening associations: Duration, attention, or relations? *Journal of Verbal Learning and Verbal Behavior, 22,* 650–666.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60.

Clark, S. E., & Hori, A. (1995). List length and overlap effects in forced-choice associative recognition. *Memory & Cognition, 23,* 456–461.

Clark, S. E., Hori, A., & Callan, D. E. (1993). Forced-choice associative recognition: Implications for global-memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 871–881.

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

Glenberg, A. M., & Bradley, M. M. (1979). Mental contiguity. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 88–97.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments and Computers, 16,* 96–101.

Hockley, W. E. (1991). Recognition memory for item and associative information: A comparison of forgetting rates. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 227–248). Hillsdale, NJ: Erlbaum.

Hockley, W. E. (1992). Item versus associative information: Further comparisons of forgetting rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1321–1330.

Hockley, W. E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory & Cognition, 27,* 657–664.

Humphreys, M. S. (1978). Item and relational information: A case for context independent retrieval. *Journal of Verbal Learning and Verbal Behavior, 17,* 175–187.

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

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

McFadden, D., & Greeno, J. G. (1968). Evidence of different degrees of learning based on different tests of retention. *Journal of Verbal Learning & Verbal Behavior, 7,* 452–457.

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

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104,* 839–862.

Nairne, J. S. (1983). Associative processing during rote rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 3–20.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

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

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

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory & Language, 43,* 67–88.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1397–1410.

Westerman, D. L. (2001). The role of familiarity in item recognition, associative recognition, and plurality recognition on self-paced and speeded tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 723–732.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25,* 1397–1410.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I. G., & Soltani, M. (1999). Recognition memory of faces: When familiarity supports associative recognition judgments. *Psychonomic Bulletin & Review, 6,* 654–661.

Appendix

High-Threshold Simulations

Each simulated participant was assigned a unique value of *Ro, Rn,* and five guessing parameters (*g1* through *g5*). All seven parameters were selected by randomly drawing from a beta distribution with the appropriate mean (see below) and a reasonably large standard deviation. The beta distribution was selected because all of the parameters in the high-threshold model are probabilities that range from 0 to 1, and the beta distribution covers that range. This distribution also possesses a shape that seems likely to correspond, at least approximately, to the true error distribution for probabilities. Specifically, when the mean is close to 0, the distribution is skewed to the right (with many values compressed against the floor). When the mean is close to 1, the distribution is skewed to the left (with many values compressed against the ceiling). When the mean is close to .50, the distribution is bell shaped.

The beta distribution is defined by two parameters, a and b, and its range is $0 < p < 1$, the appropriate interval for our purposes. For integer values of a and b, the beta distribution is given by $[(a + b)!/(a!b!)](1 - p)^{b-1}(p)^{a-1}$ and has a mean of $a/(a + b)$ and a variance of $ab/[(a + b)^2 (a + b + 1)]$. For our simulations, we fixed the sum of a and b at 10 for the two retrieval parameters because that generated what seemed to us to be sufficient error variance.

For each participant, a value of *Ro* was selected by drawing from a beta distribution with a mean and standard deviation of .70 and .138, respectively (involving beta parameters of $a = 7$, $b = 3$), and a value of *Rn* was selected by drawing from a beta distribution with a mean and standard deviation of .40 and .148, respectively (involving beta parameters of $a = 4$, $b = 6$). These mean values of *Ro* and *Rn* were chosen because they are close to the maximum likelihood estimates of *Ro* and *Rn* in the strong conditions of Experiments 1 through 3. Five guessing parameters were then selected the following way: 20 random numbers between 0 and 1 were generated from a uniform distribution and then arranged in ascending order. The 1st, 2nd, 4th, 8th, and 16th of these numbers constituted the five guessing parameters for that participant. This method ensured that the high-confident "yes" guessing parameter would be associated with the lowest guessing rate, the medium-confident "yes" guessing parameter would be associated with the next lowest guessing rate, and so on. Using this method, we drew each guessing rate from a unique beta distribution. The high-confident guessing parameter, for example, would be drawn from a beta distribution with parameters of $a = 1$ and $b = 19$. The mean guessing rates for the varying degrees of confidence would be .05, .10, .20, .40, and .80, and their respective standard deviations would be .048, .065, .087, .107, and .087. These mean values for the guessing parameters were selected because they were similar to the maximum likelihood estimates of the guessing parameters when the high-threshold model was fit to the ROC data from the strong conditions of Experiment 1 through 3. The standard deviations for these parameters and for the retrieval parameters were not selected in any principled way. Instead, they were the standard deviations that emerged when the beta distribution was programmed in a convenient way. The values were large enough to ensure considerable variability in performance across simulated participants.

After selecting the parameters for an individual participant, we began simulated recognition. For an intact trial, a random number between 0 and 1 (*u1*) was generated from a uniform distribution. If *u1* was less than *Ro,* the response was scored as a high-confident hit. If *u1* was greater than *Ro* (which means that retrieval failed), a second random number was generated (*u2*). If *u2* was less than *g1,* then the response was scored as a high-confident hit (although it was a guess). If *u2* was not less than *g1* but was less than *g2,* then the response was scored as a medium-confident hit. If *u2* was not less than *g1* or *g2* but was less than *g3,* then the response was scored as a low-confident hit. If *u2* was not less than *g1, g2,* or *g3* but was less than *g4,* then the response was scored as a low-confident miss. If *u2* was less than *g5* (but was not less than any of the other guessing parameters), then the response was scored as a medium-confident miss. Finally, if *u2* was not less than any of the five guessing parameters, then the response was scored as a high-confident miss.

The procedure for simulating a rearranged trial was similar. First, a random number between 0 and 1 (*u1*) was generated from a uniform distribution. If *u1* was less than *Rn,* the response was scored as a high-confident correct rejection. If *u1* was greater than *Ro* (which means that retrieval failed), a second random number was generated (*u2*). If *u2* was less than *g1,* then the response was scored as a high-confident false alarm. If *u2* was not less than *g1* but was less than *g2,* then the response was scored as a medium-confident false alarm. If *u2* was not less than *g1* or *g2* but was less than *g3,* then the response was scored as a low-confident false alarm. If *u2* was not less than *g1, g2,* or *g3* but was less than *g4,* then the response was scored as a low-confident correct rejection. If *u2* was less than *g5* (but was not less than any of the other guessing parameters), then the response was scored as a medium-confident correct rejection. Finally, if *u2* was not less than any of the five guessing parameters, then the response was scored as a high-confident correct rejection.

A single run of the simulation yielded 1,440 observations from intact trials (involving varying degrees of confidence) and 1,440 observations from rearranged trials (also involving varying degrees of confidence). These data were then used to construct an ROC, which was then analyzed by fitting both the high-threshold model and the signal-detection model by means of maximum likelihood estimation. The simulation was actually run 10 times, and all 10 ROCs were analyzed. For the signal-detection fits, the chi-square goodness-of-fit values (with 3 degrees of freedom) ranged from a low of 113.2 to a high of 180.1, which is to say that none of the fits were very accurate. For the high-threshold fits, the chi-square goodness-of-fit values (also with 3 degrees of freedom) ranged from a low of 1.37 to a high of 11.6, which is to say that all of the fits were quite accurate, and only occasionally did the simulated data deviate from the best fitting model to a significant degree. In every case, visual inspection of the ROC revealed what appeared to be a nearly perfectly linear ROC. Thus, pooling data over participants known to be responding in accordance with high-threshold theory does not distort the shape of the ROC even when overall performance is relatively high.

Received August 10, 1999
Revision received June 1, 2000
Accepted August 4, 2000 ■