Journal of Experimental Psychology:
Learning, Memory, and Cognition
1998, Vol. 24, No. 6, 1379–1396

0278-7393/98/$3.00

On the Difference Between Strength-Based and Frequency-Based Mirror Effects in Recognition Memory

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

A mirror effect can be produced by manipulating word class (e.g., high vs. low frequency) or by manipulating strength (e.g., short vs. long study time). The results of 5 experiments reported here suggest that a strength-based mirror effect is caused by a shift in the location of the decision criterion, whereas a frequency-based mirror effect occurs although the criterion remains fixed with respect to word frequency. Evidence supporting these claims is provided by a series of studies in which high frequency (HF) words were differentially strengthened (and sometimes differentially colored) during list presentation. That manipulation increased the HF hit rate above that for low frequency (LF) words without selectively decreasing the HF false alarm rate, just as a fixed-criterion account of the word-frequency mirror effect predicts.

In recent years, a well-known empirical regularity known as the *mirror effect* has commanded a great deal of attention. The mirror effect refers to the relationship between hit and false alarm rates in two conditions associated with different levels of recognition accuracy. Specifically, a mirror effect is said to exist when the condition associated with more accurate recognition performance is characterized by both a higher hit rate and a lower false alarm rate than the less accurate condition. This effect is so reliably observed that Glanzer, Adams, Iverson, and Kim (1993) described it as a "regularity of recognition memory."

The mirror effect can be produced by manipulating either the class or the strength of the items presented for study. The most common class manipulation is based on word frequency (high vs. low), whereas strength is usually manipulated by varying study time or number of item presentations. The consistency of the mirror effect across different methods of manipulating recognition accuracy suggests the influence of a single underlying mechanism. In terms of signal detection theory, that mechanism is often thought to be a shift in the criterion for deciding whether to respond "yes" or "no" to a test item.

Strength-Based Mirror Effects

Figure 1 illustrates the criterion-shift argument for a strength manipulation using the standard assumptions of signal detection theory. This model assumes that the decision axis represents a strength-of-evidence variable, such as familiarity. According to this account, the familiarity values associated with the target items and lure items are both normally distributed, with the mean of the target distribution being situated farther to the right on the decision axis than the mean of the lure distribution. In this example, the variances of the target and lure distributions are equal, but in practice they differ somewhat (Ratcliff, Sheu, & Gronlund, 1992). To arrive at a recognition decision, participants are assumed to set a decision criterion somewhere along the familiarity axis. Any test item with a familiarity value exceeding the criterion is judged to be old ("yes"); otherwise the item is judged to be new ("no").

The ideal placement of the decision criterion is midway between the two distributions because that is the point that maximizes the proportion of correct responses. If participants respond in a more or less optimal way, therefore, the criterion will be set farther to the right on the decision axis in the strong condition relative to its placement in the weak condition. This is essentially a formal way to represent the notion that participants in the strong condition appreciate the fact that the studied items will generate a strong sense of prior occurrence when encountered again on the recognition test. For that reason, they require a relatively high level of familiarity before declaring a test item to be old. Participants in the weak condition, by contrast, realize that the target items will not stand out quite as much, so requiring a similarly high degree of familiarity before calling an item old might be counterproductive. Thus, a less stringent requirement is adopted (which means that the criterion is shifted to the left relative to its location in the strong condition). As a result of the leftward criterion shift, the false alarm rate increases. By contrast, the hit rate will decrease compared with the strong condition because the mean of the target distribution shifts to the left twice as far as the criterion does; hence, the mirror effect.

Frequency-Based Mirror Effects

Several theories assume that the same mechanism (viz., a criterion shift) accounts for the mirror effect produced by a word frequency manipulation. In a typical word frequency experiment, participants study lists consisting of a mixture of high- and low-frequency words and then complete a

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

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

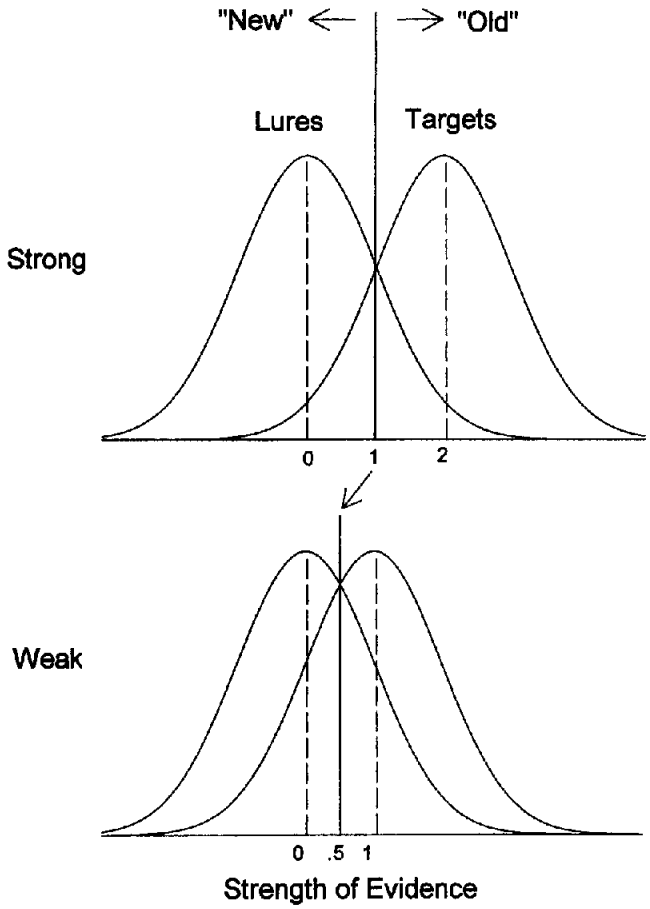

Figure 1. Hypothetical target and lure distributions for weak and strong conditions of a recognition memory experiment. The mean of the lure distribution on the strength-of-evidence axis is arbitrarily set to 0. Also illustrated is a shift in the location of the decision criterion along the strength-of-evidence axis.

yes–no or forced-choice recognition test. Low-frequency (LF) words are almost always associated with better recognition performance (i.e., a higher d') characterized by both a higher hit rate and a lower false alarm rate than high-frequency (HF) words (e.g., Glanzer & Adams, 1985). Typically, theories that attribute frequency-based mirror effects to a criterion shift assume that participants (a) are aware of the HF–LF experimental manipulation, (b) realize that LF words are more memorable than HF words, and (c) use a different decision criterion for each class of word.

Although not presented in terms of signal detection theory, Brown, Lewis, and Monk (1977) were perhaps the first to suggest that a different criterion is used for HF and LF words, thereby accounting for their differing false alarm rates. According to this account, participants appreciate the fact that LF words are more memorable than HF words. During the recognition test, therefore, the absence of a strong sense of prior occurrence for an LF test item suggests that the item did not appear on the list (because such a memorable word would otherwise seem much more familiar than it does). By contrast, if the test item is an HF word, the absence of a strong sense of prior occurrence is less diagnostic (because such a forgettable item may not seem familiar even if it did appear on the list). Thus, an unfamiliar HF lure may be considered old, whereas an equally unfamiliar LF lure would be considered new. In other words, in terms of signal detection theory, the strength-of-evidence distributions for HF and LF lures coincide, but participants use a lower decision criterion for HF words. This model is illustrated in the top panel of Figure 2. This is the same model as that shown in Figure 1, except that a frequency manipulation rather than a strength manipulation is assumed.

In their search of associative memory (SAM) model, Gillund and Shiffrin (1984) explained the mirror effect in a similar way. In their model, both the interitem association parameter (which governs the familiarity of targets) and the residual association parameter (which governs the familiarity of lures) were assumed to be greater for HF words than LF words. According to this model, the two sets of distributions for LF and HF targets and lures are arranged as in the middle panel of Figure 2. To arrive at a decision, the participant is assumed to set a decision criterion on the decision axis above which a "yes" response is given and below which a "no" response is given.

Note that the use of a single decision criterion would not result in a mirror effect. Instead, both the hit and false alarm rates would be greater for HF words (although d' would favor LF words). According to the SAM model, the mirror effect for word frequency emerges because of the following:

> We assume that two criteria are selected by the subject, one for HF words and one for LF words. This assumption is sensible for pure lists, but we assume it holds for mixed lists and mixed tests, as well. Such an assumption could be justified on the basis that subjects can ascertain the word frequency of the tested word and can use this knowledge to set a criterion (Gillund & Shiffrin, 1984, p. 34).

Hirshman (1995) also recently argued that the word-frequency mirror effect can be explained on the basis of a frequency-specific criterion shift.

Although a two-criterion model can explain the word frequency mirror effect, such an account is weakened by evidence suggesting that participants are usually unaware that LF words are more memorable than HF words in a recognition procedure. Accurate knowledge of differential item memorability is required for the criterion to be shifted in the appropriate direction (cf. Hintzman, Caulton, & Curran, 1994). In fact, however, participants usually seem to believe just the opposite. That is, according to participants' estimations, HF words are more memorable than LF words (Greene & Thapar, 1994; Wixted, 1992). Nevertheless, if participants do appreciate the differential memorability of HF and LF words in spite of evidence to the contrary, a frequency-specific criterion shift may play a role in the production of the word frequency mirror effect. A mirror effect produced in part by a shift in the decision criterion will henceforth be referred to as a Type I mirror effect.

Other models account for the word frequency mirror effect without appealing to a criterion shift. According to these models, the mirror effect arises because of the way in which the HF and LF target and lure distributions are

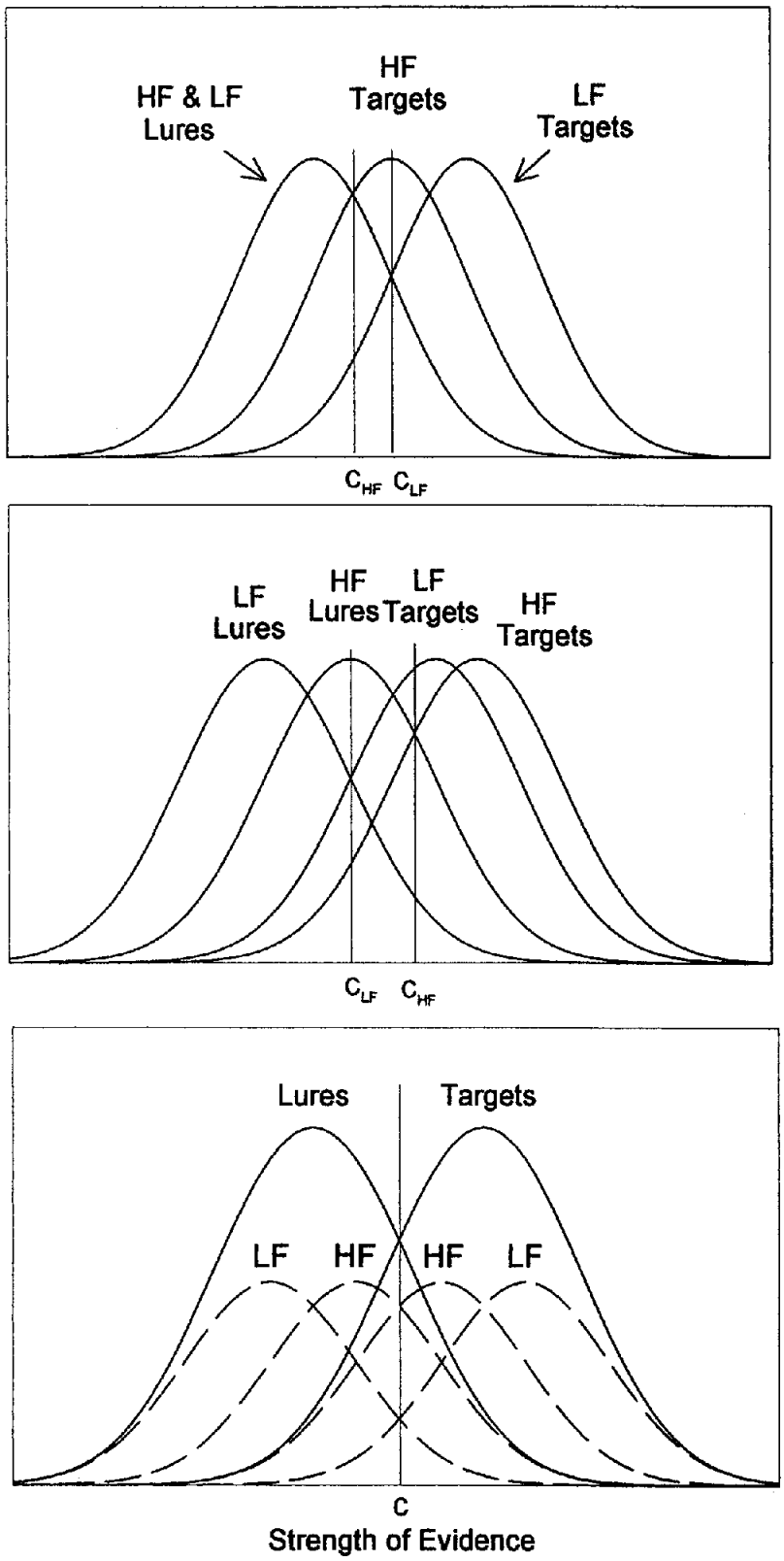

Figure 2. Top panel: Arrangement of frequency-specific target and lure distributions based on a subjective memorability theory advanced by Brown, Lewis, and Monk (1977). Middle panel: Arrangement of frequency-specific target and lure distributions based on the theory advanced by Gillund and Shiffrin (1984). Bottom panel: Arrangement of frequency-specific target and lure distributions that would yield a mirror effect in the absence of a criterion shift. C = criterion; HF = high frequency; LF = low frequency.

arrayed along the decision axis. The bottom panel of Figure 2 illustrates this idea. With respect to the lures, the HF distribution is situated farther to the right on the strength-of-evidence axis than the LF distribution. With respect to the targets, the reverse is true. Glanzer and Bowles (1976) were perhaps the first to propose a model of this kind, but similar models have been proposed by others as well (e.g., Hintzman, 1988). According to Glanzer and Bowles (1976), HF lures are more familiar than LF lures because the former are more likely to be derivatively marked (or activated) by associative spread during list presentation than the latter. LF targets, on the other hand, are more familiar than HF targets because the meaning of an LF word accessed during study is very likely to match the meaning accessed at test. Because HF words typically have more meanings, the meaning accessed during study is less likely to match the one accessed during the recognition test. More recently, Glanzer and Adams (1985, 1990) argued that LF words receive more attention during list presentation than HF words do, which could also account for the stronger sense of prior occurrence for LF words. Whatever the correct explanation might be, if the target and lure distributions are arrayed like those shown in the bottom panel of Figure 2, then a word frequency mirror effect would arise even if the decision criterion remained constant (which is plausible if participants fail to appreciate the word frequency manipulation). Note that the underlying word frequency distributions in the bottom panel of Figure 2 are shown as dashed contours because they are, according to this model, transparent to the participant during a recognition experiment. A mirror effect that occurs in spite of the fact that the decision criterion remains fixed will henceforth be referred to as a Type II mirror effect.

Evidence consistent with the Type II account of the mirror effect can be found in Hoshino (1991) and Hirshman and Arndt (1997). Both of these studies showed that mirror effects do not always arise with respect to word frequency, as they should if participants are always aware of the underlying distributions and adjust their decision criteria accordingly. Hirshman and Arndt (1997), for example, found that when participants were asked to make concreteness ratings for each word during encoding, hit rates for HF and LF words were unexpectedly similar (and in one case were actually higher for HF words), but false alarm rates were still reliably higher for HF words. If participants were aware of the effect of the concreteness rating on the location of the target distributions, they should have adjusted the decision criterion accordingly (which would have lowered false alarm rates for HF words).

On the other hand, the experimental manipulation used by Hirshman and Arndt (1997) was rather subtle. Indeed, the results of the concreteness rating task (viz., the differential strengthening of HF targets) were unexpected even to the experimenters. Thus, it would not be surprising to find that the participants in that experiment were entirely unaware of the memorial consequences of the rating task. If so, then they would not be in a position to adjust the decision criterion appropriately. Instead, they might adjust the decision criterion in the way that they usually do (i.e., they might use a high criterion for LF words and a low criterion for HF words), in which case a higher false alarm rate for HF words would still be expected. Consequently, several of the experiments reported below were designed to selectively strengthen HF targets in ways that were, across experiments, increasingly obvious to the participant.

Likelihood Ratio Models

All of the research reported here was designed to test models like the ones shown in Figures 1 and 2. An important feature of these models is that they all assume a strength-of-evidence decision axis, such as familiarity. By contrast, some theories, like Glanzer's attention/likelihood theory (ALT), assume a log likelihood ratio decision axis. Likelihood ratio models assume that a decision about a recognition test item is not based directly on that item's familiarity but is instead 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 response is "yes"; otherwise the response is "no." Given a log likelihood ratio decision axis, both strength-based and frequency-based mirror effects can occur even though the decision criterion remains fixed across conditions. This is most easily seen by computing beta, a standard measure of response bias, for the two conditions of interest (strong vs. weak or HF vs. LF). Beta is equal to the ratio of the height of the target distribution to the height of the lure distribution at the indifference point, and the log of that value represents the location of the decision criterion on the log likelihood ratio scale. For both the strong and weak conditions in Figure 1, for example, beta is equal to 1 (such that log beta equals 0). Thus, in both conditions, the criterion is placed at 0 on the log likelihood ratio axis. Mirror effects arise because the target and lure distributions are arranged like the model shown in the bottom panel of Figure 2, except that now the decision axis is a log likelihood ratio scale rather than a strength-of-evidence scale (Glanzer & Adams, 1985). Although the main purpose of the research described below is to test models of the mirror effect that assume a strength-of-evidence axis, we also consider the implications of our findings for likelihood ratio models in general and, in a later section, for ALT in particular.

Experiment 1

As a starting point, the first experiment was designed to produce mirror effects in two ways, one based on a strength manipulation and the other based on a frequency manipulation. Strength was manipulated across lists and frequency was manipulated within lists. More specifically, each participant studied one strong list (in which words were presented three times each) and one weak list (in which words were presented once each). For both lists, half the words were HF and the other half were LF. These manipulations (strength and frequency) produced nearly identical looking mirror effects, but we will argue that those effects arise for entirely different reasons.

Method

Participants. The participants were 36 undergraduates of the University of California, San Diego who were enrolled in a lower division psychology course. Participation in the experiment satisfied a course requirement.

Materials and design. In this and the following experiments, a pool of words to be used as targets and lures was compiled from word norms (Nelson, McEvoy, & Schreiber, 1994) that provide information on word frequency from Kučera and Francis (1967). The word pool consisted of 216 words, of which 108 were LF and 108 were HF. LF words used in the following experiments occurred from 0 to 3 times per million, whereas HF words occurred 40 times or more per million. The means (and standard deviations) for the word frequency conditions were as follows: For LF words, $M = 1.6$ ($SD = 1.0$); for HF words, $M = 98.9$ ($SD = 65.7$). Of the 108 words on each list, 96 were randomly chosen, for each participant, for each of two memory tests. Of the 96 words used on each test, 48 were targets and 48 were lures of which half also were LF and half HF.

A second pool of words compiled directly from the Kučera and Francis (1967) norms contained words of mixed frequencies with no replications of the original 216-item word pool. This second pool was used for a distractor task that was performed between list presentation and the recognition test.

Procedure. Participants were tested individually. After reading and signing an informed consent form, participants were seated in front of a computer. Each participant completed the experiment alone in the room without distraction. Participants were first given a brief reaction time (RT) task to familiarize them with the use of the right and left mouse buttons. This task consisted of 20 trials in which an arrow appeared in the center of the screen. If the arrow pointed to the left, the participant was to press the left mouse button, and if it pointed to the right, the participant was to press the right mouse button.

After the RT task, participants were presented with a list of words in one of two encoding conditions. A random half of the participants received the weak encoding condition first and the strong second, and the other half received the reverse order. In the weak encoding condition, 48 targets from the word pool (half LF and half HF) were presented once, one at a time, for 500 ms with an interstimulus interval (ISI) of 250 ms, in a random order that was unique for each participant. In the strong encoding condition, the target words were presented three times each, with each presentation occurring randomly throughout the list. Therefore, in this condition, participants were shown 144 (48×3) word presentations. Presentation time and ISI were the same as in the weak encoding condition. Participants were instructed to read each word aloud as it appeared on the screen.

After list presentation, participants were given a short distractor task in which words from the distractor pool were presented (spelled backwards) at the center of the screen, one at a time for 500 ms with an ISI of 500 ms. Participants were instructed to pronounce the words as they would be pronounced if spelled in the correct, forward order. The distractor task, which ran for 20 s, was designed to minimize the contribution of short-term store.

Immediately after the distractor task, participants were given a yes–no recognition test. They were informed that the test would be timed and that they should respond as quickly as possible without sacrificing accuracy. They were also informed that they would be asked to make a Remember–Know judgment for each yes response and to make a confidence rating on a 1 to 5 scale (*complete guess* to *absolutely certain*) for each response. The recognition test consisted of the 48 targets randomly intermixed with the remaining 48 words from the word pool for this list as lures (half of which were HF and half of which were LF).

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

Results and Discussion

Table 1 shows the mean d' scores for each condition. For each participant, a d' value was computed based on that participant's hit and false alarm rates. On those occasions when a hit rate was 1.0 or a false alarm rate was 0, values of $1 - 1/N$ or $1/N$ (respectively) were used instead, where N refers to the maximum possible hits and false alarms for a given condition (Macmillan & Creelman, 1991). The mean d' scores in Table 1 show that the strength manipulation and word frequency manipulation both had the expected effects. An analysis of variance for d' showed significant main effects of encoding strength, $F(1, 34) = 58.25$, $MSE = .439$, and word frequency, $F(1, 34) = 176.02$, $MSE = 0.274$. The interaction between frequency and strength was not significant, $F(1, 34) < 1$, $MSE = 0.242$ (all statistical analyses used an alpha level of .05).

Mean hit and false alarm rates for both encoding conditions are also shown in Table 1. Overall, the hit rate was higher and the false alarm rate was lower in the strong encoding condition compared with the weak encoding condition. Similarly, the hit rate was higher and the false alarm rate was lower for LF words compared with HF words. In short, clear mirror effects were obtained for both encoding strength and word frequency.

Statistical analyses supported the aforementioned observations. An analysis of variance for hits showed significant main effects of strength of encoding, $F(1, 34) = 27.13$, $MSE = 0.040$, and word frequency, $F(1, 34) = 62.63$, $MSE = 0.033$, with no significant interaction between encoding condition and word frequency, $F(1, 34) = 0.13$,

Table 1
Recognition Accuracy (d'), Hit Rates (Hits), and False Alarm Rates (FA) for the Weak and Strong Conditions of Experiment 1

Frequency and dependent measure	Weak	Strong	Mean
High frequency			
d'	0.96	1.60	1.28
Hits	.568	.696	.632
FA	.255	.184	.219
Low frequency			
d'	1.82	2.37	2.10
Hits	.742	.859	.801
FA	.157	.124	.140
Mean			
d'	1.39	1.99	1.69
Hits	.655	.777	.716
FA	.206	.154	.180

MSE = 0.016. An analysis of variance for false alarms showed a significant main effect of strength of encoding, $F(1, 34) = 5.62$, $MSE = 0.035$, and a significant effect of word frequency, $F(1, 34) = 40.10$, $MSE = 0.011$, with no significant interaction between encoding condition and word frequency $F(1, 34) = 2.98$, $MSE = 0.009$. Note that the order in which participants experienced the two between-list conditions (strong vs. weak) was included as a between-subjects factor in all statistical analyses. No significant effects of order were obtained in this experiment or in the following ones.

Finally, as would be expected, the proportion of "Remember" responses was higher in the conditions generating more accurate recognition performance. In the strong condition, 82% of the hits were Remember responses, whereas the corresponding value in the weak condition was 67%. Similarly, in the LF condition, 84% of the hits were Remember responses whereas the corresponding value in the HF condition was 67% (cf. Gardiner & Java, 1990).

Assuming a strength-of-evidence decision axis, the mirror effect produced by the between-list strength manipulation presumably occurred because of a shift in the location of the decision criterion. That is, because the lures in the two cases were physically identical (i.e., in both cases the lures consisted of words drawn randomly from the word pool), their familiarity characteristics were presumably the same. If so, then the differing false alarm rates arose because of a shift in the decision criterion, as depicted in Figure 1. Such a shift would make sense because, following a strong list, participants presumably realize that the target items will generate a strong sense of prior occurrence on the subsequent recognition test. Thus, they can avoid making false alarms while still responding correctly to most of the targets by setting a high criterion relative to the weak condition. Of course, it is possible that the differing false alarm rates in the strong and weak conditions occurred because the strength manipulation influenced the familiarity of the lures. However, recent evidence presented by Shiffrin, Huber, and Marinelli (1995), which is considered in more detail later, supports the intuitively appealing idea that strengthening targets does not affect the familiarity of the lures.

The main question of interest is whether the word-frequency mirror effect arises for the same reason (viz., a conscious shift in the decision criterion); that is, perhaps participants realize that LF words are more memorable than HF words. During the recognition test, therefore, they use a high criterion for LF words for the same reason that they use a high criterion during the recognition test that follows the strong list. If so, then the word frequency mirror effect would be classified as a Type 1 mirror effect as well. In Experiment 2, we tested the criterion-shift account by differentially strengthening HF words during list presentation.

Before turning to that experiment, we briefly consider how the results of Experiment 1 might be interpreted in terms of a likelihood ratio model. As indicated earlier, both strength-based and frequency-based mirror effects, such as those shown in Table 1, can be explained without assuming a criterion shift if we drop the assumption of a strength-of-evidence axis and instead assume a log likelihood ratio decision axis. The location of the decision criterion on this scale can be estimated by computing log beta, where beta is the standard measure of bias mentioned earlier. With the group hit and false alarm rates shown in Table 1 to illustrate this point, beta (which ideally would equal 1.0) is equal to 1.28 and 1.25 in the HF and LF conditions, respectively, and 1.29 and 1.26 in the weak and strong conditions, respectively; the corresponding log beta values are 0.25, 0.22, 0.26, and 0.23, respectively. Thus, although participants exhibited a slight "no" bias throughout, the location of the decision criterion on the log likelihood ratio decision axis was always about the same and was always close to 0. As described in detail elsewhere (e.g., Glanzer & Adams, 1985), the mirror effect arises because of the relative locations of the target and lure distributions, not because of a criterion shift.

Experiment 2

Experiment 2 was similar in certain respects to an experiment recently reported by Hirshman and Arndt (1997). During encoding, participants in Hirshman and Arndt's experiment were instructed to rate the concreteness of the items presented for study. For reasons that are not entirely clear, that manipulation differentially strengthened HF targets such that the HF hit rate equaled or exceeded the LF hit rate. If the word frequency mirror effect usually occurs because of an adjustment of the decision criterion, then, when the usual disadvantage for HF targets is overcome by some experimental manipulation, the decision criterion should be adjusted upward (and the false alarm rate for HF words should decrease). This prediction is illustrated in the upper panel of Figure 3. Note that this is essentially the same model as that shown in the upper panel of Figure 2, except that the HF decision criterion is now situated to the right of the LF criterion because the HF targets were differentially strengthened. Contrary to this prediction, Hirshman and Arndt found the usual false alarm rate effect (i.e., HF words still had a higher false alarm rate than LF words).

However, as indicated earlier, the experimental manipulation used by Hirshman and Arndt (1997) was a subtle one. The effect of the concreteness-rating task on the HF hit rate was not anticipated by the experimenters and was probably not noticed by the participants. If participants did not realize that HF targets were selectively strengthened by the concreteness manipulation, then they would not be in a position to adjust the decision criterion in the manner illustrated in the upper panel of Figure 3. Instead, they might adjust the decision criterion as they always do (i.e., using a high criterion for LF words and lower criterion for HF words), in which case a higher false alarm rate for HF words would still be expected.

In Experiment 2, we used a procedure that would also differentially strengthen HF targets, but the manipulation was intentionally conspicuous. During list presentation, participants studied a list consisting of an equal number of HF and LF words. In one condition, the HF and LF words were each presented once (as usual), followed by a yes–no

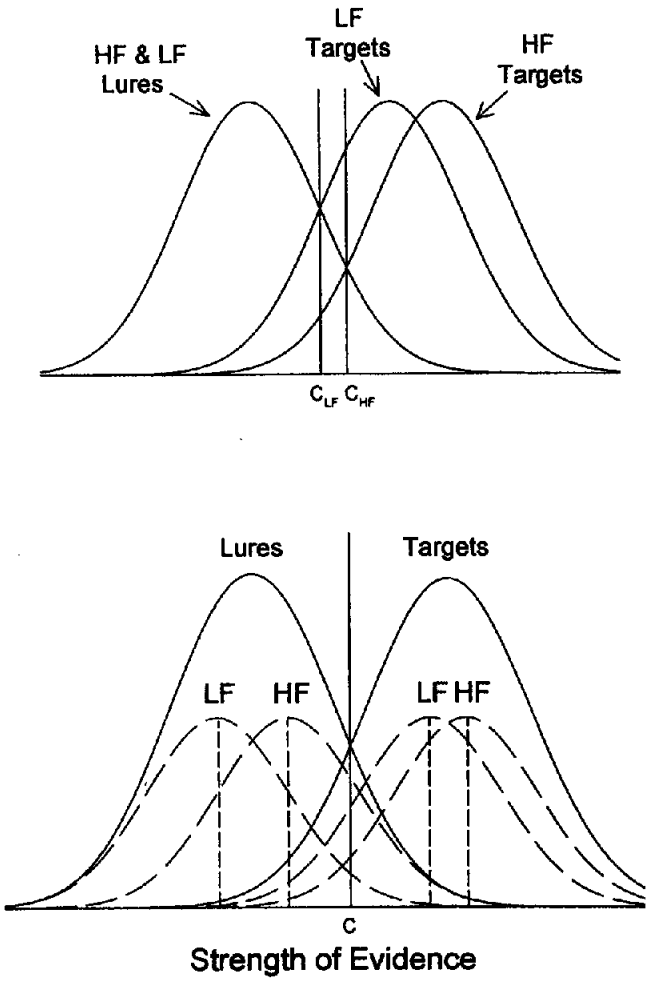

Figure 3. Upper panel: The effect of differentially strengthening HF targets on the HF target distribution and the HF decision criterion according to a Type I account of the word frequency mirror effect. Lower panel: The effect of differentially strengthening HF targets on the HF target distribution according to a Type II account of the word frequency mirror effect. C = criterion; HF = high frequency; LF = low frequency.

recognition test. In another condition, the HF words were differentially strengthened during list presentation by presenting them 5 times each (whereas the LF words were only presented once each). If participants use a different decision criterion for LF and HF words, one would expect that by increasing the hit rate for HF words above that for LF words in an obvious way, the false alarm rate for HF lures would be lower than the false alarm rate for LF lures (thereby preserving the mirror effect). As illustrated in the upper panel of Figure 3, this is a straightforward prediction made by the criterion-shift (i.e., Type I) model of the word-frequency mirror effect. On the other hand, if the word-frequency mirror effect is a Type II mirror effect (bottom panel of Figure 2), then the differential strengthening of HF targets should not result in a lower false alarm rate for HF words relative to LF words. The effect of differentially strengthening the HF targets according to a Type II account is illustrated in the lower panel of Figure 3. Note that, in this figure, the HF target distribution is situated to the right of the LF target distribution because the HF targets were differentially strengthened. This is essentially the same single-criterion model as that shown in the bottom panel of Figure 2, except that now the HF targets are more familiar than the LF targets.

Note that, collapsed across frequency, Experiment 2 can be construed as involving a strength manipulation. Even though only HF words are strengthened, overall encoding strength in the strong condition (collapsed across word frequency) should be greater than overall strength in the weak condition. Thus, if the mirror effect is preserved, the increased hit rate in the strong condition should be associated with a decreased false alarm rate relative to the weak condition (in which all items were presented only once), as in Experiment 1.

Method

Participants. The participants were 36 undergraduates of the University of California, San Diego who were enrolled in a lower division psychology course.

Materials and design. The materials used in this experiment were identical to those used in Experiment 1.

Procedure. The weak and strong conditions were also the same except that in the strong condition, HF targets were presented five times each, randomly intermixed with the LF targets, which were presented only once each. Otherwise, the procedure was identical to that used in Experiment 1.

Results and Discussion

Table 2 shows the mean d' scores for Experiment 2. Collapsed across the word frequency manipulation, the strong condition (in which HF words were selectively repeated) resulted in a higher mean d' than the weak condition (1.95 vs. 1.44, respectively). Within strength conditions, the mean d' for LF words exceeded that for HF

Table 2
Recognition Accuracy (d'), Hit Rates (Hits), and False Alarm Rates (FA) for the Weak (HF × 1) and Strong (HF × 5) Conditions of Experiment 2

Frequency and dependent measure	Weak	Strong	Mean
High frequency			
d'	0.98	1.94	1.48
Hits	.546	.832	.689
FA	.229	.208	.219
Low frequency			
d'	1.90	1.95	1.93
Hits	.760	.718	.739
FA	.157	.103	.130
Mean			
d'	1.44	1.95	1.70
Hits	.653	.775	.714
FA	.193	.156	.175

Note. HF = high frequency.

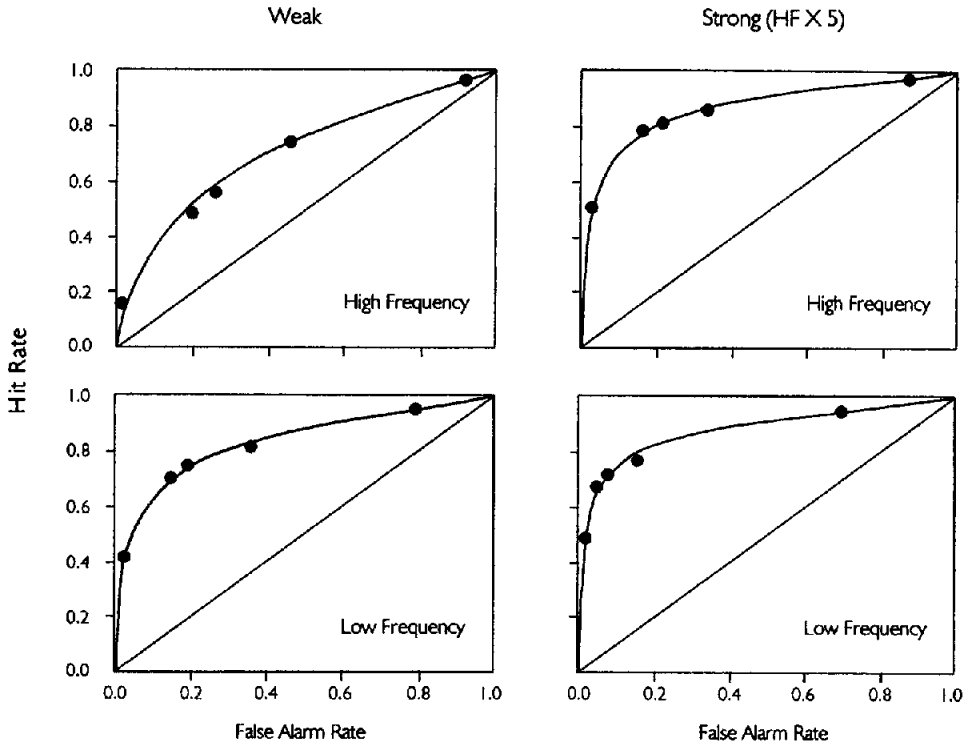

Figure 4. Left panel: Confidence-based receiver operating characteristics (ROCs) for the HF and LF words from the weak (i.e., standard) condition of Experiment 2. Right panel: Confidence-based ROCs for the HF and LF words from the strong condition of Experiment 2 (in which HF targets were differentially strengthened). HF = high frequency; LF = low frequency.

words in the weak condition (as usual), but the values were nearly identical in the strong condition. An analysis of variance for d' supported these observations. A significant main effect was obtained for encoding condition, $F(1, 34) = 49.24$, $MSE = 0.370$, and word frequency, $F(1, 34) = 61.66$, $MSE = 0.252$, and the interaction between condition and word frequency was also significant, $F(1, 34) = 34.05$, $MSE = 0.425$.

Table 2 also shows the hit and false alarm rates for both conditions. Note that the hit and false alarm rates in the standard weak condition exhibit the usual word-frequency mirror effect. That is, in the weak condition, LF words are associated with a higher hit rate and lower false alarm rate than HF words. Quite a different pattern emerges in the strong condition in which the hit rate for HF words was higher than for LF words, but the false alarm rate for HF words was still greater than that for LF words (i.e., no mirror effect was observed). Indeed, the false alarm rates (with respect to word frequency) remained similar to those in the weak encoding condition. Note that, while differentially strengthening HF words did not *selectively* lower the false alarm rate for HF words, the overall false alarm rate collapsed across frequency did decrease.

An analysis of variance for hits revealed a significant main effect of strength of encoding, $F(1, 34) = 42.35$, $MSE = 0.025$, a main effect of word frequency, $F(1, 34) = 8.83$, $MSE = 0.021$, and a significant interaction between strength of encoding and word frequency, $F(1, 34) = 72.13$, $MSE = 0.027$. These analyses merely underscore the fact that, relative to HF words, the hit rate was higher for LF words in the weak condition and lower in the strong condition. An analysis of variance for false alarms revealed a nearly significant main effect of encoding strength, $F(1, 34) = 4.12$, $MSE = 0.024$, $p = .0504$, a main effect of word frequency, $F(1, 34) = 51.31$, $MSE = 0.011$, and, importantly, no significant interaction between encoding strength and word frequency, $F(1, 34) = 1.49$, $MSE = 0.013$.

The confidence-based receiver operating characteristics (ROCs) were also examined to determine if the selective strengthening of HF targets distorted the usual curvilinear function. The left two panels of Figure 4 show the HF and LF ROCs for the weak condition, and the right two panels show the corresponding ROCs for the strong condition (in which the HF targets were differentially strengthened). For the strong condition, the curves are essentially identical, which is what would be expected if the strengthening manipulation merely shifted the HF target distribution to the right on the strength-of-evidence axis (without otherwise distorting the distribution).[1] The confidence-based ROC data

[1] We thank Thomas Nelson for pointing out the importance of this analysis.

from all of the experiments reported in this article were also examined, and all were equally unremarkable.

The most important finding from this experiment was the fact that the HF false alarm rate remained greater than the LF false alarm rate even when HF target words were strengthened to the point that the HF hit rate was greater than the LF hit rate. Although this finding is just what a Type II account of the word frequency mirror effect would predict (as illustrated in the lower panel of Figure 3), it seems more difficult to reconcile with what the Type I account predicts (as illustrated in the upper panel of Figure 3). If participants usually adjust the decision criterion as a function of word frequency because they realize that LF words are more memorable than HF words, why would they fail to adjust the criterion in the opposite direction when HF words are differentially strengthened in an obvious way?

The pattern of results shown in Table 2 is similar to a pattern recently reported by Maddox and Estes (1997). In their experiment, participants were preexposed to novel items (e.g., random syllable or digit triads). Half the items were seen only once during the preexposure phase, and the other half were seen several times each. Thus, the former items were analogous to LF words, and the latter items were analogous to HF words. Following the preexposure phase, some of the preexposed items were presented on a study list. A rapid rate of presentation was used in an effort to minimize differential attention to the LF and HF items. On the subsequent recognition test, the HF items exhibited both a higher hit rate and a higher false alarm rate than the LF items. Presumably, this occurred because HF targets and lures were more familiar than their LF counterparts, and a single decision criterion was in effect (as in the lower panel of Figure 3). Thus, whereas Maddox and Estes (1997) overcame differential attention by rapid presentation (cf. Hoshino, 1991), we compensated for differential attention by presenting the HF words multiple times during list presentation (and achieved the same result).

The findings reported here appear to be inconsistent with a new theory of the mirror effect advanced by Greene (1996). According to this account, the word frequency mirror effect is an uninteresting by-product of the fact that participants attempt to distribute their "yes" responses evenly between HF and LF words during a recognition test. More specifically, participants are assumed to appreciate the fact that the list consists of an equal mix of HF and LF words. On the subsequent recognition test, therefore, they are assumed to reserve half their "yes" responses for HF words and half for LF words. If half the test items are HF words and half are LF words (as is usually the case), then a response strategy of this kind would necessarily produce a mirror effect. However, this account also predicts that the selective strengthening of HF words should serve merely to improve the accuracy of those "yes" responses reserved for HF words. That is, the HF hit rate should increase and the HF false alarm rate should decrease, but the overall probability of saying "yes" to a HF word should not change. Instead, the results of Experiment 2 suggest that the HF hit rate increases without a corresponding decrease in the false alarm rate (such that considerably more than 50% of the "yes" responses are made to HF words).

One way to reconcile this finding with Greene's (1996) theory is to assume that the strengthening manipulation caused participants to mistakenly believe that the list (and, therefore, the subsequent recognition test) involved many more HF words than LF words. If participants tried to match the proportion of "yes" responses to the perceived proportion of HF and LF words, then they would say "yes" more often to HF words than to LF words. Whether or not participants were actually mistaken about the proportion of HF and LF words in Experiment 2 is unknown. However, results to be discussed later (specifically, results from Experiments 4 and 5) weigh against this interpretation.

Earlier, we noted that a mirror effect produced by a criterion shift on the strength-of-evidence axis (e.g., Figure 1) corresponds to a fixed criterion on the log likelihood ratio decision axis (i.e., the criterion remains fixed at a point close to zero, but the mirror effect occurs anyway). Here, the situation is reversed. The absence of a mirror effect for word frequency in the strong condition of Experiment 2 suggests that the criterion is fixed on the strength-of-evidence axis as a function of word frequency (as in the lower panel of Figure 3). By contrast, if a log likelihood ratio decision axis is assumed, the criterion must shift. Once again, this is most easily seen by computing log beta for the LF and HF words in both the weak and strong conditions. For the weak condition in which neither LF nor HF words were strengthened, the log beta values based on the group hit and false alarm rates were 0.26 and 0.27 for LF and HF words, respectively. That is, as in Experiment 1, the criterion was at approximately the same place on the log likelihood ratio axis for both LF and HF words. For the strong condition (as illustrated in the lower panel of Figure 3), by contrast, beta is greater than 1 for LF words (such that log beta is positive) and less than 1 for HF words (such that log beta is negative). The actual log beta values computed from the group hit and false alarm rates in this condition were 0.63 and -0.13, which is to say that the LF and HF criteria were located at different points on the log likelihood ratio axis. Why participants might choose to use the same criterion for HF and LF words in the weak condition but different criteria in the strong condition (i.e., a "no" bias for LF words and a "yes" bias for HF words) is not clear. However, as discussed in more detail later, one version of Glanzer's ALT can accommodate this finding.

Assuming a strength-of-evidence decision axis, the data from Experiment 2 are consistent with the predictions of a Type II account of the word frequency mirror effect. According to this account, the criterion is fixed with respect to word frequency in both the weak and strong conditions. Nevertheless, a slight modification of the Type I (criterion-shift) account can explain the results as well. In Experiment 2, we attempted to differentially strengthen HF targets in a way that was unlikely to be missed by a participant who appreciates the HF versus LF manipulation (as a Type I account of the word-frequency mirror effect assumes). However, it is possible that participants did not consider the difference between HF and LF words during study but

distinguished between them only during the recognition test phase. That is, during study, participants may have only realized that some words were strengthened (without appreciating the fact that the HF words in particular were being strengthened). On the subsequent recognition test, however, they may have assessed the memorability of each test item in the usual way. That is, they may have adopted a relatively high criterion for LF words (because they realized those words are quite memorable) and a relatively low criterion for HF words (for the opposite reason). If so, the HF words would have a higher false alarm rate (because of a criterion shift) even though the HF words were strengthened.

Experiment 3

Experiment 3 was designed to make the differential strengthening manipulation even more obvious to the participant. In both conditions of Experiment 3, HF words were differentially strengthened by presenting them five times each (whereas the LF words were presented only once). However, in one condition of Experiment 3, the LF words (both targets and lures) were presented in one color, such as green, and the HF words (both targets and lures) were presented in another color, such as red. In the other condition, all words were presented in a single color (white). These conditions will be referred to as *cued* and *uncued*, respectively.

In the cued condition of Experiment 3, participants were provided with all of the information needed to make a frequency-specific adjustment of the decision criterion. Even if participants failed to realize that the HF words were strengthened, they presumably could not have missed the fact that the red words were presented five times each, whereas the green words were presented only once. Thus, when confronted with a red (HF) item on the subsequent recognition test, participants should be able to use their knowledge of the fact that the red words were selectively strengthened and adjust their decision criterion accordingly (thereby lowering the FA rate for the red HF words). That is, the participant should use a conservative criterion for the HF (i.e., red) words in the cued condition of Experiment 3 (as illustrated in the upper panel of Figure 3).

Method

Participants. The participants were 36 undergraduates of the University of California, San Diego who were enrolled in a lower division psychology course. Participation in the experiment satisfied a course requirement.

Materials and design. The words were drawn from the same source (and in the same way) as Experiment 1.

Procedure. Except for the difference in word color, both conditions of Experiment 3 were identical to the strong condition of Experiment 2; that is, in both conditions, the HF words were presented five times each during list presentation (randomly distributed throughout the list), and LF words were presented only once each. In one condition (the cued condition), the HF targets and lures were presented in one color (either red or green), and the LF targets and lures were presented in another color (green or red, depending on the color of the HF words). In the other condition (the uncued condition), all of the words were presented in the same color (white). Thus, the uncued condition of Experiment 3 was identical in every respect to the strong condition of Experiment 2. A random half of the participants received the cued condition first and the uncued condition second, whereas the other half received the reverse order (although no order effects were observed). Each list was followed by a 20-s distractor and yes–no recognition test as in Experiment 2.

Results and Discussion

Table 3 shows the mean d' scores for Experiment 3. In both the cued and the uncued conditions, d' for LF words slightly exceeded that for HF words in spite of the fact that HF words were differentially strengthened by presenting them five times each. The difference, however, was not quite significant, $F(1, 34) = 3.32$, $MSE = 0.329$, $p = .077$.

Table 3 also shows the hit and false alarm rates for HF and LF words in both conditions. The pattern is the same whether or not the words were differentially colored. In both cases, the hit rate for HF words exceeded that of the LF words, and the false alarm rate for HF words also exceeded that of LF words. Statistical analyses revealed a main effect of word frequency on hits, $F(1, 34) = 12.81$, $MSE = 0.022$, and false alarms, $F(1, 34) = 30.94$, $MSE = 0.023$. The effect of condition was not significant in either case, and no interactions involving the condition factor approached significance.

These findings replicate the results of Experiment 2. Strengthening the HF words to the point that they were associated with a significantly higher hit rate than LF words did not reverse the usual false alarm rate advantage for LF words even when the words were colored in such a way as to make a differential criterion shift rather easy. Thus, it appears that participants did not take advantage of the color information provided them.

These findings are not easily reconciled with a Type I (criterion shift) account of the word-frequency mirror effect. According to the Type I account, participants use a relatively

Table 3
Recognition Accuracy (d'), Hit Rates (Hits), and False Alarm Rates (FA) From Experiment 3 for the Noncued Differential-Strength Condition and the Differential-Strength Condition Cued by Color

Dependent measure	Noncued	Cued	Mean
High frequency			
d'	1.67	1.58	1.63
Hits	.768	.764	.766
FA	.241	.262	.252
Low frequency			
d'	1.76	1.74	1.75
Hits	.721	.684	.703
FA	.164	.141	.153
Mean			
d'	1.72	1.66	1.69
Hits	.745	.724	.734
FA	.203	.202	.202

high criterion for LF words because they realize those words are likely to be recognized had they appeared on the list. A lower criterion is used for HF words because participants also realize that these words might not seem terribly familiar even if they had appeared on the list. If so, it is hard to imagine why they would not take advantage of the further information provided by correlating the color of the word with its encoding condition. A Type II account of the word frequency mirror effect, on the other hand, is not challenged by these findings.

Experiment 4

The fact that participants did not appear to adjust the decision criterion within a recognition test even when provided with the information needed to do so is surprising. By adjusting the decision criterion on the basis of color information, they would have been able to respond in a more optimal way than is possible when only a single decision criterion is used throughout.

In Experiments 1 and 2, the results suggested that participants do indeed shift the decision criterion as a function of strength *between* lists. That is, following a strong list, a relatively high decision criterion is used (and it remains fixed throughout the recognition test). Following a weak list, a lower decision criterion is used, and it too remains more or less fixed throughout the recognition test. Within a list, however, participants appear to be reluctant to shift the decision criterion on an item-by-item basis (at least according to the results of Experiment 3).

On the other hand, in the color-cued condition of Experiment 3, participants were faced with competing sources of information. That is, the items that were differentially strengthened were precisely the items that ordinarily are more difficult to recognize (viz., HF words). Conceivably, that conflict caused participants to disregard the color information during the recognition test in favor of the frequency information. To investigate this possibility, Experiment 4 also involved a differential strength manipulation, but the manipulation was no longer frequency specific. That is, in one condition, half the words were strengthened by presenting them five times each, and the other half were not. The strong words, which now comprised both HF and LF words, were presented in one color, and the weak words, which also comprised HF and LF words, were presented in another color. On the subsequent recognition test, the targets were presented in the same color as they appeared on the list. In addition, half the lures were presented in the same color as the strong targets and half in the same color as the weak targets. Note that, other than a difference in color, the lures in the strong and weak conditions were physically identical (i.e., a random half were red, the other half were green). Thus, any difference in the false alarm rate to red and green words would be most easily explained on the basis of a criterion shift occurring during the recognition test itself.

It is easy to imagine why participants might use a different decision criterion for red and green words. That is, just as they presumably do for a between-list strength manipulation, participants faced with a within-list strength manipulation might decide to use a high criterion for words they know would be strong had they appeared on the list (e.g., red words) and a lower criterion for words they know would be relatively weak even if they had appeared on the list (e.g., green words). On the other hand, if participants are reluctant to shift the decision criterion on an item-by-item basis during a recognition test, which is what the results of Experiment 3 suggest, then the strengthening manipulation should produce a large effect on the hit rate without affecting the false alarm rate. This prediction is illustrated in Figure 5 (which is similar to Figure 1, except that the criterion does not shift as a function of strength). Because a single decision criterion is in effect, no false alarm rate difference should be observed.

Method

Participants. The participants were 36 undergraduates of the University of California, San Diego who were enrolled in a lower division psychology course. Participation in the experiment satisfied a course requirement.

Materials and design. The words were drawn from the same source (and in the same way) as in the previous experiments.

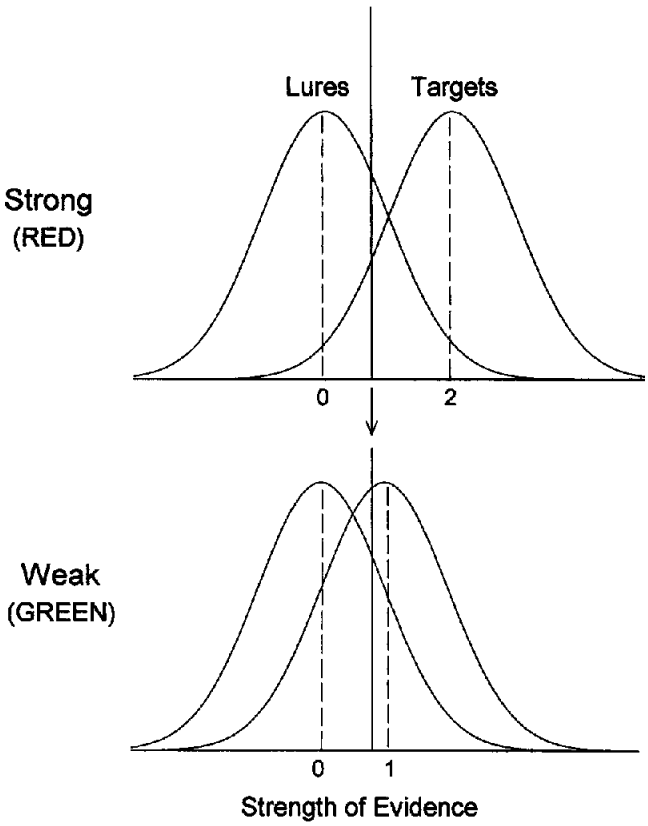

Figure 5. Effect of strengthening half the items in a list assuming the decision criterion remains constant across conditions. This example also assumes that the items in the strong condition, presented in red, are perceptually distinct from those of the weak condition, presented in green (otherwise a criterion shift on the basis of strength would be impossible).

Procedure. As in the preceding experiments, participants studied two lists under two conditions, and recognition memory was tested after each. One condition of Experiment 4 was identical to the cued condition of Experiment 3. That is, HF words were presented in one color (and were differentially strengthened by presenting them five times each during list presentation), and LF words were presented in another color. For half the participants, the HF words were red and LF words were green. For the other half, the colors were reversed. The other condition of Experiment 4 was similar except that the strengthening manipulation was not correlated with word frequency. That is, half the words were differentially strengthened by presenting them five times each (half of these were HF words, the other half LF words), and half were presented only once (again, half of these were HF words, the other half LF words). The strong words were always presented in one color (e.g., red), and the weak words in another color (e.g., green). In addition, half the lures were presented in red and the other half in green, and in both cases half were HF words and half were LF words. Thus, in this condition, the lures were physically identical. In all other respects, the procedure was the same as that used in the previous experiments. The two conditions of this experiment will be referred to as the correlated mixed-strength (HF words selectively presented five times each) and uncorrelated mixed-strength conditions (half the words presented five times each).

Results and Discussion

Table 4 shows the mean d' scores for both conditions of Experiment 4 along with the corresponding hit and false alarm rates. The results from the correlated mixed-strength condition replicate the results of Experiment 3 in every respect. More specifically, the differential strengthening of HF targets resulted in a d' score approximately equal to that produced by LF words. Nevertheless, and in spite of the fact that HF and LF words were presented in different colors, HF words were associated with a significantly higher hit rate and a significantly higher false alarm rate. An analysis of variance performed on the data from this condition revealed a main effect of word frequency for hits, $F(1, 34) = 22.42$, $MSE = 0.022$, and false alarms, $F(1, 34) = 13.72$, $MSE = 0.035$. These findings are most easily reconciled with a theory that assumes a fixed decision criterion.

The results from the uncorrelated mixed-strength condition reveal that strengthening half the words on the list resulted in a significantly higher d' for those words (not surprisingly). More important, the strong words were associated with a significantly higher hit rate relative to the weak words, but the false alarm rates were about the same. That is, whereas a criterion shift model would predict a significantly lower false alarm rate for lures presented in the same color as the strong targets, those lures were actually associated with a slightly higher false alarm rate (though the difference in false alarm rates did not approach significance). An analysis of variance performed on the data from this condition revealed a main effect of strength on hits, $F(1, 34) = 32.89$, $MSE = 0.021$, no main effect of strength on false alarms, $F(1, 34) = 0.71$, $MSE = 0.011$, and, unexpectedly, a significant interaction between strength and frequency, $F(1, 34) = 4.97$, $MSE = 0.008$. This interaction reflects the fact that, in the strong condition, the false alarm rate to LF lures decreased slightly, whereas for HF lures it increased slightly (an effect that did not replicate in the next experiment).

Table 4
Recognition Accuracy (d'), Hit Rates (Hits), and False Alarm Rates (FA) From Experiment 4 for the Differential and Mixed-Strength Conditions

Condition	d'	Hits	FA
Correlated mixed-strength list			
HF × 5	1.66	.774	.247
LF × 1	1.65	.656	.132
Uncorrelated mixed-strength list			
Weak × 1	1.24	.578	.180
Strong × 5	2.01	.821	.195

Note. HF = high frequency; LF = low frequency.

The findings from the uncorrelated mixed-strength condition are especially surprising. When strength is manipulated across lists, the strong list is associated with both a higher hit rate and a lower false alarm rate (as in Experiments 1 and 2). The lower false alarm rate presumably arises because participants in the strong condition appreciate the fact that, during the recognition test, the targets will seem quite familiar relative to lures. Thus, a relatively high level of familiarity is required before a test item is declared to be old (i.e., a conservative criterion is used). In the weak condition, participants are also presumably aware of the fact that targets will not seem so familiar, so they do not have the luxury of using such a conservative decision criterion. The use of a more liberal criterion increases the false alarm rate (see Figure 1). The same arguments apply to the within-list strength manipulation when color information identifies items as belonging to one condition or the other. That is, just as a participant in a between-list strength experiment knows that strong targets will seem quite familiar relative to lures on the recognition test, participants in the within-list strength experiment know that strong targets (identified by color) will seem quite familiar. Thus, the same kind of criterion shift should be observed. In contrast to this prediction, the data suggest that no criterion shift takes place.

Why are participants reluctant to shift the decision criterion within a list but not between lists? The data do not speak directly to this issue, but one explanation might be that participants are reluctant to expend the mental energy required to shift the decision criterion on an item-by-item basis. That is, in the between-list strength experiment, participants need only set a decision criterion once following the weak list and once again following the strong list. In the within-list strength experiment, on the other hand, many more shifts in the decision criterion would be needed. During the recognition test of Experiment 4, test items were sometimes red and sometimes green (and they alternated in more or less random fashion). Each time a red item appeared, one criterion would need to be set. If the next item happened to be green, a different criterion would be needed. When the next red item appeared, the original criterion would need to be reestablished (and so on throughout the recognition test). This kind of moment-by-moment shift in

the decision criterion may be something participants are not inclined to do.

The results from Experiment 4 are also not in line with predictions based on Greene's (1996) response strategy account of the mirror effect. According to that theory, participants divide their "yes" responses evenly between the two word classes (in this case, red and green words). If so, a mirror effect for strength should have been observed. These findings also weigh against an alternative version of Greene's theory that was advanced to explain the results of Experiment 2 (according to which participants mistakenly believe that the repeated items constitute more than 50% of the items on the list). If participants mistakenly believed that the repeated red items were more numerous than their nonrepeated green counterparts, and if they devoted more "yes" responses to red words as a result, then the increased hit rate should have been accompanied by an increased false alarm rate (just as was true of HF words in Experiment 2). Instead, the strengthening manipulation selectively increased the hit rate for red words.

These findings are also not easily explained by a model that assumes a likelihood ratio decision axis. For the false alarm rate to remain more or less constant across conditions, the criterion must have shifted on the log likelihood ratio axis as a function of strength. Using the group hit and false alarm rates to illustrate this point, beta for the weak condition was 1.50 (log beta was 0.40) and for the strong condition was 0.95 (log beta was −0.05). Thus, assuming a log likelihood ratio decision axis, the criterion was placed at 0.40 for the weak condition (producing a relatively strong "no" bias) and −0.05 for the strong condition (producing a slight "yes" bias). Why participants would be biased in different ways when strength is manipulated within list (but not between list) is not easily explained. Instead, it seems simpler to assume that participants merely used a single decision criterion situated along a strength-of-evidence axis for all of the test items. What, then, does the change in beta (a standard measure of response bias) represent? According to this view, bias changed in a descriptive sense only. No change in bias actually occurred in a psychological sense (cf. Hirshman & Arndt, 1997). That is, the participants did not change their decision strategy at all as a function of strength (the way they might if, say, payoffs for correct "yes" and "no" responses were changed). Instead, they evaluated the familiarity of every item (strong or weak) against a single fixed decision criterion.

Experiment 5

The next experiment was a replication of Experiment 4, except that the instructions provided to the participants were much more explicit. Although the participants in Experiment 4 were told that some of the words on the list would be strengthened by repetition, they were not told that the strong words would be presented in one color and the weak words in another color. We assumed that this would be obvious to the participant (and informal postsession interviews confirmed this assumption). Nevertheless, in light of our relatively surprising findings, we decided to spell out all of the details of the procedure to the participant in advance of the experiment. Under such conditions, participants would be in the best possible position to make a strength-specific criterion shift during the recognition test. Thus, the only significant difference between the uncorrelated mixed-strength condition of Experiment 4 and Experiment 5 was that, in the latter experiment, participants were told in advance in which color the strengthened targets would be presented and in which color the nonstrengthened targets would be presented. In addition, participants were informed that on the subsequent recognition test, half of the lures would be presented in the same color as the strengthened targets, and half would be presented in the same color as the nonstrengthened targets.

Method

Participants. The participants were 31 undergraduates of the University of California, San Diego who were enrolled in a lower division psychology course. Participation in the experiment satisfied a course requirement.

Materials and design. The words were drawn from the same source (and in the same way) as in the previous experiments.

Procedure. The procedure was the same as that used in the uncorrelated mixed-strength condition of Experiment 4 with the following exceptions: Participants studied one list of words consisting of 104 targets, they were not given a reaction time test prior to the experiment, no Remember–Know judgments were obtained, and participants were not given an opportunity to change their minds once a decision was made. As in Experiment 4, half the items were strengthened by repetition (and were presented in one color) and half were not strengthened (and were presented in another color). On the recognition test, half the lures were presented in the same color as the strong targets, and half were presented in the same color as the weak targets.

The instructions to the participants were particularly detailed and included examples of what the list would look like and what the recognition test would be like. The instructions were as follows:

> In this experiment, you will first see a long list of words (104 words in all). The words will be presented one at a time on the computer screen. After the words are presented, a recognition test will be given. During this test, words will again be presented one at a time and, for each one, you will be asked to decide whether or not it appeared on the list. Some of the words will have appeared on the list (so the answer would be "yes") and others will not have appeared on the list (so the answer would be "no").
>
> The list words will be presented at a fairly rapid rate, and you should read each word ALOUD as it is presented. You will notice that some of the list words will be displayed in red letters and others will be displayed in green letters. The red words will each be presented several times (scattered throughout the list) and the green words will be presented only once. On the recognition test that follows the list, you will have to decide which red words appeared on the list (and which red words did not) and which green words appeared on the list (and which green words did not). Click on OK for a short example of what the list will look like . . .

The sample session consisted of 10 target items (half strong in red, half weak in green) and, on the subsequent recognition test, those 10 targets were randomly intermixed with 10 lures (5 of which were red and 5 of which were green). After answering any questions the participant might have, the experiment proper began.

Results and Discussion

Table 5 shows the results of this experiment (d', hit rate, and false alarm rate). As expected, overall performance was better for LF words than for HF words and better for strong words than for weak words. An analysis of variance on d' scores revealed a main effect of word frequency, $F(1, 30) = 77.38$, $MSE = 0.156$, and a main effect of strength, $F(1, 30) = 89.25$, $MSE = 0.222$. The interaction between strength and frequency was not significant, $F(1, 30) = 0.77$, $MSE = 0.142$.

The hit rates shown in Table 5 were also unsurprising. LF words were associated with a higher hit rate than HF words, $F(1, 30) = 17.77$, $MSE = 0.009$, and strong words were associated with a higher hit rate than weak words, $F(1, 30) = 118.96$, $MSE = 0.008$. The interaction between strength and frequency was not significant, $F(1, 30) = 0.46$, $MSE = 0.006$.

The false alarm rates shown in Table 5 exhibit the usual word frequency effect (viz., a higher false alarm rate for HF words), but the rates were nearly identical as a function of strength. An analysis of variance performed on these data revealed a main effect of word frequency, $F(1, 30) = 30.69$, $MSE = 0.006$. The effect of the strength manipulation was not significant, $F(1, 30) = 0.88$, $MSE = 0.013$, and neither was the interaction between strength and frequency, $F(1, 30) = 0.046$, $MSE = 0.006$.

These findings suggest that even when participants are apprised of all of the details of the experiment in advance, they do not necessarily make a strength-specific adjustment of the decision criterion during the recognition test. If they did, the higher hit rate in the strong condition should have been accompanied by a lower false alarm rate in that condition. It should be acknowledged that, unlike in Experiment 4, the false alarm rate in the strong condition of Experiment 5 was slightly lower than that of the false alarm rate in the weak condition (which is what the criterion-shift account predicts), but the result was not significant. Conceivably, a small criterion shift did occur in this experiment, a shift that would have been detected if power were increased.

Table 5
Recognition Accuracy (d'), Hit Rates (Hits), and False Alarm Rates (FA) From Experiment 5 for the Word Frequency and Nondifferential-Strength Manipulations

Frequency and dependent measure	Weak	Strong	Mean
High frequency			
d'	1.59	2.33	1.96
Hits	.703	.884	.793
FA	.183	.167	.175
Low frequency			
d'	2.16	3.02	2.59
Hits	.783	.945	.864
FA	.107	.085	.096
Mean			
d'	1.88	2.68	2.28
Hits	.743	.915	.829
FA	.145	.126	.136

Even if that were true, however, it is worth noting that the effect of the strength manipulation on d' in this experiment was actually larger than produced by the between-list strength manipulations in Experiments 1 and 2. Thus, if anything, the criterion-shift account predicts that a larger difference in false alarm rates should be observed in this experiment. Instead, the false alarm rate difference was considerably smaller and was not significant.

As shown in Table 5, the word frequency manipulation (which affected performance over essentially the same range as the strength manipulation) produced a large effect on the false alarm rate. A criterion-shift account of this finding would need to explain why participants shifted the decision criterion on the basis of word frequency within a list but not on the basis of strength within a list. This would be especially surprising given that, if anything, discriminating HF from LF words (which must occur if the criterion is adjusted in a frequency-specific way) is more difficult than discriminating, say, the red–strong condition from the green–weak condition. A simpler account of the present findings holds that participants do not ordinarily shift the decision criterion within a list for either strength or frequency manipulations and that the difference in HF and LF false alarm rates arises for reasons other than a criterion shift.

General Discussion

The five experiments reported here were designed to address a simple question: Does a shift in the decision criterion play a role in producing the word-frequency mirror effect, as several theories assume it does (Brown et al., 1977; Gillund & Shiffrin, 1984; Hirshman, 1995)? If so, one would expect to find that differentially strengthening HF words during list presentation would decrease the false alarm rate for HF words below that for LF words (the opposite of what is usually observed). The results reported here suggest that this does not happen. Instead, when the HF hit rate exceeds the LF hit rate because of differential strengthening, the HF false alarm rate remains significantly higher than the LF false alarm rate. Surprisingly, this holds true even when HF and LF words are presented in different colors, which gives the participant unambiguous perceptual information that can be used to establish different criterion settings for HF and LF words. This finding suggests that participants may be disinclined to shift the decision criterion on an item-by-item basis during a recognition test even when conditions for doing so are optimal. Why, then, do HF words have a higher false alarm rate than LF words? Presumably because LF lures generate a lower sense of prior occurrence (e.g., they are less familiar) than HF lures, as a Type II account of the word-frequency mirror effect assumes (lower panel of Figure 2).

The story may be somewhat different for the strength-based mirror effect. When strength is manipulated across lists (as in Experiment 1), a mirror effect is usually observed. That is, not only is the hit rate significantly higher in the strong condition, but the false alarm rate is significantly lower as well. Note that, unlike the lures in a word-frequency manipulation, the lures in a strength manipulation

are physically identical on average (i.e., in both conditions, the lures are randomly selected items that did not appear on the list). Thus, the lower false alarm rate associated with the strong condition presumably arises because participants, quite reasonably, use a high criterion following a strong list and a lower criterion following a weaker list (i.e., this pattern is a Type I mirror effect). To do that, participants need only set the decision criterion twice, once following the strong list and once again following the weak list. When strength is manipulated within list, and condition is cued by perceptual information (such as color), however, no mirror effect is observed. Instead, the hit rate in the strong condition exceeds that of the weak condition, but the false alarm rates are about the same. Again, this appears to suggest that participants are reluctant to shift the decision criterion moment-to-moment during the recognition test.

Subjective Memorability

All of these findings weigh against the idea that, during the recognition test, participants consciously assess the memorability of each test item and set the decision criterion accordingly (Brown et al., 1977).[2] That is, with respect to word frequency, some have argued that participants realize that LF words are more memorable than HF words and therefore require a higher sense of prior occurrence for LF words than HF words before declaring the item to be old. If so, LF words would be associated with a lower false alarm rate. Prior research already posed some difficulty for this account because participants do not seem to appreciate the fact that LF words are more memorable than HF words in the first place (Greene & Thapar, 1994; Wixted, 1992). Hintzman et al. (1994) also presented evidence against the subjective memorability model. Instead of evaluating the memorability of each item during the recognition test and setting the decision criterion accordingly, participants appear to set the criterion once based on more global properties of the list and use it throughout the recognition test.

The results of Experiments 4 and 5 suggest that participants also fail to use memorability information that is available to them when strength (rather than frequency) is manipulated within a list. In these experiments, some items in the list were strengthened by repetition, and others were not. If participants used word color to assess memorability and adjusted the decision criterion accordingly, then the false alarm rate to red words should have been less than that for green words. More specifically, participants should have required a high sense of prior occurrence for red words (because they know that those words would seem very familiar had they appeared on the list) and a lower sense of prior occurrence for green words (because those words might not seem very familiar even if they did appear on the list). Evidence for such behavior was not obtained in either experiment even though participants almost certainly had accurate knowledge of which words would be more memorable had they appeared on the list.

Although participants apparently do not use word memorability to adjust the location of the decision criterion on an item-by-item basis during the recognition test, they do appear to use global information about the strength of the list to set the decision criterion initially. That is, following a strong list, participants presumably realize that the target items will seem quite familiar and therefore set a relatively high decision criterion. Once set, it presumably remains there except for, perhaps, random fluctuation (and perhaps some drift during the course of testing). Following a weak list, a different criterion is used, which also remains essentially fixed throughout the recognition test. Because a less stringent criterion is used in the weak condition, a higher false alarm rate is observed. Thus, subjective memorability does affect the placement of the decision criterion, but only when memorability is manipulated between lists (thereby minimizing the frequency with which the criterion must be shifted).

Prior Research Involving Differential Strength Manipulations

Two prior studies have shown that the location of the target distribution can be shifted upward without producing a corresponding change in the location of the decision criterion. In a study reported by Shiffrin et al. (1995), participants studied a single long list of words composed of items drawn from many different semantic and orthographic categories (21 categories in all). The various categories differed in length (i.e., number of exemplars drawn from a category) and strength (number of exemplar presentations). The unprovable but seemingly reasonable assumption was that, with so many different categories, participants would adopt a single decision criterion and use it throughout (rather than adopting a different decision criterion for each of the 21 categories). Given that assumption, any change in the category-specific false alarm rate as a function of category strength or category length could be attributed to changes in the properties of the lure distribution. However, while the hit rate increased with category strength (obviously), the false alarm rate was unaffected by that manipulation. That is, the false alarm rate to categories involving strong targets was about the same as the false alarm rate to categories involving weak targets. Thus, when steps are taken to decrease the likelihood that participants will shift the decision criterion as a function of strength, the false alarm rate does indeed remain constant for both strong and weak lists. Our research suggests that the criterion also tends to remain fixed even when steps are taken to facilitate (rather than impede) a strength- or frequency-specific criterion shift within a list.

The findings reported by Shiffrin et al. (1995) also suggest that strengthening target items does not affect the characteristics of the lure distribution. If, for example, strengthening target items increased or decreased the mean familiarity of the corresponding lures, then (assuming the criterion remained fixed) the false alarm rate should have either increased or decreased as a result. Instead, the false alarm

[2] Our results do not rule out the possibility (or even speak to the issue) of an *unconscious* criterion shift accounting for the word-frequency mirror effect.

rate remained constant. Thus, the change in false alarm rates that occurs when strength is manipulated between lists (as in Experiment 1) appears to be due to a shift in the location of the decision criterion (and not because of a change in the characteristics of the lures across conditions).

Hirshman and Arndt (1997) also showed that the target distribution can be affected without a corresponding change in the location of the decision criterion. Specifically, when participants were asked to rate words for concreteness during list presentation, the hit rate for HF words selectively increased (sometimes beyond that of LF words) for reasons that are not entirely clear. In spite of that selective increase, the false alarm rate for HF words did not selectively decrease (i.e., it remains higher than that of LF words). This should not happen if participants possess full knowledge of the locations of the target and lure distributions and set the decision criterion accordingly. On the other hand, participants may have been entirely unaware of the fact that a concreteness rating selectively increased the strength of HF words. If that were the case, they would not be in a position to adjust the decision criterion in the appropriate way. Our research shows, however, that even when the selective strengthening of HF words is conspicuous, participants still do not change the criterion in effect for HF words. Experiment 5 contained some evidence that participants might do so when the details of the experiment were clearly explained to the them at the outset (i.e., the false alarm rates change in the appropriate direction under those conditions), but the effect was not significant.

The Possibility of Within-List Criterion Shifts

The findings reported here should not be taken to imply that participants never shift the decision criterion within a list. For example, Wixted (1992) used lists consisting of HF, LF, and rare words. The rare words used in that experiment were so rare that they were tantamount to nonwords. Almost certainly, participants discriminated these words from the others on the list. However, no mirror effect for rare words was observed relative to HF words. Instead, rare words were associated with a higher hit rate and a higher false alarm rate. This is the kind of effect one would expect if (a) participants *did* shift their decision criteria for these words and (b) they did so in an inefficient way because of a mistaken idea about the memorability of those words. That is, participants rated those words as being very low in memorability when in fact they were almost as memorable as LF words. Perhaps because of this, they adjusted their decision criterion for rare words downward (the kind of manipulation that would make sense for words that were difficult to remember), thereby increasing both the hit rate and the false alarm rate. Conceivably, participants fail to adjust the criterion within a list when the two classes involve semantically similar items. When they are sufficiently dissimilar (e.g., words vs. nonwords, words vs. pictures, etc.), within-list criterion shifts may readily occur. Thus, our point is not that such criterion shifts never occur. Rather, our findings suggest that they do not readily occur when word frequency or strength is manipulated within a list (but the word-frequency mirror effect occurs anyway).

Glanzer's Attention Likelihood Theory

Although the studies were not designed specifically to test this theory, the findings presented here bear on what is probably the best known account of the mirror effect, Glanzer's attention likelihood theory (ALT). ALT holds that participants make recognition decisions on the basis of likelihood ratios instead of on the basis of a decision criterion situated along a strength-of-evidence axis. When the likelihood that a test item was drawn from the target distribution exceeds the likelihood that it was drawn from the lure distribution, the participant responds "old"; otherwise, the response is "new." To compute such a likelihood, the memory system must know the properties of both the target and lure distributions (mean, variance, and mathematical form).

In accounting for the word-frequency mirror effect, the theory usually assumes that participants are aware of the locations of the HF and LF target and lure distributions and that they compute likelihood ratios on the basis of this information (although these computations are not necessarily assumed to take place on a conscious level). Table 6 shows hypothetical data generated by ALT. The mathematical details of the theory are presented in Glanzer et al. (1993), and the same parameter values used there to illustrate the theory were used here as well. More specifically, for the standard condition (in which HF and LF words were each presented once), p(New), the proportion of word features already marked prior to study, was set to 0.10; n(B), the number of HF features sampled during study or test, was set to 40; and n(A), the number of LF features sampled during study or test, was set to 60. Note that the theory clearly predicts a word-frequency mirror effect (because the same parameter values were used in our calculations, these are the same hit and false alarm rates reported by Glanzer et al., 1993). Table 6 also shows what this version of the theory predicts when HF words are differentially strengthened and participants possess accurate knowledge of the locations of the HF and LF target and lure distributions. These predictions were obtained by increasing n(B) from 40 to 120 during study (i.e., a greater number of HF features were sampled during list presentation) and decreasing it again to 40 during the recognition test. In contrast to what we found,

Table 6
Hit Rates (Hits) and False Alarm Rates (FA) Predicted by Glanzer's Attention Likelihood Theory for the Standard Condition and Strengthened HF Target Condition

	Standard condition		Strengthened HF targets	
Word frequency	Hits	FA	Hits	FA
High frequency	.650	.371	.867	.206
Low frequency	.724	.248	.724	.248

Note. Rates were predicted assuming that participants were aware of the word frequency manipulation. HF = high frequency.

the theory predicts that a mirror effect should still be observed. That is, an increase in the HF hit rate should be accompanied by a selective decrease in the false alarm rate. This does not occur even when perceptual information is provided to make such a criterion shift as easy as possible. Thus, this version of ALT is ruled out by the present findings.

Although the predictions of ALT are usually computed assuming full knowledge of the locations of the HF and LF target and lure distributions, Glanzer et al. (1993) point out that ALT still predicts a word-frequency mirror effect even if participants are not aware of that experimental manipulation. Table 7 shows the predicted hit and false alarm rates under these conditions. Under this version of the model, participants are assumed to evaluate the likelihood ratio for each test item based on target and lure distributions that are the average of the corresponding HF and LF distributions. Again, because we used the same parameters as Glanzer et al. (1993), the values shown in the table for the standard condition are the same as those reported by Glanzer et al. Indeed, for the standard condition, the predicted hit and false alarm rates are unaffected by whether or not participants appreciate the HF–LF manipulation. Also shown in Table 7 are the predictions of this version of ALT when the HF words are differentially strengthened. Note that the theory now predicts the observed pattern of results: Increasing the hit rate of HF words above that of LF words does not change the fact that the HF false alarm rate is higher than the LF false alarm rate. Thus, this version of ALT (viz., the one that assumes that participants are not aware of the HF–LF manipulation) is consistent with the findings reported here.

One feature of the results reported in this article is not easily reconciled with even this version of ALT. As indicated above, if participants are not aware of the HF–LF manipulation, both the HF hit rate and the HF false alarm rate are predicted to exceed the corresponding values for LF words when HF words are differentially strengthened. However, our results show that the same result occurs even when participants are provided with unambiguous perceptual information that can be used to make the discrimination. If participants can readily compute likelihood ratios for individual items on the basis of their subjective understanding of the location of the relevant target and lure distributions, why would they choose to ignore important information about where those distributions might be? That is, when faced with, say, a red test item, why would participants use the same estimates of the locations of the target and lure distributions to compute a likelihood ratio as they would when faced with a green test item when they know perfectly well that a red item, if it appeared on the list, was presented five times (which corresponds to a target distribution high on the evidence axis)? This aspect of our findings seems difficult to explain if recognition memory operates as ALT assumes, but the data do not necessarily refute ALT. Conceivably, a version of this theory will assume that perceptual information about the locations of the relevant target and lure distributions is ignored by participants even though other kinds of information is not (e.g., information about word frequency).

Table 7
Hit Rates (Hits) and False Alarm Rates (FA) Predicted by Glanzer's Attention Likelihood Theory for the Standard Condition and Strengthened HF Target Condition

	Standard condition		Strengthened HF targets	
Word frequency	Hits	FA	Hits	FA
High frequency	.650	.371	.867	.206
Low frequency	.724	.248	.589	.142

Note. Rates were predicted assuming that participants were unaware of the word frequency manipulation. HF = high frequency.

Conclusion

Several models assume that a mirror effect arises when one condition yields a higher d' than another condition because of a change in the location of the decision criterion. This is assumed to be true whether memory performance is affected by an encoding strength manipulation or a word frequency manipulation. However, our findings suggest that although a criterion shift may occur in the former case (so long as strength is manipulated between lists), it does not occur in the latter. Therefore, mirror effects can arise for at least two reasons. For one kind of mirror effect (Type I), a criterion shift plays a role in that it accounts for the difference in false alarm rates in the two conditions. In the other kind of mirror effect (Type II), the criterion is fixed and the difference in false alarm rates in the two conditions arises because the means of the two lure distributions differ on the strength-of-evidence axis.

References

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency, and negative recognition. *Quarterly Journal of Experimental Psychology, 29,* 461–473.

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

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

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

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

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

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 21–31.

Greene, R. L. (1996). Mirror effect in order and associative information: Role of response strategies. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 687–695.

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

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

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 275–289.

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

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

Hoshino, Y. (1991). A bias in favor of the positive response to high-frequency words in recognition memory. *Memory & Cognition, 19,* 607–616.

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

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

Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 539–559.

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

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

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

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

Received April 15, 1997
Revision received March 19, 1998
Accepted March 19, 1998 ■