Journal of Experimental Psychology:
Learning, Memory, and Cognition
2002, Vol. 28, No. 6, 1095–1110

Copyright 2002 by the American Psychological Association, Inc.
0278-7393/02/$5.00 DOI: 10.1037//0278-7393.28.6.1095

On the Nature of the Decision Axis in Signal-Detection-Based Models of Recognition Memory

Holly E. R. Morrell, Santino Gaitan, and John T. Wixted
University of California, San Diego

Most models of recognition memory involve a signal-detection component in which a criterion is placed along a decision axis. Older models generally assume a familiarity-decision axis, but newer models often assume a likelihood ratio axis instead because it allows for a more natural account of the ubiquitous mirror effect. In 3 experiments reported here, item strength was differentially manipulated to see whether a mirror effect would occur. Within a list, the items from 1 category were strengthened by repetition, but the items from another category were not. On the subsequent recognition test, the hit rate was higher for the strong category, but the false-alarm rates for the weak and strong categories were the same (i.e., no mirror effect was observed). This result suggests that the decision axis represents a familiarity scale and that participants adopt a single decision criterion that they maintain throughout the recognition test.

Nearly every model of recognition memory that has been advanced over the past 40 years has included some role for signal-detection theory. In its simplest form, signal-detection theory holds that a decision about whether an item was recently encountered on a list depends on its level of familiarity. If the item's level of familiarity exceeds a criterion value, then it is judged to be old; otherwise it is judged to be new. As illustrated in Figure 1A familiarity values associated with the old and new items (i.e., the targets and lures, respectively) are often assumed to be normally distributed, with the mean of the target distribution located at a higher point on the familiarity axis than the mean of the lure distribution. The decision criterion, which specifies the familiarity value above which an item is declared to be old, can be placed anywhere along the familiarity axis. In Figure 1, it is placed exactly halfway between the means of the target and lure distributions, which is the placement that yields unbiased responding and that maximizes the proportion of correct responses.

The shaded area in Figure 1 represents the proportion of lures that are incorrectly judged to be old because they are associated with familiarity values that fall above the criterion. That proportion is known as the false-alarm rate, a variable that will be the main focus of much of this article. Because the decision criterion is placed midway between the target and lure distributions, the false-alarm rate for the situation depicted in Figure 1 would be about 16% (whereas the hit rate would be 84%). Such responding is said to be unbiased because responses of "old" and "new" are given equally often. A subject with a liberal response bias would have the criterion placed farther to the left, in which case both the hit rate and the false-alarm rate would be higher (i.e., the subject would say "old" more often than "new").

An important assumption of the detection theory just described is that the decision axis represents a strength-of-evidence variable, such as familiarity. Although many signal-detection models of recognition memory make precisely that assumption, another class of detection models does not. These models assume that the decision axis represents a log likelihood-ratio scale. According to a likelihood-ratio model, recognition decisions are based on a statistical computation: If the computed odds that the item appeared on the list are high enough (usually greater than even), then the item is declared to be old; otherwise it is declared to be new. Figure 1B illustrates the likelihood-ratio model, and it obviously looks a lot like the familiarity model shown in Figure 1A. The only real difference is the decision axis, which now represents a log-likelihood scale.

To see how the familiarity and log likelihood-ratio scales are related to each other, consider a test item that generates a familiarity value that happens to fall exactly at the mean of the target distribution in Figure 1A. We might think of this item as generating a moderately high level of familiarity, and an unbiased subject would certainly declare such an item to be old. To do so, the subject need not know anything about the shapes of the underlying distributions. The fact that the item's level of familiarity exceeds the criterion is sufficient to arrive at a decision.

Likelihood-ratio models assume that the operative psychological variable is an odds ratio associated with the test item, not its level of familiarity. The odds ratio is equal to the likelihood that the item was drawn from the target distribution divided by the likelihood that it was drawn from the lure distribution. Graphically, this value is given by the height of the target distribution divided by the height of the lure distribution (h_L) at the point on the familiarity axis where the item falls. The height of the target distribution at its mean is 7.38 times the height of the lure distribution. As such, a test item that generates this level of familiarity (i.e., a familiarity of μ_{Target}) is 7.38 times as likely to have been drawn from the target distribution as the lure distribution. Whenever the odds are greater than even, as they are in this case, an unbiased subject would declare the item to be old. Note that the log of 7.38 is 2.0, so the same item that generates a level of familiarity

Holly E. R. Morrell, Santino Gaitan, and John T. Wixted, Department of Psychology, University of California, San Diego.

We thank Robert Kelly for his support and ideas and Chara Matsuo and Damon Comp for their invaluable assistance.

Correspondence concerning this article should be addressed to John T. Wixted, Department of Psychology, University of California, San Diego, La Jolla, California 92093–0109. Email: jwixted@ucsd.edu

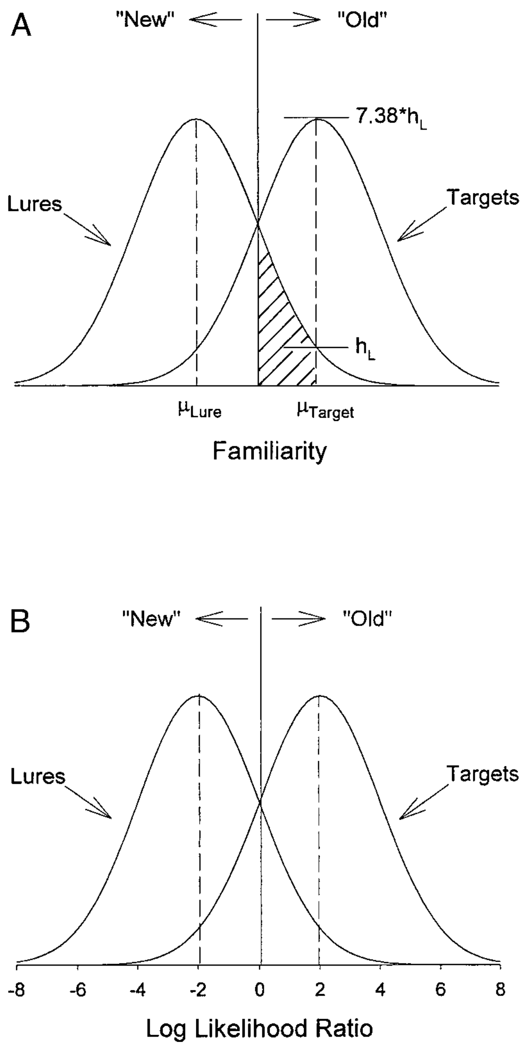

Figure 1. A: Standard familiarity-based, signal-detection model of recognition memory. The decision axis represents a familiarity scale that ranges from low to high. B: Likelihood-ratio version of the signal-detection account of recognition memory. The decision axis represents a log likelihood-ratio scale that ranges from minus infinity to plus infinity.

falling at the mean of the target distribution in Figure 1A generates a log likelihood ratio of 2.0 in Figure 1B. Similarly, an item whose familiarity value falls midway between the means of the target and lure distributions is associated with a likelihood ratio of 1.0 (i.e., the heights of the two distributions are equal at that point), which translates to a log likelihood ratio of 0 in Figure 1B. Obviously, likelihood ratio models assume that the memory system has knowledge of, and can perform, computations on the underlying distributions.

The illustrations shown in Figure 1 are idealized, equal-variance detection models, but prior analyses of recognition memory receiver operating characteristic data suggest that the standard deviation of the target distribution is actually about 1.25 times that of the lure distribution (Ratcliff, Sheu, & Gronlund, 1992). This fact introduces no particular complications for the familiarity version of the model, but it can introduce complications for likelihood-ratio models. The main complication is that when the target and lure distributions are of unequal variance, there may not be a unique familiarity value associated with each likelihood ratio. For example, Gaussian target and lure distributions actually intersect twice if their standard deviations differ (instead of intersecting just once, as shown in Figure 1), so two different familiarity values would correspond to a log likelihood ratio of 0. The likelihood-ratio models considered later in this article do not encounter this problem because they involve various non-Gaussian distributions that have unique likelihood-ratio solutions for every familiarity value, even though the target distribution has greater variance than the lure distribution. However, the basic properties of these models correspond closely to those of the equal-variance Gaussian model, so we used that simple model here to illustrate the properties of likelihood-ratio models in general.

The familiarity-and-likelihood ratio accounts are equally able to explain, say, a hit rate of .84 and a false-alarm rate of .16 (as shown in Figure 1), but Figure 2 illustrates one important difference between the two. Figure 2 shows what the expected result would be if participants maintained the same decision criterion as d' (the standardized distance between the means of the target and lure distributions) changed from low to high. Figure 2A shows the familiarity-based model, and Figure 2B shows the likelihood-ratio model. Note that the definition of the decision criterion differs between the two accounts. In the strength version, the criterion is a particular level of familiarity (such that items yielding higher familiarity levels than that are judged to be old). In the likelihood-ratio version, the criterion is a particular odds ratio (such that items yielding higher odds than that are judged to be old).

In the familiarity-based model shown in Figure 2A, the criterion is placed 0.75 standard deviations above the mean of the lure distribution, whether the targets are weak or strong (i.e., the criterion is fixed at a particular point on the familiarity axis). As such, the hit rate increases considerably as a function of strength, but the false-alarm rate remains constant. If the familiarity of the lures does not change as the targets are strengthened (which is the simplest assumption), then the only way that the false-alarm rate would change is if the criterion moved across conditions.

In the likelihood-ratio model, the criterion is placed at the point where the odds that the item was drawn from the target or lure distributions are even (i.e., where the odds ratio is 1.0). The point of even odds occurs where the heights of the two distributions are equal, and that occurs where the distributions intersect. In the familiarity model, the point of intersection occurs at one level of familiarity in the weak case and at a higher level of familiarity in the strong case. However, these two familiarity values both translate to a value of 0 on the log likelihood-ratio axis (i.e., log[1.0] = 0). Thus, from a likelihood-ratio point of view, the criterion remains fixed at 0 across the two strength conditions.

Note the different patterns of hit and false-alarm rates yielded by these two fixed-criterion models. Unlike the familiarity model, the likelihood-ratio model naturally predicts a mirror effect (i.e., as the hit rate goes up, the false-alarm rate goes down), and that pattern happens to be a nearly universal finding in the recognition memory literature (Glanzer, Adams, Iverson, & Kim, 1993). The mirror effect is not only typically observed for strength manipulations but also for manipulations of word frequency, concreteness, and a variety of other variables (Glanzer & Adams, 1990). The ability of

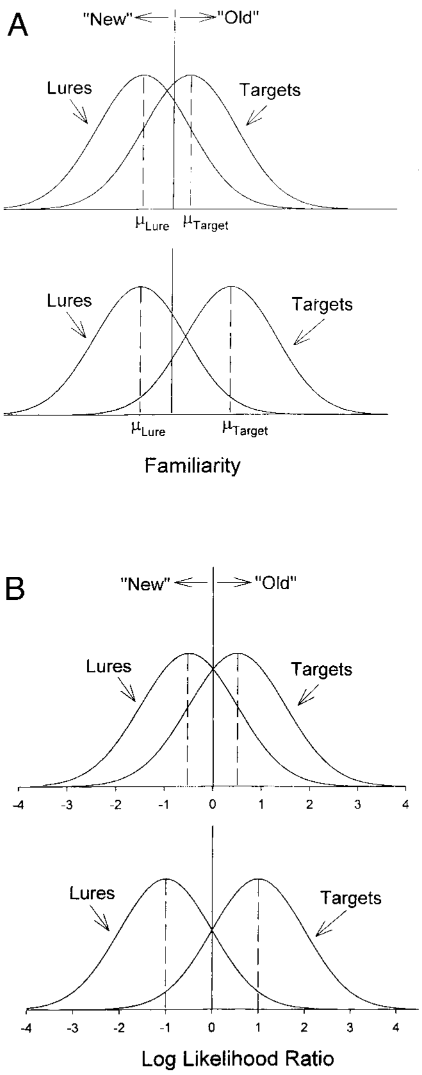

Figure 2. A: Familiarity-based signal-detection model of weak and strong conditions over which the decision criterion remains fixed on the decision axis. Weak: Hit rate = .60; false-alarm rate = .23. Strong: Hit rate = .89; false-alarm rate = .23. B: Likelihood-ratio signal-detection model of weak and strong conditions over which the decision criterion remains fixed on the decision axis. Weak: Hit rate = .69; false-alarm rate = .31. Strong: Hit rate = .84; false-alarm rate = .16.

the likelihood-ratio model to account for the mirror effect in a natural way is one of its most attractive features.

Many models include a familiarity-based, signal-detection process, and, as such, it is probably fair to say that none provides a completely natural account of the mirror effect. These models include search of associative memory (SAM; Gillund & Shiffrin, 1984), theory of distributed associative memory (TODAM; Murdock, 1982, 1983), and MINERVA (Hintzman, 1984, 1988), as well as other models that assume that recognition decisions reflect a combination of retrieval-based and familiarity-based responses (Mandler, 1980; Yonelinas, 1999). Three other major models of memory include a likelihood ratio signal-detection process, and, as such, all provide a very natural account of the mirror effect. These include theories known as attention-likelihood theory (ALT; Glanzer et al., 1993), retrieving effectively from memory (REM; Shiffrin & Steyvers, 1997), and subjective-likelihood theory (SLT; McClelland & Chappel, 1998). Indeed, recent theorizing about recognition memory appears to reflect a shift away from the idea that the decision axis represents a familiarity scale toward the idea that it represents a likelihood-ratio scale.

The focus of the present article is on strength-based mirror effects because a closer examination of these effects may help to distinguish between the two general classes of model mentioned above. When strength is manipulated between lists, a mirror effect is always observed (possibly without exception in the published literature). In one recent example described by Stretch and Wixted (1998a), words on the study list were presented three times each in a strong condition and once each in a weak condition. Obviously, overall recognition performance was better following the strong list than it was following the weak list. Moreover, as is typically the case, a mirror effect was observed. Familiarity-based detection models do not predict this pattern, but they can explain it by assuming that subjects used a higher decision criterion in the strong condition relative to the weak condition. Nevertheless, it could be argued that the likelihood-ratio models account for the result in a more natural way because they do not need to assume that a criterion shift occurs.

Stretch and Wixted (1998a) also manipulated strength within list. In each list, half the words were colored red, and they were presented five times each. The other half were colored blue, and they appeared only once each. The red and blue words were randomly intermixed, and the red word repetitions were randomly scattered throughout the list. On the subsequent recognition test, the red and blue targets were randomly intermixed with red and blue lures. Obviously, the hit rate in the strong (red) condition was expected to exceed the hit rate in the weak (blue) condition, and it did. Of interest was whether the false-alarm rate in the strong condition (i.e., to red lures) would be less than the false-alarm rate in the weak condition (i.e., to blue lures). The likelihood-ratio models discussed above predict such a mirror effect for the same reason it is predicted for the between-list strength manipulation (Figure 2B).

The familiarity-based detection model makes a different prediction. If participants used the same decision criterion (i.e., the same criterion familiarity level) for red items as they do for blue items, then the false-alarm rate to red lures would equal the false-alarm rate to blue lures. After all, red and blue lures are presumably equally familiar (in both cases they are simply words randomly drawn from the same word pool), so if the same familiarity criterion is used for both, their respective false-alarm rates should be the same (Figure 2A). The results reported by Stretch and Wixted (1998a) were consistent with the predictions of the familiarity-based models. In two experiments in which strength was cued by color, the hit rate to the strong items was much higher than the hit rate to the weak items, but the false-alarm rates were nearly identical. Thus, at least with respect to strength manipulations, the between-list paradigm yields data most easily reconciled with a likelihood-ratio model, but the within-list paradigm yields data most easily reconciled with a familiarity-based model.

Because neither the fixed-criterion model correctly anticipates the pattern obtained by varying strength within and between lists, the question that arises is, Which version is better able to accommodate the result it does not predict? Consider the familiarity account first. From this perspective, the between-list strength data suggest that participants adjust the decision criterion as a function of strength such that a higher criterion is used in the strong condition relative to the weak (thereby accounting for the lower false-alarm rate in the strong condition). Participants might behave that way because, having just devoted a relatively large amount of time to the study of each item on the list in the strong condition, they know that the targets will be highly familiar on the upcoming recognition test. If so, it would make sense for them to demand a high sense of familiarity (i.e., to use a higher criterion) before declaring an item to be old. But if they adjust the decision criterion as a function of strength between lists, why would they not do the same when strength is manipulated within list? Stretch and Wixted (1998a) suggested that participants might be willing to adjust the criterion between lists because that requires only one criterion setting per list. Adjusting the criterion within list, by contrast, would require constant back-and-forth movement of the decision criterion throughout the recognition test, and that may not be something participants are very inclined to do (cf. Wixted & Stretch, 2000); hence, no within-list mirror effect. Thus, although the familiarity model does not necessarily predict the differing patterns of results obtained from between- and within-list strength manipulations, it is relatively easy to reconcile this model with those patterns.

At first glance, it seems extremely difficult to reconcile equivalent false-alarm rates across within-list strength conditions with any likelihood ratio account. To do so, one would need to argue that participants used different criterion odds ratios for strong and weak items (for no sensible reason) and that they relied on criteria that were just different enough to yield identical false-alarm rates. The tortured nature of that account contrasts with the extremely simple and appealing account offered by the familiarity model: In the within-list case, participants set a single criterion familiarity value and evaluated all recognition test items (red or blue) against it.

On the other hand, it seems possible that one uninteresting procedural difference between the between- and within-list strength experiments is actually responsible for the differing patterns of results. In the within-list experiment, but not in the between-list experiment, information about strength was carried by the color of a word. The color of a word presented on a computer screen is probably not a property that ordinarily commands a great deal of attention or consideration. Because color is a property that is secondary to what participants typically attend to and think about during list presentation (such as the pronunciation and meaning of the word), perhaps they were inclined to simply ignore the color manipulation. If color was ignored in spite of the fact that it provided useful information, then the results from the within-list strength experiment would not contradict the likelihood ratio account after all. Instead, the equivalent false-alarm rates as a function of strength would reflect the fact that there was really only one effective strength condition for the lures, not two.

The purpose of this research was to test the familiarity and likelihood-ratio signal-detection theories by using a within-list strength manipulation along semantic and perceptual dimensions that are (presumably) already the primary focus of the subject's attention. All three experiments used the same basic design. In each case, half the items on the list were drawn from one category (A) and the other half were drawn from another category (B), and participants were alerted to this fact so that they would not overlook it. During list presentation, the items from one category were strengthened by presenting them five times each, whereas the items from the other category were presented only once each. For half of the participants, the A items were presented five times each and the B items only once, whereas for the other half, the B items were presented five times each and the A items only once. On the subsequent recognition test, the targets from Categories A and B were randomly intermixed with an equal number of lures from Categories A and B, and these were all individually presented for an old or new recognition decision.

In Experiment 1, the lists consisted of words drawn from two distinctly different semantic categories (A = locations and B = professions). Experiments 2 and 3 were similar except that the two types of items making up the lists were differentiated to an even greater degree than in Experiment 1. Specifically, half of the items on the list were words drawn from one semantic category and half were pictures drawn from a different semantic category (specifically, A = profession words and B = bird pictures). The question of interest was whether a mirror effect would now be observed as it always is when strength is manipulated between lists (and as likelihood-ratio models naturally predict).

Experiment 1

In addition to testing the predictions of these two general signal-detection accounts, the first experiment also provided a test of an apparently widespread (and otherwise quite reasonable) assumption that participants readily adjust the decision criterion during the course of a recognition test (e.g., Miller & Wolford, 1999). In the past, researchers working within the familiarity-based version of detection theory have assumed that criterion shifts occur when item strength or class is manipulated within a list. In fact, some studies have been specifically designed to prevent within-list criterion shifts of the kind we tried to facilitate in Experiment 1. Shiffrin, Huber, and Marinelli (1995), for example, presented participants with lists consisting of 25 different semantic and orthographic-phonemic word categories, and they manipulated both length (number of exemplars per category) and strength (number of repetitions per item) between categories. Like us, they were interested in the effects of manipulations like these on the false-alarm rates to lures drawn from the various categories represented on the list. Unlike us, they did not want their participants to be able to adjust the criterion as a function of these manipulations. By using so many categories, their hope was that participants would be unable to keep track of which categories had been strengthened and which had not (and so would be unable to shift the criterion as a function of strength). By contrast, we used only two categories and made it quite clear which one was strengthened and which one was not.

Method

Participants. The participants were 24 undergraduate volunteers from the University of California, San Diego. All were compensated with course credit.

Materials and design. Stimuli were drawn at random for each participant from two pools of 40 words. Each pool consisted of words from a single semantic category: professions (e.g., *plumber, doctor, policeman,* etc.) or geographical locations (*Spain, Canada, Australia,* etc.).

Procedure. On arrival, participants were given a brief description of the experiment and explicit instructions according to a prepared script. Participants were specifically informed about the strength manipulation and the nature of the two semantic categories. That is, they were informed that they would see words belonging to two categories (locations and professions) and that the items from one of these categories would be strengthened by repetition. After signing a consent form, each participant was shown to a small room where he or she completed the experiment individually and without distraction.

Participants were initially shown a study list that consisted of 20 words randomly drawn from each semantic category. These words were presented for 500 ms at 250-ms intervals (presentation time and interstimulus interval were determined by a series of pilot tests). Each semantic category was randomly assigned to one of two encoding conditions, strong or weak, and this assignment was counterbalanced across participants. In the strong condition, words were presented five times each, randomly scattered throughout the list, whereas in the weak condition words were only presented once. Thus, participants saw a total of 120 randomly intermixed word presentations.

Immediately following the study list, participants were given a yes–no recognition test. This test consisted of 20 targets and 20 lures from each semantic category, yielding a total of 80 randomly intermixed test items. Each word was presented individually and participants were required to decide whether they had seen that word before. Participants used a mouse to select one of six response options that corresponded with six levels of confidence: "definitely no," "no," "maybe no," "maybe yes," "yes," and "definitely yes." They were instructed to respond as quickly as possible without sacrificing accuracy.

Results and Discussion

The main results of Experiment 1 are summarized in Table 1.[1] Hit and false-alarm rates were computed for each subject by summing "yes" responses across the three confidence levels. Collapsed across semantic category, the hit rate to strong targets greatly exceeded the hit rate to weak targets (.96 and .77, respectively), $t(23)=6.40$, $p <.01$, which certainly came as no surprise. By contrast, the false-alarm rates to lures from the weak and strong categories were nearly identical (.23 and .22, respectively) and did not differ significantly. Thus, unless the tiny difference in false-alarm rates is taken seriously, no mirror effect was observed in this experiment. Given the large difference in hit rates and the negligible difference in false-alarm rates, sensitivity (d') was much greater in the strong condition relative to the weak, $t(23) = 6.30$, $p < .01$.

For half of the participants, professions was the strong category and locations was the weak, and these data are presented in bold typeface in Table 1. For the other half, locations was the strong category and professions was the weak, and these data are presented in italics in Table 1. These data show that the pattern averaged across categories (discussed earlier) was the same as the pattern observed for the categories considered separately. Thus, for example, when professions was the strong category, the hit rate was higher compared with when it was the weak category, $\mathrm{t}(22) = 4.46$, $p < .01$, but the false-alarm rates to professions did not change much as a function of strength (and the difference did not approach significance). A similar story applies to the location hit and false-alarm rates. The difference in hit rates as a function of strength was highly significant, $\mathrm{t}(22) = 4.60$, $p < .01$, but the small difference in false-alarm rates did not approach significance.

The findings presented above suggest that even when strength information is correlated with a property that participants are undoubtedly processing during list presentation, evidence of a strength-based criterion shift was not forthcoming. Instead, the data are most easily reconciled with the idea that participants responded on the basis of familiarity and maintained the same decision criterion throughout the course of the recognition test (in accordance with the familiarity-based model illustrated in Figure 2A).

Table 1
Hit and False-Alarm (FA) Rates and d′ Scores for the Weak and Strong Conditions From Experiment 1

Stimulus type	Weak			Strong		
	Hit	FA	d'	Hit	FA	d'
Professions	**.75**	**.25**	**1.45**	*.96*	*.21*	*2.52*
Locations	*.80*	*.20*	*1.81*	**.96**	**.22**	**2.60**
Overall	.77	.23	1.63	.96	.22	2.56

Note. Bold values represent results from one group of participants, and italicized values represent results from the other group.

Likelihood-ratio models are challenged by this result because those models predict a mirror effect if participants are assumed to adopt the same criterion odds ratio for both categories. As it stands, the equivalent false-alarm rates coupled with very different hit rates suggest that if the likelihood-ratio model is correct, participants relied on vastly different decision criteria for locations and professions. This can be conveniently illustrated by computing log beta (a standard measure of bias) for the strong and weak categories. Log beta is simply the log of the criterion likelihood ratio, and, ideally, it would equal 0 for both categories (and a mirror effect would be observed). On the basis of the group hit and false-alarm rates, log beta was equal to -1.24 for the strong condition and was approximately 0 for the weak condition. Thus, by this measure, performance was unbiased for the weak items and exhibited a liberal bias for the strong items. Although descriptively accurate in the sense that participants provided more old responses than new responses for items from the strong category only, it seems unlikely that participants would strategically shift their decision criterion in this manner.

Because the main finding of Experiment 1 relies on a null result (i.e., the lack of a significant difference in false-alarm rates), a natural question to ask is whether power to detect an effect was sufficient. To answer that question, we first used the difference in hit rates between conditions to estimate the expected false-alarm effect size. The mean difference between hit rates in Experiment 1 was .18, and a similar difference in false-alarm rates yielded an exceedingly large effect size of 1.31 (see Cohen, 1988, for effect size norms). Subsequent power calculations indicated that we had a nearly 100% chance of detecting such a false-alarm rate difference. On the other hand, as noted by McClelland and Chappel (1998), between-list strength manipulations frequently yield a larger effect on hit rates than on false-alarm rates. Thus, we also computed power by using an effect size that was independent of the effect on hit rates and which corresponded to what Cohen

[1] If any false-alarm rate was zero, it was replaced by a value equal to the inverse of $2n$. If any hit rate equaled one, it was replaced by one minus the inverse of $2n$, where $n = 20$.

(1988) called a medium effect size (namely, 0.5 *SD*). A medium effect size in this case translates into a .07 difference in false-alarm rates across strength conditions. The probability of detecting such an effect was .65 by using a two-tailed test and .77 by using a one-tailed test (but the false-alarm rate differences were not significant even by a one-tailed test).

Experiment 2

Stretch and Wixted (1998a) cued within-list strength by color, and the present Experiment 1 did so by semantic category, but in neither case was a mirror effect observed. Although hit rates increased as a function of strength in both cases, the false-alarm rates were unaffected (in contrast to what reliably occurs in the analogous between-list case). In Experiment 2, in an effort to make the within-list strength manipulation even more salient to participants, we used study lists consisting of words from one semantic category (professions) and pictures from another (animals). For half of the participants, the words were strengthened by presenting them five times each, whereas for the other half, the pictures were strengthened in the same way.

As described thus far, the familiarity-based version of signal-detection theory assumes that all recognition decisions are based on values that lie on a unidimensional psychological scale (namely, familiarity). However, there is no guarantee that stimuli as different as pictures and words will be evaluated along the same dimension. If they are not, then, presumably, participants would treat these items as belonging to two separate lists although they are randomly intermixed. Under such conditions, a mirror effect should be observed, just as it always is in a between-list strength paradigm. More specifically, when pictures are strong, participants should use a high criterion along whatever dimension is used to decide whether pictures are old or new (yielding a low false-alarm rate); when pictures are weak, they should use a low criterion (yielding a higher false-alarm rate). Moreover, they should do this regardless of whether the intermixed words are weak or strong because words may be evaluated along a different psychological dimension.

On the other hand, Hintzman, Curran, and Caulton (1995) presented surprising but compelling evidence suggesting that, as different as they might seem, words and pictures (specifically, drawings of objects) are evaluated along the same unidimensional psychological scale, at least when the items are intermixed. That being the case, one might expect participants to be as reluctant to shift the decision criterion as a function of strength here as they were when strength was cued by color or semantic category.

Method

Participants. The participants were 24 undergraduate volunteers from the University of California, San Diego, all of whom were compensated with course credit.

Materials and design. Targets and lures were drawn at random from two categories (words and icons) that each comprised 40 items. The pool of words was composed of the 40 professions used in Experiment 1. The icons were all colored drawings of various animals selected from Broderbund ClickArt 200,000 (1997).

Procedure. Experiment 2 was identical to Experiment 1 except that animal icons were used instead of location words. The instructions to participants clearly indicated the nature of the list (animal icons and professions words) and the strengthening manipulation (i.e., they were told that items from one class would be presented multiple times). Each participant then saw the study list composed of randomly intermixed words and icons (20 of each). Items from one category were presented once each, whereas items from the other category were presented five times each. The category chosen for strengthening was counterbalanced across participants such that half of the participants (randomly selected) had the icons strengthened and the other half of the participants had words strengthened. After the study list was presented, participants were given a yes–no recognition test containing 20 targets and 20 lures from each category, all randomly intermixed.

Results and Discussion

The results of Experiment 2 are shown in Table 2. For the most part, these data exhibited the same pattern observed in Experiment 1. With regard to the overall data (collapsed across words and icons), false-alarm rates did not differ significantly across conditions, but the hit rate was much higher in the strong condition compared with the weak condition, *t*(23)=2.98, *p* < .01. For words alone, the false-alarm rate did not change from the weak to the strong condition, but the hit rate increased significantly, *t*(22)=2.49, *p* < .05. For icons considered alone, the change in hit rates approached significance (*p* = .07), whereas the change in false-alarm rates did not. Overall, *d*′ was higher in the strong condition relative to the weak, but the difference was not quite significant, *t*(23)=1.78. This was probably due to the fact that the performance of many participants was at ceiling, as well as the fact that false-alarm rates showed a slight (and probably spurious) increase as a function of strength. When sensitivity values were computed by using false-alarm rates averaged across strength (.175 for words and .115 for icons), *d*′ was significantly higher in the strong condition for all comparisons: *t*(22)=2.39, *p*=.03, for words; *t*(22) = 2.16, *p* = .04, for icons; and *t*(23) = 2.6, *p* = .02, overall.

Given the null effect on false-alarm rates as a function of strength, we again performed power calculations. The mean difference in hit rates between conditions (.09), and a similar difference in false-alarm rates (.086), corresponded with a medium effect size of .5, so we used .086 as the mean difference on which to base our power analyses. Those analyses revealed that we had a 65% chance of detecting a significant effect using a two-tailed test and a 77% chance using a one-tailed test. The observed difference in false-alarm rates between conditions was still not significant, even by one-tailed test (*p* = .2), and the direction of the effect was opposite to what likelihood-ratio models predict.

Table 2
Hit and False-Alarm (FA) Rates and d′ Scores for the Weak and Strong Conditions From Experiment 2

	Weak			Strong		
Stimulus type	Hit	FA	*d*′	Hit	FA	*d*′
Words	**.85**	**.16**	**2.10**	*.95*	*.19*	*2.53*
Icons	*.87*	*.10*	*2.45*	**.95**	**.13**	**2.89**
Overall	.86	.13	2.28	.95	.16	2.69

Note. Bold values represent results from one group of participants, and italicized values represent results from the other group.

Although Experiment 2 seemed to have adequate power to detect a difference between false-alarm rates as a function of strength, it could be argued that the strength manipulation was not altogether effective in that d' was not significantly higher in the strong condition relative to the weak (unless false-alarm rates were first averaged across the weak and strong conditions). In light of that, we designed Experiment 3 to test for a within-list mirror effect by using pictures and words with longer lists (to avoid ceiling effect problems) and more participants (to increase power).

Experiment 3

In Experiment 3, we doubled the number of targets and lures from each category on the assumption that recognition performance would decrease with list length (thereby reducing the possibility of a ceiling effect). In addition, we doubled our sample size to increase power to detect any false-alarm rate difference that might exist as a function of strength. Otherwise, Experiment 3 was identical to Experiment 2.

Method

Participants. The participants were 48 undergraduate volunteers from the University of California, San Diego, all of whom were compensated with course credit.

Materials and design. Targets and lures were drawn at random from two categories (words and icons) that each contained 80 items. The pool of words was composed of 80 professions, 40 of which were used in Experiment 1. The icons were all colored photographs of various animals selected from Broderbund ClickArt 200,000 (1997).

Procedure. Participants were given precise instructions according to the script prepared for Experiment 2. After signing a consent form, each participant was shown to a small room where he or she completed the task individually and without distraction. Each participant was presented with a study list composed of randomly intermixed words and icons (40 of each). The items in one category (i.e., the strengthened category) were shown five times each, yielding a total of 240 item presentations. After study list presentation, participants were given a standard yes–no recognition test containing 40 targets and 40 lures from each category.

Results and Discussion

The results of Experiment 3 are shown in Table 3. The overall data (collapsed across words and icons) showed essentially the same pattern that was observed in the previous experiments. The overall hit rate increased significantly as a function of strength, $t(47) = 4.44$, $p < .01$, but the difference in false-alarm rates did not approach significance. For icons alone and for words alone, between-subject comparisons reveal that hit rates increased with strength, $t(46) = 2.71$, $p < .01$, and $t(46) = 3.01$, $p < .01$, respectively, but the false-alarm rates did not differ significantly in either case. All differences in d' values between the weak and strong conditions were highly significant at the two-tailed level: $t(46) = 3.20$, $p < .01$, for words; $t(46) = 2.44$, $p < .02$, for icons; and $t(47) = 3.63$, $p < .01$, overall.

Table 3
Hit and False-Alarm (FA) Rates and d' Scores for the Weak and Strong Conditions From Experiment 3

	Weak			Strong		
Stimulus type	Hit	FA	d'	Hit	FA	d'
Words	**.76**	**.32**	**1.27**	*.86*	*.28*	*1.89*
Icons	*.73*	*.13*	*2.00*	**.84**	**.12**	**2.63**
Overall	.74	.23	1.63	.85	.20	2.26

Note. Bold values represent results from one group of participants, and italicized values represent results from the other group.

False-alarm rates were significantly greater for words than for icons, $t(47) = 7.13$, $p < .01$, but the hit rates did not differ significantly. Why this class-based difference emerged is unclear, although it is not especially surprising because we did not try to equate the words and icons for preexperimental familiarity. Still, this difference did increase error variance in the within-subject false-alarm t tests, reducing power to detect a difference. The increased error variance results from the fact that a within-subject t test involves computing a difference score (weak false-alarm rate minus strong false-alarm rate) for every participant and then testing the null hypothesis that the difference is zero. For some participants, the subtraction involved a weak false-alarm rate for words minus a strong false-alarm rate for icons (which tended to yield a relatively large positive value), and for others the subtraction involved a weak false-alarm rate for icons minus a strong false-alarm rate for words (which tended to yield a large negative value).

We therefore conducted another test on the overall .026 false-alarm rate difference (weak minus strong) after accounting for the expected difference in false-alarm rates because of class. On average, the word false-alarm rate minus icon false-alarm rate (collapsed across strength) was .182. Thus, for cases in which the weak minus strong computation involved a weak false-alarm rate for words minus a strong false-alarm rate for icons, we subtracted an additional .182. For cases in which the weak minus strong computation involved a weak false-alarm rate for icons minus a strong false-alarm rate for words, we added an additional .182. This had the effect of reducing error variance while leaving the mean difference in overall false-alarm rates at .026. Even after this adjusted analysis, the difference between weak and strong false-alarm rates was still not significant, $t(47) = 1.03$.

The mean difference in hit rates (.11) corresponded with a large effect size of .91. Our power calculations indicated that we had an 83% chance of detecting a similar difference in false-alarm rates at the two-tailed level ($d = .6$). We would have a 90% chance of detecting a difference at the one-tailed level, but the observed difference in false-alarm rates between conditions was not significant, even by a one-tailed test ($p = .2$). A medium effect size of .5 corresponds to a false-alarm rate difference of .087, and, as in the previous experiments, the probability of detecting such an effect was .65 for a two-tailed test and .77 for a one-tailed test.

Although these analyses show no convincing evidence of a criterion shift, further analyses of the confidence data suggested that a small criterion shift may have actually occurred. Because participants supplied a confidence rating for each old or new decision, it was possible to compute false-alarm rates for responses exceeding each confidence criterion. The false-alarm rates shown in Table 3 involve responses to lures that received a rating of + or higher (i.e., +, ++, or +++ responses on the confidence rating scale). However, other false-alarm rates can be computed by cumulating responses from a different point on the confidence scale. For example, a false-alarm rate corresponding to a conservative criterion setting can be computed by using only the high-confident, +++ responses (i.e., the number of +++ responses to lures divided by the total number of lures). In a similar fashion, a false-alarm rate corresponding to a liberal criterion setting can be computed by using all responses of -- or greater (i.e., the number of --, -, +, ++ and +++ responses to lures divided by the total number of lures).

Stretch and Wixted (1998b) showed that when strength is manipulated between list, all of the confidence criteria (not just the old/new decision criterion) shift, and the degree to which they do increases considerably the more liberal the criterion setting is (a pattern uniquely predicted by a likelihood ratio account). An idealized version of the prediction made by likelihood-ratio models is depicted in Figure 3. The figure shows not only how the old/new decision criterion would shift on the familiarity axis as a function of strength in order maintain a constant 1:1 likelihood ratio but also how various confidence criteria would shift to maintain constant likelihood ratios of greater than or less than even odds.[2] The leftmost confidence criterion in the upper panel of Figure 3 separates "new" (--) responses from "certain new" (---) responses, and, in this example, it is placed at the point on the familiarity axis where the height of the target distribution is 1/9 that of the lure distribution. Thus, a test item that generates that level of familiarity is 9 times as likely to be a lure as a target. To maintain the same criterion when conditions change from strong to weak (upper panel to lower panel in Figure 3), this criterion must shift to a much lower point on the familiarity axis. As a result, if false-alarm rates are computed by using the leftmost criterion (by summing --, -, +, ++, and +++ responses to lures), the effect of the strength manipulation should be particularly noticeable. In this hypothetical example, the liberal false-alarm rate that we computed by using the leftmost confidence criterion increases from about 55% to about 90% as conditions change from strong to weak. The traditional false-alarm rate, which is computed by using the old/new criterion, increases by a smaller amount (from 16% to 31% in this example), and the conservative false-alarm rate, computed by using only the rightmost criterion, actually decreases slightly (from 2.5% to 1.5%). Thus, the effect of the strength manipulation on false-alarm rates should increase the more liberal (i.e., the more to the left) the criterion happens to be. Indeed, that pattern is reliably observed in between-list strength manipulations (Stretch & Wixted, 1998b), and the important point for present purposes is that the same pattern was observed in Experiment 3 (although none of the false-alarm rate differences was statistically significant). Specifically, the conservative false-alarm rates in Experiment 3 (based only on +++ responses) were identical in the strong and weak conditions (.049 in both cases), whereas the traditional false-alarm rates (based on +, ++, and +++ responses) differed by a larger, albeit nonsignificant, amount (.200 vs. .226, respectively, as shown in Table 3), and the liberal false-alarm rates (based on --, -, +, ++, and +++ responses) differed by a still larger amount (.584 vs. .663, respectively). This latter difference was also not significant, but it came close after adjusting for baseline differences in liberal false-alarm rates to icons and words, $t(47) = 1.88$. Although none of the false-alarm rate differences reached significance, the data appear to reflect an attenuated version of the pattern that is observed when strength is manipulated between list (Stretch & Wixted, 1998b). The confidence data from Experiments 1 and 2 showed no such pattern.

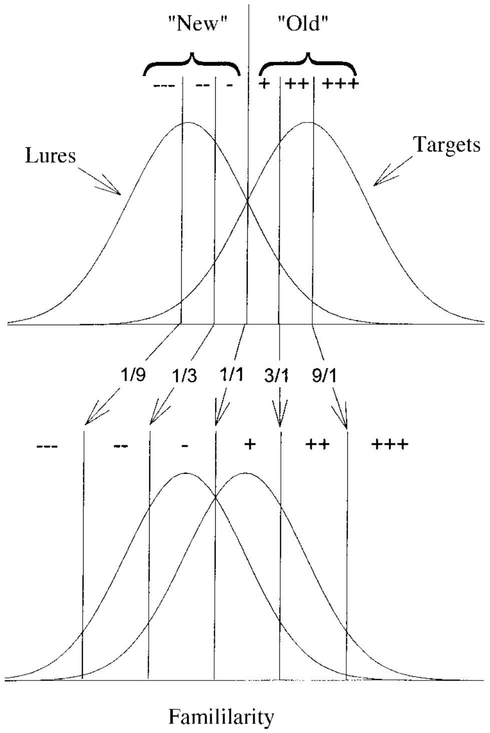

Figure 3. Illustration of the movement of the confidence criteria as a function of d' according to likelihood-ratio models. Strong ($d' = 2.0$); weak ($d' = 1.0$).

We also analyzed the ROC data pooled over participants and fit a detection model based on the logistic approximation to the normal distribution, exactly as Stretch and Wixted (1998b) did for their between-list strength data. The basic model involved nine parameters: d_{weak}, r_{weak}, d_{strong}, r_{strong}, (where d_{weak} and d_{strong} are parameters analogous d', and r is the ratio of the standard deviations of the lure distribution to the standard deviation of the target distribution) and five confidence criterion parameters (namely, the criteria separating +++/++, ++/+, +/-, -/--, and --/--- responses). If all five confidence criteria remained fixed as a function of strength, then the fit would not be significantly improved by adding five additional confidence criterion parameters (i.e., five criteria for the weak condition and five more for the strong). However, when this was done, the fit was significantly improved, $\chi^2(5) = 27.29$, mostly due to the effect of allowing the most liberal of the five confidence criteria to differ between the weak and strong conditions. Thus, on the basis of all of these analyses, the safest conclusion is that a criterion shift did occur, one that was too small to be detected at the level of false-alarm rates but that could

[2] If the familiarity scale in Figure 3 were translated into a log likelihood-ratio scale, the confidence criteria would be fixed on the decision axis as a function of strength (just as the old or new criterion is in Figure 2B). What would change are the variances of the target and lure distributions in the weak condition.

be detected with the higher power afforded by ROC analysis.[3] Moreover, the small shift that apparently did occur is the kind that would be expected if likelihood-ratio models were correct in that the magnitude of the effect was larger the more liberal the criterion setting happened to be.

The overall conclusion from these three experiments is that strength manipulations within list tend not to produce a criterion shift (Experiments 1 and 2), or they produce a shift that is greatly attenuated compared with the between-list case (Experiment 3). Such a result is more easily explained by familiarity versions of detection theory than by likelihood ratio accounts. Still, whenever the confidence criteria do shift (such as when strength is manipulated between list and, to some extent, in Experiment 3), it appears that they shift in the manner predicted by likelihood-ratio models and that presents an interesting theoretical puzzle that needs to be explained.

General Discussion

Taken together, the results of these experiments demonstrate that when strength is manipulated for different categories of items within the same list, the hit rate for the strong items far exceeds that of the weak items, but the false-alarm rates to lures drawn from the weak and strong categories do not differ (i.e., no mirror effect is observed). Whereas Stretch and Wixted (1998a) reported this result for words presented in one of two colors, the present research shows that the same result obtains even when the relevant categories are semantic (e.g., strong locations vs. weak professions) or both semantic and perceptual (e.g., strong location words vs. weak animal pictures). This is important because it shows that the otherwise pervasive strength-based mirror effect is much less likely to occur for within-list strength manipulations, even when strength is conspicuously correlated with item properties that are presumably the focus of a subject's attention.

Shiffrin et al. (1995) reported results like these for lists involving many categories, so many, in fact, that participants were unlikely to have been able to use a different decision criterion for the different categories even if they had wanted to (indeed, that study was specifically designed to prevent category-specific criterion shifts). In their experiment, strengthening some categories on the list increased the hit rate but did not affect the false-alarm rate. The present results demonstrate that category-specific false-alarm rates are largely unaffected by strength, even when the lists involved only two highly distinct categories that would have allowed participants to easily adjust the decision criterion had they been so inclined. Still, they did not.

Any null result could represent nothing more than a failure to detect a real effect, and that caveat applies to the null effect of strength on false-alarm rates observed here. Indeed, a close look at the confidence data suggested that the nonsignificant change in false-alarm rates seen in Experiment 3 probably reflected a small criterion shift. Thus, the safest conclusion to draw from these three experiments is that participants are reluctant to shift the criterion when strength is conspicuously manipulated within list, and that any shift that might occur is surprisingly small even when extraordinary steps are introduced to make it happen. The effect is so small, in fact, that it often does not show up even in the raw false-alarm rates (much less than in the statistical analyses of those rates). Across the six conditions in the three experiments reported here, the false-alarm rate increased with strength three times (for locations in Experiment 1 and for both words and icons in Experiment 2) and decreased in strength three times (for professions in Experiment 1 and for both words and icons in Experiment 3). In no case did the observed differences approach significance, although the hit rate differences were always significant. Similarly, in two within-list strength experiments reported by Stretch and Wixted (1998a), the false-alarm rate increased as a function of strength once (Experiment 4) and decreased once (Experiment 5). Again, neither effect was significant. This pattern contrasts with what is observed in between-list strength manipulations where a significant increase in the hit rate is almost always accompanied by a significant decrease in the false-alarm rate (e.g., Murnane & Shiffrin, 1991; Stretch & Wixted, 1998a, Experiment 1).

The pattern of results observed for within-list strength manipulations is most easily reconciled with familiarity-based signal-detection models that assume that subjects maintain a single old/new decision criterion throughout the course of a recognition test. By contrast, these results seem more difficult to reconcile with likelihood ratio accounts (specific versions of which are discussed in more detail below). As such, one general implication of these findings is that the decision axis in recognition memory may represent a familiarity scale, as models like SAM (Gillund & Shiffrin, 1984), TODAM (Murdock, 1982, 1983) and MINERVA (Hintzman, 1984, 1988) assume, not a likelihood ratio scale, as models like ALT (Glanzer et al., 1993), REM (Shiffrin & Steyvers, 1997) and SLT (McClelland & Chappell, 1998) assume. Our findings are also easily reconciled with some recent models of the mirror effect (Reder et al., 2000; Sikstrom, 2001) and with random walk or diffusion models (Link, 1992; Ratcliff, 1978, 1988), which contain a view of the decision criterion more similar to that envisioned by familiarity-based detection models than by likelihood-ratio-based detection models.

The familiarity model that best accommodates the data reported here is illustrated in Figure 2A. The model assumes that participants set and maintain a single decision criterion throughout the course of the recognition test even though items from different categories that obviously differ in strength appear on the test. That participants would be reluctant to shift the criterion as a function of strength is somewhat surprising given that they apparently do just that when strength is manipulated between list. In the between-list case, the strong condition is always associated with a higher hit rate and a lower false-alarm rate, and the lower false-alarm rate presumably reflects an upward shift of the criterion on the familiarity axis. Thus, although participants appear to understand that a higher criterion makes sense when the targets are strong, they prefer not to act on that knowledge when strength is manipulated within list.

Why not? The answer may be that to do otherwise would require constant criterion shifts throughout the course of the recognition test. One criterion would be needed if the first test item happened to be a profession, a different criterion would be needed if the next item happened to be a location, and then the original criterion

[3] Although this extra power derives from assumptions that may or may not be correct, those assumptions being that the logistic approximation is accurate, that the responses are all independent, and that pooling over participants yields representative data.

would need to be reinstated if the next item was another profession (and so on throughout the recognition test). Absent any strong incentive to do otherwise, participants may simply find it much easier to adopt a single decision criterion and judge all test items in relation to it (cf. Wixted & Stretch, 2000).

Likelihood-Ratio-Like Behavior

As indicated earlier, Stretch and Wixted (1998b) showed that when strength is manipulated between list, the confidence criteria shift on the decision axis in the manner predicted by likelihood-ratio models (as illustrated in Figure 3). The experiments reported in the present article, as well as earlier experiments reported by Stretch and Wixted (1998a), suggest that the criteria are much less likely to shift when strength is manipulated within list, contrary to what likelihood-ratio models predict. Why should the between-list strength manipulation yield results that seem to cry out for a likelihood-ratio interpretation, whereas the within-list strength manipulation yields results that weigh against that interpretation? The answer may be that participants arrive at their decisions in a manner consistent with classical (i.e., familiarity-based) detection theory, but they also bring to the experimental situation a history of learning that is relevant to the task at hand. According to the standard detection model, the confidence criteria are positioned on the familiarity axis in advance of the recognition test, and the familiarity of each test item is assessed relative to those criteria. In the simplest version of this model, the criteria do not shift item-by-item during the course of a recognition test, which is why manipulations of strength within list tend not to affect the false-alarm rate. When strength is manipulated between list, by contrast, participants appear to be more willing to shift the confidence criteria, and the reason why they shift them in a way that corresponds to predictions of the likelihood-ratio account may have to do with their prior history of learning.

Imagine, for example, that participants have expressed varying degrees of confidence in recognition decisions in the past and that they have encountered consequences for doing so (e.g., they have occasionally discovered whether they were correct after expressing high confidence in an “old” response). As argued in detail by Wixted and Gaitan (in press), what that learning history would teach assuming that participants are sensitive to it, is that the odds of being correct for a particular level of familiarity differ depending on the conditions of learning. When the conditions of learning are favorable (i.e., when d' is high), a moderately high level of familiarity might be associated with relatively high odds (say, 9/1) of being correct for responding “old” (as in the upper panel of Figure 3). When the conditions of learning are less favorable, experience would teach that a higher level of familiarity would be needed for those same high odds to be in effect (as in the lower panel).

Note that these learned odds are very similar to likelihood ratios, but they are not the computed odds that a particular item is a target. Instead, they are the learned odds of being correct for particular levels of familiarity given the parameters of the learning situation and that learning dictates where the confidence criteria should be placed in advance of a recognition test. In the between-list strength manipulation, this would result in a fanning of the confidence criteria as strength decreases (because participants have learned from past experience that the criteria need to be set that way to maintain the appropriate odds of being correct for varying degrees of confidence). In the within-list case, the criteria would be set prior to the recognition test (just as they always are) and then remain fixed whether the test item is weak or strong.

Note that in this view of the situation, item-by-item likelihood ratios are not computed, but the confidence criteria fan out on the decision axis anyway. Furthermore, according to this model, the processes that give rise to the appropriate placement of the confidence criteria and the processes that determine whether participants also shift the criteria on an item-by-item basis during the course of a recognition test are different (whereas those processes are one and the same in likelihood-ratio models). Participants are apparently not inclined to adjust the criteria item by item, perhaps because doing so involves more mental effort than they prefer to exert. Even so, participants appear to have some idea of where to set the confidence criteria after having studied a list in order to maintain a particular likelihood of being correct for each level of confidence that might be expressed in subsequent recognition decisions (knowledge that they readily reveal when strength is manipulated between lists).

Implications for Likelihood-Ratio Models

As we have indicated throughout this article the fact that the false-alarm rate often remains constant when strength is manipulated within list is not consistent with the generic likelihood-ratio account depicted in Figure 2B. We turn now to a discussion of three specific likelihood-ratio models to see how the present results bear on them.

Attention-likelihood theory. The within-list results appear to be especially difficult to reconcile with a likelihood-ratio model such as ALT (Glanzer & Adams, 1990; Glanzer et al., 1993). ALT is a feature-sampling model that was basically designed to explain the word-frequency mirror effect. According to this theory, words are represented by an array of features, some of which are marked (due to preexperimental exposure to the words) and some of which are not. When a word is presented on a study list, additional features are marked in proportion to how much attention the word receives. A key assumption of this model is that low-frequency (LF) words receive more attention (and therefore have a greater proportion of features marked during study) than high-frequency (HF) words. By contrast, LF and HF lures (i.e., new items) are assumed to have an equal number of marked features.

In this model, the number of marked features determines an item’s level of familiarity, and that number is distributed across items according to the binomial. The binomial lure distributions for HF and LF items have the same mean and variance, whereas the LF target distribution has a higher mean (and variance) than the HF target distribution. If participants used a single decision criterion on the familiarity axis and responded on the basis of whether the test item exceeded that criterion, then no mirror effect would be observed. Instead, LF words would have a higher hit rate than HF words, but the false-alarm rates would be the same.

Why, then, is a mirror effect observed? According to ALT, a mirror effect is produced for the reasons already discussed in relation to Figure 2B. Specifically, during the recognition test, participants are assumed to first identify the test word as belonging to one of two classes (LF or HF) and to then generate a likelihood ratio from their knowledge of the parameters of the corresponding target and lure distributions. For example, a particular test item with familiarity *f* (corresponding to the number of marked features) might first be identified as an LF word. According to the model, a likelihood ratio would then be computed on the basis of the ratio of the height of the LF target distribution to the height of the LF lure distribution at the point, *f,* on the familiarity axis. When the likelihood ratio exceeds one, an unbiased subject would respond "old," otherwise a "new" response would be given. The same process occurs for HF words, except that now the HF target and lure distributions are used in the likelihood-ratio computations. Because the mean of the HF target distribution is lower than that of the LF target distribution, a HF lure with a familiarity value of *f* will generate a higher log-likelihood ratio than an LF lure associated with the same level of familiarity. Thus, although HF and LF lures are equally familiar, on average, HF lures will yield a higher false-alarm rate than will LF lures.

Note that this model easily explains why a word-frequency mirror effect is observed when HF and LF words are intermixed in the same list. Even under those conditions, participants compute likelihood ratios using the appropriate frequency-specific distributions. A very similar analysis explains why a mirror effect should have been observed for within-list strength manipulations of the kind we examined here. In this case as well, the lure distributions for both the strong and weak categories are the same (just as the lure distributions for HF and LF words are assumed to be), and the target distribution for the strengthened category is situated higher on the familiarity axis than the nonstrengthened category, just as the target distribution for LF words is relative to HF words. In the strength manipulations used in this study, the subject should be acutely aware of which category is strong and which is weak, so they should have no trouble using their knowledge of the location of the strong distribution to compute likelihood ratios for the strong items and their knowledge of the location of the weak distribution to compute likelihood ratio for the weak items. If they did, a mirror effect would be observed (for the exact same reason it is observed for word-frequency manipulations).

The fact that no mirror effect was observed for the within-list strength manipulations suggests one of the following two possibilities:

1. Participants used different criterion odds ratios for strong and weak items.

2. In their likelihood-ratio computations, participants are inclined to compute odds ratios by using a single, composite target distribution (and single lure distribution) throughout the course of the recognition test.

The first possibility seems too implausible on its face to offer a viable account. Why would participants exhibit a liberal bias for strong items and a more conservative bias for weak items, with the difference in response bias being just large enough to maintain equal false-alarm rates? The absence of any principled reason for response bias differences like this renders this account untenable. Still, this account is logically possible, and it might be worth exploring if some principled explanation ever does emerge.

The second possibility is analogous to the idea that participants are simply disinclined to shift the decision criterion on the familiarity axis as a function of strength within list. In the likelihood-ratio account, the corresponding assumption would be that participants use a single psychological representation of the target distribution and a single psychological representation of the lure distribution to compute the odds for all of the test items. Thus, for example, if professions were the strong category and locations the weak category, participants would not use a target distribution with a high mean to compute likelihood ratios for professions and a separate target distribution with a lower mean to compute likelihood ratios for locations. Instead, perhaps participants are inclined to use a single composite target distribution for both professions and locations. Because the lure distributions for professions and locations are the same, this would yield the observed pattern involving an effect on hit rates without a corresponding effect on false-alarm rates.

This explanation seems viable, but it is somewhat puzzling and contrasts with what is ordinarily assumed. It seems puzzling because ALT assumes that participants are already expending considerable computational effort on each test item to compute a likelihood ratio. Participants also clearly know that one category is strong and the other is weak. Given that, it seems odd that they would not use their knowledge of strength to compute strength-specific likelihood ratios (as they do when strength is manipulated between list). Still, it is conceivable that although significant computational effort is put into every recognition decision, the extra effort that would be required to use different target distribution statistics for strong and weak items is not something participants are willing to expend.

Perhaps more troubling than this is the fact that ALT already assumes that participants change the psychological representation of the target distribution on an item-by-item basis during the course of a recognition test to compute likelihood ratios. Specifically, as indicated above, the mirror effect produced by manipulating word frequency within list is explained by assuming that participants use one target distribution with a high mean when assessing low-frequency test items and another target distribution with a lower mean when assessing high-frequency test items. Although the high- and low-frequency lure distributions have the same mean, the use of different target distribution statistics in the likelihood-ratio computations would yield a mirror effect. But if participants readily change target distribution statistics on an item-by-item basis as a function of word frequency, it seems odd that they would be so reluctant to do the same as a function of strength.

Glanzer et al. (1993) noted that ALT can be shown to predict a word-frequency mirror effect even if participants are unaware of the word-frequency manipulation or are unaware of the fact that LF targets have a higher mean familiarity than do HF targets (and so use only one target distribution and one lure distribution in their likelihood-ratio computations). The key process that gives rise to this prediction is the greater attention that LF words receive both at study and at test. However, making a similar move here (i.e., making the assumption that participants are unaware of the categorical distinction) seems less satisfying because participants are very definitely aware of the two categories of items on the list, and they know perfectly well which category is strong and which is weak. Still, for the present results to be reconciled with ALT, one must assume that participants choose not to take advantage of knowledge they clearly posses when strength is manipulated within list (and even though they are always assumed to be making item-by-item likelihood-ratio computations anyway).

Subjective-likelihood theory. SLT does not assume that participants are privy to the shapes of the underlying distributions in the same way that ALT does. Instead, feature-based computations yield estimates of the likelihood of encountering the features of the recognition test item based on prior learning. One likelihood is estimated on the assumption that the test item appeared on the list and another is estimated on the assumption that it did not.

Words in this model are represented by arrays of features that can take on binary values (0 or 1). Memory for each list item consists of a *detector,* which is an array of feature probability estimates (one probability estimate per feature). Each estimate reflects the probability that a feature will be associated with a value of 1, and these estimates change with learning. Imagine, for example, an item consisting of six features with values of {1, 1, 1, 0, 0, 0}. Initially, the detector for this item might have probability estimates of {.5, .5, .5, .5, .5, .5}. After studying the item once, these values might change to {.6, .6, .6, .4, .4, .4}, and after studying the item a second time these values might change to {.7, .7, .7, .3, .3, .3}. Thus, the detector for a particular word learns to increase its estimates of encountering a 1 for a feature each time that value actually is encountered and to decrease its estimates of encountering a 1 each time a 0 is encountered.

On the recognition test, the target represented by {1, 1, 1, 0, 0, 0} might be presented again for an old/new decision. According to the model, the memory system would estimate the likelihood of encountering each of those six features based on the current state of the detector, which is {.7, .7, .7, .3, .3, .3}. Thus, the probability of encountering a 1 in the first position is .7, as is the probability of encountering a 1 in the second and third positions. The probability of encountering a 1 in each of the last three positions is .3, which means that the probability of encountering a 0 (which is what was actually observed) is .7 for all three positions. The overall likelihood of encountering the pattern {1, 1, 1, 0, 0, 0} given that it is the word that produced the image is therefore $.7^6$, or .118. A second likelihood-ratio computation would be performed on the assumption that the item is new, which, in the simplest case, would involve a probability array of {.5, .5, .5, .5, .5, .5}. This would yield a likelihood estimate of $.5^6$, or .016. The likelihood ratio for this item would be .118/.016, or 7.37.

A similar set of computations would be performed for lures, one of which might be represented by {1, 1, 0, 0, 1, 0}. When compared against the detector set to {.7, .7, .7, .3, .3, .3}, the probabilities of encountering each feature of this lure would be .7, .7,.3, .7, .3, and .7 such that the overall likelihood of encountering this set of features given that the item appeared on the list would be .0216. The likelihood of encountering this array of features given that the item is a lure would again be $.5^6$, or .016. Thus, the likelihood ratio for this lure would be 1.35.

With further study (i.e., with each additional presentation of the target item on the list), the detector changes its estimates to more closely match the target item. Thus, after a third presentation of an item represented by {1, 1, 1, 0, 0, 0}, the detector's estimates would be {.8, .8, .8, .2, .2, .2}. Repeating the computations described above would now yield a likelihood ratio of $.8^6/.5^6$, or 16.38 for this target. Doing the same for a lure represented by {1, 1, 0, 0, 1, 0} yields a likelihood ratio of $(.8^3)(.2^3)/.5^6$, or 0.256. Thus, strengthening the target item increased the likelihood ratio associated with the target and decreased the likelihood ratio associated with the lure. Hence, the strength-based mirror effect.

The computations just presented capture what would happen if the list consisted of just one item (such that only one detector was stored). For lists involving multiple items (i.e., for the typical case), each list item is associated with its own detector. When a word is presented for a recognition decision, the model assumes that the word's representation is compared to every detector, with computations for each performed in the manner just illustrated. Each comparison yields a likelihood ratio, and the highest likelihood ratio is used to make the decision (i.e., the response is "old" if the highest likelihood ratio exceeds the criterion and is "new" otherwise). Although this is a global matching model, it predicts a between-list mirror effect for strength manipulations essentially for the reasons discussed above in relation to the simplest case (where the probe item was matched against only one detector).

Does this model also predict a strength-based mirror effect for the within-list manipulation? That probably depends on assumptions about the details of the global matching process. In the typical case, each recognition test item is theoretically evaluated against all of the detectors created by the study list. However, if the list consists of items from more than one category, the comparison process may be category specific. If, say, icons are preferentially matched against the detectors for icons and words are preferentially matched against the detectors for words, then the machinery of the model for the within-list case is the same as that for the between-list case (and a mirror effect would definitely be expected). On the other hand, if all recognition test items are matched against all detectors regardless of any other considerations, it may be possible for the model to accommodate the kind of result we observed in Experiments 1 through 3. Evidence bearing on this issue (discussed in detail below) appears to suggest that test items are matched preferentially against category-specific memory traces (or detectors) in mixed lists. If so, then SLT predicts a mirror effect for the within-list strength manipulation for the same reason it predicts one for the between-list strength manipulation.

Retrieving effectively from memory. As both McClelland and Chappell (1998) and Shiffrin and Steyvers (1997) have noted, REM and SLT are a lot alike. Two significant differences are that items in REM are represented by arrays of features that are not binary (instead, values are drawn from a geometric distribution) and that the likelihood ratios computed for each test item against every stored list item are averaged (rather than taking the highest of them, as in SLT). With regard to the first difference, the array of features representing an item can assume any positive integer value. In REM, these were determined by drawing values from a geometric distribution with its single defining parameter set to 0.45. The memory traces formed by studying words on a list are called *images,* and they basically consist of increasingly accurate copies of the list item (not probability estimates such as those comprising the detectors in SLT). For example, a list word might be represented by the feature values {1, 2, 1, 3, 1, 2}. The image it creates might be represented by {0, 2, 0, 0, 1, 0} after one presentation of the item and by {1, 2, 1, 3, 1, 0} after several more presentations. When the target item is presented for a recognition decision, the features of the test item are compared against the features of the image. For each feature comparison, two values are computed: the likelihood that the features match (or mismatch) given that the image was created by this test item and the likelihood that the features match (or mismatch) given that this test item is a lure. The ratio of these two values is the likelihood ratio.

This model predicts a strength-based mirror effect for between-list strength manipulations for essentially the same reason that SLT does. Moreover, like SLT, REM is a global matching model that assumes that each recognition test item is compared against all of the images created by the list items. Whether REM necessarily predicts a mirror effect for within-list strength manipulations depends on whether the comparison process is category specific. When professions and locations are intermixed on a list, for example, and a profession is presented for an old or new recognition decision, it matters whether the probe item is compared against all of the stored images or is preferentially matched against the profession images. If the matching process is preferential (or, in the extreme, exclusive), then REM predicts a mirror effect when strength is manipulated within list for the same reason it predicts a mirror effect when strength is manipulated between list. If the matching process is not preferential (i.e., if a profession probe is matched against both the profession and location images) then simulations that we performed (see Appendix) suggest that REM can predict the results reported in Experiments 1–3 (i.e., no mirror effect). SLT can also probably accommodate the absence of a mirror effect for within-list strength manipulations in the same way, but we did not actually perform simulations to verify this.

Is the matching process preferential, in which case both REM and SLT predict a mirror effect for the within-list case? The evidence bearing on this question suggests that it is. Ohrt and Gronlund (1999), for example, investigated whether a list length effect would be observed for categories of different lengths that were intermixed in the same list. The list length effect refers to the reliable decrease in memory performance as the length of the study list increases. In Ohrt and Gronlund's (1999) Experiment 2, the study lists consisted of 10 items drawn from one semantic category (e.g., professions) randomly intermixed with 40 items drawn from another (e.g., locations). The question of interest was whether d' for the smaller category was larger than d' for the larger category. If each probe item is matched against all 50 images, regardless of whether it is a profession or location, then no effect on d' should be observed (because the effective list length is 50 for both item types). By contrast, if d' for professions turned out to be lower than that for locations, it would suggest that the matching process is preferential. Because profession probes needed to be matched against only 10 images, whereas locations probes needed to be matched against 40, a list length should be observed. Ohrt and Gronlund (1999) found a large effect on d' as a function of category length, supporting the idea that the global matching process is preferential. Shiffrin et al. (1995) reported similar results for much smaller category lengths manipulated within list even when it was probably not apparent to participants that the list contained multiple items drawn from various semantic categories.

If the global matching process is preferential, as these findings would appear to suggest, then both SLT and REM predict a mirror effect for the within-list strength manipulations used here. Preferential matching would mean that probes from the strong category are matched mainly against strong images, whereas those from the weak category are matched mainly against weak images (just as is true of the between-list situation). Still, it could be argued that because the matching process is preferential, but perhaps not exclusive, SLT and REM predict a mirror effect, but one that is reduced in the within-list case relative to the between-list case (where the matching process is necessarily exclusive). The reduced effect on false-alarm rates as a function of strength for the within-list case may simply have gone undetected in the present series of experiments. Indeed, the results of Experiment 3 suggested that a small (albeit nonsignificant) criterion shift probably did occur.

Conclusion

To conclude, it is important to emphasize what we are not claiming in this article. First, we are not claiming that participants never shift the decision criterion item-by-item in response to within-list manipulations. In fact, they almost certainly do under some conditions (and they may have done so a little in Experiment 3 here). Instead, we are claiming that they appear to be remarkably reluctant to do so even when they know they should, and it would be easy for them to do were they so inclined. Second, we are not claiming that these data falsify likelihood-ratio accounts or that familiarity-based models can explain all of the relevant data. Instead, we are claiming that the evidence reported here weighs against likelihood-ratio models and in favor of the idea that the decision axis is best construed as a strength-of-evidence variable (like familiarity). At the very least, likelihood-ratio models may need to be modified in a way that would not have been necessary had within-list mirror effects been observed.

References

Broderbund, LLC (1997). Broderbund ClipArt 200,000. [Computer Software]. Novato, CA: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) New York: Academic Press.

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

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

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

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

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

Hintzman, D. L., Curran, T., & Caulton, D. A. (1995). Scaling the episodic familiarities of pictures and words. *Psychological Science, 6,* 308–313.

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

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

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760.

Miller, M. B., & Wolford, G. L. (1999). The role of criterion shift in false memory. *Psychological Review, 106,* 398–405.

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

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review, 90,* 316–338.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 855–874.

Ohrt, D. D., & Gronlund, S. D. (1999). List-length effect and continuous memory: Confounds and solutions. In Izawa, C. (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson–Shiffrin model* (pp. 105–125). Mahwah, NJ: Erlbaum.

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

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling the accumulation of partial information. *Psychological Review, 95,* 238–255.

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

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgements in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 294–320.

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

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

Sikstrom, S. (2001). The variance theory of the mirror effect. *Psychonomic Bulletin & Review, 8,* 408–438.

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

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

Wixted, J. T., & Gaitan, S. (in press). Cognitive theories as reinforcement history surrogates: The case of likelihood-ratio models of human recognition memory. *Animal Learning & Behavior*.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107,* 368–376.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1415–1434.

Appendix

REM Simulations

Simulations were performed to determine whether REM predicts a within-list mirror effect for strength manipulations. Except as noted below, the details of the model were arranged to match the model described by Shiffrin and Steyvers (1997). As in their model, words were represented by 20 features each, and those features were determined by drawing values randomly from a geometric distribution with the parameter of that distribution (g_H) set to .45. Two list items were created in this manner (one representing Category A, and the other representing Category B). Two categorical lures were then created by randomly selecting some proportion of the features, *s*, to match the corresponding target and by randomly drawing the other 1-*s* from a geometric distribution with $g_H = .45$. Thus, *s* is a similarity parameter that determines the proportion of lure features that match its corresponding target. If $s_A = 1$, then the lure from Category A would be identical to the target from Category A, and the same would be true for the Category B items if $s_B = 1$. If $s_A = s_B = 0$, then the lures would be no more similar to their corresponding targets than the targets would be to each other.

Note that the creation of categorically similar lures is not something Shiffrin and Steyvers (1997) addressed, so this particular aspect of the simulation process was our own doing. Nevertheless, it seems like a reasonable way to represent similarity in this model. Except for this detail, the simulations closely followed Shiffrin and Steyvers. As in their simulations, the probability of encoding a feature into an image on a given trial, *u**, was set to .04, and the probability of encoding a feature correctly given that a feature was encoded, *c,* was set to .7.

During learning, the weak item (from Category A) was presented 4 times, whereas the strong item (from Category B) was presented 12 times. On the subsequent recognition test, the two targets and their corresponding similar lures were considered, one at a time, for an old or new recognition decision. The recognition test item was compared against a stored image on a feature-by-feature basis. For each comparison, the features either matched or they did not. Either way, two values were computed: the likelihood that the features match (or mismatch) given that the image was created by this test item and the likelihood that the features match (or mismatch) given that this test item did not create the image. The ratio of these two values is the likelihood ratio.

The likelihood of encountering a match for feature *k* given that the image was created by the same item being considered for a recognition decision, $p(m_{k|}\text{same})$, is, in Shiffrin and Steyvers' (1997) original model, equal to:

$$p(m_k|\text{same}) = c + (1-c)\,g(1-g)^{V_{k-1}} \tag{A1}$$

where V_k is the matching feature value (of the *k*th feature). This equation is the probability that the feature was encoded successfully from the test item during learning (*c*) plus $1 - c$ times the probability that the matching value was randomly drawn from the geometric distribution when an encoding error occurred (and so just happens to match coincidentally). The value of *g* was set 0.4, as in Shiffrin and Steyvers (1997).

The probability that the features match given that the image was created by a different item than the one under consideration, $p(m_k|\text{diff})$, is

$$p(m_k|\text{diff}) = g(1-g)^{V_{k-1}}, \tag{A2}$$

which is just the probability that the value was drawn randomly from the geometric distribution used to define the list items. This equation was also used in our simulations when the recognition test item and the stored image were from different categories. When the test item and image were from the same category, then a different equation was needed, one that takes into account the possibility that the features match because the test item and the different item that created the image are categorically similar. The probability that the different (but categorically similar) item that created the image has the same value for feature *k* as the test item is *s* (for the moment, we assume that $s_A = s_B = s$). If it does have the same value for feature *k* because of categorical similarity, then the probability that that feature was successfully encoded into the image is *c*. If it was not successfully encoded (which occurs with probability $1 - c$), then the probability that the features would match by chance anyway is $g(1-g)^{Vk-1}$. On the other hand, if the different item that created the image has a different value for feature *k* as the test item, which occurs with probability $1 - s$, then the probability that the features would match by chance anyway is $g(1-g)^{Vk-1}$. Thus,

$$p(m_k|\text{diff}) = s\,\{c + (1-c)\,g(1-g)^{Vk-1}\} + (1-s)\,g(1-g)^{Vk-1}, \tag{A3}$$

Table A1
Within-List Retrieving-Effectively-From-Memory Simulations

Lures	Hit	FA	*d′*
Dissimilar (A)	.78	.29	1.32
Dissimilar (B)	.79	.28	1.38

Note. Parameters were set to $s_A = 0$, $s_B = 0$, $t_A = 4$, and $t_B = 4$. FA = false alarm.

which can be simplified to

$$p(m_k|\text{diff}) = sc + (1 - sc)\, g(1 - g)^{V_k-1}. \tag{A4}$$

This equation states that, with probability *sc*, the features match because the test item and the image are from the same category and the feature was successfully encoded (and the memory system is aware that *s* is the probability of feature overlap for categorically similar items), and with probability $1 - sc$ the features match because the matching value happened to be drawn randomly from a geometric distribution (with *g* set to 0.4). Thus, for features of the test item and image that happen to match, the likelihood ratio is:

$$\frac{p(m_k|\text{same})}{p(m_k|\text{diff})} = \frac{c + (1 - c)\, g(1 - g)^{V_k-1}}{sc + (1 - sc)\, g(1 - g)^{V_k-1}}. \tag{A5}$$

This is certainly not the only way to modify REM to handle item similarity, but it does seem reasonable in that similarity is captured by degree of feature overlap, and the resulting likelihood function has desirable properties. For example, when $s = 1$ (i.e., when the lures are identical to the targets in every respect), the numerator and denominator are equal (as they should be) and the likelihood ratio is 1.0. When $s = 0$ (i.e., when the items and lures are not at all categorically similar), the equation reduces to that used by Shiffrin and Steyvers (1997) to model recognition memory for random lists of words.

If the feature of the test item that is being compared to a feature of an image happens not to match, then simpler equations apply. If the image was created by the test item but the features do not match, then this must have occurred with probability $1 - c$ (i.e., the probability that an encoding error occurred). If the image was created by a different item, then the denominator is 1.0. Thus, the likelihood ratio is $(1 - c)/1$, or $1 - c$.

When a given recognition test item is evaluated against a stored image, likelihood ratios are computed for each of the 20 features in the manner described above, and they are all multiplied together to yield a final value for that comparison process. The recognition test item is then evaluated against the next stored image to yield another likelihood ratio and so on until the item has been compared with all of the stored images. In our simulations, if the test item and the image with which it was being compared were drawn from different categories, *s* was set to 0. If they were drawn from the same category, *s* was set to 0.3. After all the likelihood ratios were computed for a given test item (one likelihood ratio per image), the values were averaged. If the average value exceeded 1.0, the decision was considered "old," otherwise the decision was considered "new."

Table A2
Within-List Retrieving-Effectively-From-Memory Simulations

Lures	Hit	FA	*d′*
Similar (A)	.67	.30	0.83
Dissimilar (B)	.80	.29	1.39

Note. Parameters were set to $s_A = .3$, $s_B = 0$, $t_A = 4$, $t_B = 4$. FA = false alarm.

Table A3
Within-List Retrieving-Effectively-From-Memory Simulations

Lures	Hit	FA	*d′*
Similar (A)	.62	.30	0.83
Similar (B)	.62	.31	0.81

Note. Parameters were set to $s_A = .3$, $s_B = .3$, $t_A = 4$, and $t_B = 4$. FA = false alarm.

Simulations involved two list items, one for each simulated category (longer lists could have been used, but the outcome would be the same). The recognition test involved four items, the two targets, one representing an old item from Category A and the other representing an old item from Category B, and the two lures, one representing a new item from Category A and the other representing a new item from Category B. All simulations involved 5,000 trials. In the first, *s* was set to 0 for both categories ($s_A = s_B = 0$), which means that the targets and their respective lures were no more related to each other than the two targets themselves, and t_A and t_B (learning trials for the first and second category, respectively) were set to four. The hit and false-alarm rates for this initial simulation in Table A1 show that, before similarity is considered, REM yields typical values (and the numbers provide a basis for comparison with the simulations described next). The results of when *s* was set to 0.3 for the Category A item (such that its corresponding lure shared 30% of its features) and to 0 for the Category B item (such that its corresponding lure only shared features due to coincidental draws from the geometric distribution) in Table A2 show the expected reduction in *d′* when the lures are categorically similar to the targets. This is an expected effect, and it indicates that the *s* parameter is affecting performance in a reasonable way. Table A3 shows the results when both the Category A list item and the Category B list item have similar lures ($s = .3$ for both). The *d′* values for both are reduced relative to the case where $s = 0$ (as in Table A1), which is to be expected.

The next table presents the results of most interest. What happens when the list items from one of the categories is differentially strengthened? This was modeled by increasing t_B to 12, the results of which are shown in Table A4.

Table A4
Within-List Retrieving-Effectively-From-Memory Simulations

Strength	Hit	FA	*d′*
Weak (A)	.55	.24	0.84
Strong (B)	.77	.24	1.45

Note. Parameters were set to $s_A = .3$, $s_B = .3$, $t_A = 4$, $t_B = 12$. FA = false alarm.

Table A5
Within-List Retrieving-Effectively-From-Memory Simulations Involving Exclusive Category-Specific Matching

Strength	Hit	FA	*d′*
Weak (A)	.52	.15	1.09
Strong (B)	.70	.09	1.86

Note. Parameters were set to $s_A = .3$, $s_B = .3$, $t_A = 4$, $t_B = 12$. FA = false alarm.

The predicted pattern corresponds exactly to the pattern we observed in Experiments 1–3. Thus, REM can accommodate our results, but a nonpreferential global matching process must be assumed. That is, in the simulations described above, recognition test items were matched against all of the encoded images, even when the test item and the image were from different categories. If the matching process is instead exclusive (such that recognition test items from Category A are matched only against images created by Category A items), then a different pattern emerges. In that case, a mirror effect is predicted because the situation is no different from a between-list strength situation. Table A5 shows the predicted results assuming an exclusive matching process.

Now, a mirror effect emerges. Even so, the false-alarm rate difference is less than the hit rate difference. Thus, in the within-list case, if the matching process is preferential (and prior evidence suggests that it is), but not exclusive, the predicted false-alarm rate effect would be even smaller than this. Conceivably, the smaller effect on false-alarm rates might go undetected in the kinds of experiments we performed.

Received June 5, 2001
Revision received April 5, 2002
Accepted April 11, 2002 ■