Journal of Experimental Psychology:
Learning, Memory, and Cognition
1992, Vol. 18, No. 4, 681-690

0278-7393/92/$3.00

Subjective Memorability and the Mirror Effect

John T. Wixted
University of California, San Diego

The mirror effect refers to the common finding that hit and false alarm rates on a recognition test are inversely related. The present research investigated the generality of the mirror effect (to rare words) and tested whether the effect might be grounded in accurate estimates of word memorability. The first 2 experiments showed that although high- and low-frequency words exhibit a mirror effect, rare words do not. Furthermore, contrary to expectations, Ss consistently (and mistakenly) predicted that memorability was directly correlated with frequency of usage. These findings weigh against the idea that the mirror effect arises because of a S's ability to reject low-frequency lures on the grounds that such words would have been remembered had they appeared previously. Instead, the rejection of lures from different frequency categories may be determined by their semantic or phonemic overlap with list targets, and an analysis along these lines may help to explain why rare words constitute an exception to the otherwise ubiquitous mirror effect.

The *mirror effect* is an increasingly well-established recognition phenomenon that refers to the parallel relationship between a subject's ability to correctly classify previously seen and unseen items. In general, conditions that facilitate the correct identification of "old" items also facilitate the correct rejection of "new" items (Glanzer & Adams, 1985). The best example of this phenomenon can be found in studies concerned with recognition memory for high- and low-frequency words. In these studies, low-frequency words are almost always associated with higher hit rates and lower false alarm rates than high-frequency words (e.g., Glanzer & Bowles, 1976; Rao & Proctor, 1984).

One intuitively appealing model of the mirror effect is illustrated in Figure 1. This figure depicts hypothetical familiarity distributions for both high- and low-frequency words under two conditions. The two distributions on the left correspond to words that did not appear on the list (new words) and the two on the right correspond to words that did appear on the list (old words). With regard to new items, the familiarity of low-frequency words is presumably less than that of their high-frequency counterparts. However, according to several theories (e.g., Glanzer & Bowles, 1976; Mandler, 1980), these distributions are reversed for old items. As a result, low-frequency words will be easily recognized when they are old (and therefore very familiar) and easily rejected when they are new (and therefore very unfamiliar).

Glanzer and Bowles (1976) conducted a particularly detailed analysis of the model depicted in Figure 1. In this experiment, subjects studied lists of high- and low-frequency words followed by a two-alternative, forced-choice recognition test involving all possible combinations of old and new items. In agreement with the familiarity model, they found that performance was best on trials involving a choice between old and new low-frequency words (L+ vs. L−, respectively) and worst on trials involving a choice between old and new high-frequency words (H+ vs. H−, respectively). Performance on mixed trials (H+ vs. L− or L+ vs. H−) was intermediate, presumably because of the intermediate separation between the relevant familiarity distributions.

Despite its intuitive appeal, recent research conducted by Glanzer and Adams (1990) casts some doubt on a familiarity-based account of the mirror effect. In one of their experiments, Glanzer and Adams presented subjects with a list of words, half of which were spelled in forward order and half of which were spelled in reversed order. In a subsequent yes/no recognition test, the reversed words were associated with higher hit rates and lower false alarm rates than the untransformed words (i.e., the mirror effect was obtained). The authors argued that this result cannot be accommodated by simple strength theories, such as those based on familiarity, because the effect was evident within a given stimulus class (e.g., low-frequency words). Under these conditions, lures from either condition (i.e., spelled in forward or backward order) should be equally familiar and, therefore, equally likely to occasion false alarms.

Although it might be possible to defend a pure strength theory even in this case, Glanzer and Adams (1990) prefer an alternative explanation of why negative recognition (the correct rejection of lures) mirrors positive recognition. Their theory is rooted in an idea first espoused by Brown (1976) and is based on the notion of subjective memorability (cf. Gentner & Collins, 1981). In its simplest version, the theory holds that subjects are aware of the fact that certain items (e.g., low-frequency words) are more memorable than other items (e.g., high-frequency words). On a recognition test, lures judged to be memorable can be correctly rejected on the grounds that they would have been remembered had they actually appeared on the list (not because they are unfamiliar). Lures judged to be less memorable are correspondingly more difficult to confidently reject because they may have appeared

I thank Julie Dea, Patricia DeAlva, Doug Sheres, and Greg Winterstein for their assistance in data collection and Thomas Nelson, John Brown, and an anonymous reviewer for their insightful comments.

Correspondence concerning this article should be addressed to John T. Wixted, Department of Psychology, C-009, University of California, San Diego, La Jolla, California 92093.

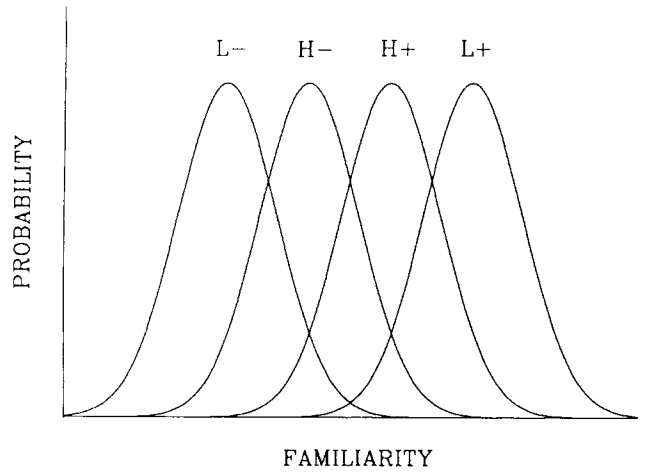

Figure 1. Hypothetical familiarity distributions for old and new low-frequency words (L+ and L−, respectively) and old and new high-frequency words (H+ and H−, respectively).

on the list and been forgotten. Thus, false alarm rates for memorable words should be lower than those for nonmemorable words.

Although this account seems plausible, direct evidence that the mirror effect is grounded in an accurate subjective analysis of word memorability is lacking. Moreover, the familiarity-based analysis depicted in Figure 1, which does not assume knowledge of memorability, is consistent with the large majority of studies relevant to the mirror effect. Therefore, the present research was designed to evaluate the viability of both the familiarity-based and subjective memorability accounts of the mirror effect. With regard to the familiarity account, the first two experiments examined memory for high frequency, low frequency, and rare words. On the basis of a model such as that shown in Figure 1, one might expect to find relatively few false alarms for rare (and, therefore, very unfamiliar) words and many false alarms for high-frequency (and, therefore, familiar) words. Moreover, if the mirror effect held, then one might also expect to find correspondingly high hit rates for rare words and low hit rates for high-frequency words. Surprisingly, the results of two experiments instead showed that rare words were associated with high false alarm rates, and they did not exhibit a mirror effect with respect to high-frequency words.

The last three experiments investigated whether the obtained pattern of results could be explained on the basis of subjective memorability. That subjects might be able to correctly predict the memorability of high-and low-frequency words is not an unreasonable hypothesis. Most adults have had substantial experience with even low-frequency words, and it would not be surprising to discover that they have learned something about the memorability of words that differ in frequency of usage. Moreover, the substantial literature on metamemory suggests that people often correctly predict what they are likely to remember on a later memory test (e.g., Nelson, 1988). However, with regard to rare words, relevant experience is quite limited and subjective memorability estimates may therefore be considerably off target. If the mirror effect arises because of accurate subjective memorability estimates for high- and low-frequency words, it may also be undermined by inaccurate subjective memorability estimates for rare words.

Memory for Rare Words

Previous research on the subject of memory for rare words has been consistent in one respect, namely, that recognition memory for rare words is less accurate than that for low-frequency words (Mandler, Goodman, & Wilkes-Gibbs, 1982; Rao & Proctor, 1984; Schulman, 1976; Zechmeister, Curt, & Sebastian, 1978). However, in other respects relevant to the model depicted in Figure 1, the findings have been less consistent. In agreement with a familiarity-based account, for example, Mandler et al. (1982) found that false alarm rates for extremely rare words were lower than the rates observed for both high- and low-frequency words. This result would be expected if the familiarity distribution for new rare words was located to the far left in Figure 1. For the mirror effect to emerge, the familiarity distribution for old rare words would need to fall to the far right in Figure 1. Instead, Mandler et al. (1982) found that the hit rate for rare words was lower than that for both high- and low-frequency words.

Other studies concerned with memory for rare words have produced results that are less contrary to the mirror effect, but that appear to conflict with the idea that subjects respond on the basis of familiarity per se. For example, Rao and Proctor (1984) found that the false alarm rate for rare words exceeded that for low-frequency words and approached that of high-frequency words. If subjects were responding on the basis of item familiarity alone, such a result would seem to imply that extremely rare words are as familiar as high-frequency words when neither has yet appeared on a list. On the surface, such an idea seems unlikely. On the other hand, in at least two of five conditions, memory for rare words did exhibit a mirror effect.

Although the findings to date are somewhat inconsistent, recognition memory for rare words may hold important and theoretically interesting implications for the mirror effect and, more generally, for models that attempt to explain recognition memory on the basis of item familiarity (e.g., Gillund & Shiffrin, 1984; Mandler, 1980). The first experiment reported below differed from previous research on memory for rare words in that a forced-choice recognition procedure was used. The use of a forced-choice procedure allows direct comparisons between items that differ in word frequency but not in list status (e.g., H− vs. L− or H+ vs. L+). The design was essentially identical to that used by Glanzer and Bowles (1976), except that rare words were included in the analysis. The second experiment used the standard yes/no recognition procedure to evaluate the generality of the findings obtained from the forced-choice procedure.

Experiment 1

Subjects in this experiment were exposed to lists of high-frequency, low-frequency, and rare words, followed by a two-

alternative, forced-choice recognition test involving all possible combinations of new and old words from the three frequency categories. Also included were "null" trials involving a forced choice between two items that appeared on the list or between two items that did not appear on the list (cf. Glanzer & Bowles, 1976). Following an analysis similar to that depicted in Figure 1, and assuming that the mirror effect holds for rare words, the predictions of a familiarity-based model are relatively straightforward. Because new rare words (R−) will presumably be the least familiar, the distribution for these words falls to the far left. If the mirror effect obtains, then the familiarity distribution for old rare words (R+) should fall to the far right. Thus, for example, given a choice between R+ and any other alternative (e.g., L+, H+, H−, L−, or R−) subjects should choose the former, whereas given a choice between R− and any other alternative subjects should choose the latter.

Method

Subjects. Seventy-two undergraduates at the University of California, San Diego, participated as subjects in the experiment to satisfy an introductory psychology course requirement.

Materials and design. A large pool of high-frequency, low-frequency, and rare words was compiled using Francis and Kucera (1982), Thorndike and Lorge (1944), and the *Oxford English Dictionary* (OED). The high-frequency words all occurred more than 40 times per million in the Francis and Kucera (1982) corpus, whereas the low-frequency words occurred between 1 and 3 times per million. The rare words were drawn from both Thorndike and Lorge (1944) and the OED. With regard to the former source, the words selected occurred with a frequency of less than once per 7 million. Words thought to be familiar to undergraduates despite their low frequency of usage were not included. With regard to the latter source, an effort was made to select words that did not appear in either Thorndike and Lorge (1944) or Francis and Kucera (1982) and that would presumably be unfamiliar to most undergraduates. A pool of 600 words, 200 from each frequency category, was constructed in this manner.

The words in each category were equated for length and pretested for undergraduates' knowledge of word meaning. Forty subjects rated each word on a 5-point scale ranging from 1 (*no knowledge of word meaning*) to 5 (*exact knowledge of word meaning*). The mean ratings for high-frequency, low-frequency, and rare words were 4.99, 4.31, and 1.41, respectively. Thus, the rare words were indeed quite unfamiliar.

For each subject, a single list of 150 words was constructed by randomly selecting 50 words from each of the three word pools (high frequency, low frequency, and rare). A different random order was used for every subject. Following list presentation, the words from the list were rerandomized and presented again on the forced-choice recognition test. An additional 50 words from each pool were randomly selected and paired with these test items to serve as distracters. During the recognition test, subjects received 10 repetitions of 15 different trial types. Nine of these were standard in the sense that they involved one item from the list and one item not from the list (e.g., H+ vs. L−), whereas six were null trials involving a choice between two items that appeared on the list (e.g., H+ vs. L+) or between two items that did not (e.g., H− vs. L−). The order in which these recognition trials were presented was randomly determined, and a different random order was used for every subject.

Procedure. All subjects were tested individually. After signing a consent form, subjects were informed that they would be viewing a long list of words on the screen and that the list would be followed by a recognition test. Following an instruction screen that introduced the list, the 150 words were presented one at a time at the center of a computer screen. Each word remained on the screen for 2.5 s and was followed by a 0.5-s interstimulus interval. After all 150 items were presented another instruction screen appeared informing the subject of the nature of the two-alternative, forced-choice recognition test that would follow. No mention was made of the null trials (cf. Glanzer & Bowles, 1976). On each of 150 recognition trials, two words appeared on the screen and the subject selected one of them by moving the cursor to that word (using a "mouse") and clicking once with the left button. After each selection, the words disappeared and two new words were presented for a recognition decision.

Results and Discussion

Forced-choice responses. Table 1 lists the proportion of correct responses for each of the nine recognition trials involving a choice between one old item and one new item. Thus, for example, the first entry represents the proportion of correct responses on trials involving a choice between an old high-frequency word (High+) and a new high-frequency word (High−). Also shown is the mean proportion correct for each target category averaged over distracter category (last column) and the mean proportion correct for each distracter category averaged over target category (bottom row).

A within-subjects analysis of variance (ANOVA) performed on the data in Table 1 revealed a main effect for target category (High+, Low+, Rare+), $F(2, 142) = 15.37$, $MS_e = 0.79$, as well as a main effect for distracter category (High−, Low−, Rare−), $F(2, 142) = 6.62$, $MS_e = 0.78$ (all statistical tests used an α level of .05). The interaction between target and distracter category did not approach significance. With regard to the targets, the main effect evidently derived from the reduced probability of recall for High+ relative to either Low+ or Rare+. Although an advantage for low-frequency words over high-frequency words was to be expected, the same advantage for the rare words is somewhat surprising. Pairwise Bonferroni t tests contrasting High+ versus Low+ and High+ versus Rare+ were both significant, $t(71) = 5.19$ and $t(71) = 3.86$, respectively. The very small difference between Low+ and Rare+ was not significant. With regard to the distracters, the main effect derived from the performance advantage on trials involving Low− relative to both High− and Rare−, $t(71) = 2.72$ and $t(71) = 4.31$, respectively. The small difference between High− and Rare− did not approach significance.

The pattern of results on trials involving high- and low-frequency words is in accordance with expectations and con-

Table 1
Proportion of Correct Recognition Judgments in Experiment 1

	High⁻	Low⁻	Rare⁻	*M*
High⁺	.774	.775	.735	.761
Low⁺	.819	.861	.818	.833
Rare⁺	.810	.879	.807	.832
M	.801	.838	.787	

forms to the mirror effect. That is, subjects were more likely to correctly choose low-frequency targets relative to high-frequency targets and less likely to choose low-frequency lures relative to high-frequency lures. However, the relatively poor performance on trials involving Rare– was unexpected given the relatively good performance on trials involving Rare+. Had the mirror effect held, performance on trials involving Rare– should have matched that on trials involving Low–. In addition to limiting the generality of the mirror effect, these results are difficult to reconcile with a theory of recognition memory based solely on item familiarity. Presumably, the rare items that did not appear on the list (i.e., Rare–) were the least familiar items of all, yet they were more likely to be mistakenly chosen as having been seen before than the low-frequency words (and slightly more so than the high-frequency words).

Table 2 lists performance on the six null trials, three of which involved a choice between two old items and three of which involved a choice between two new items. These trials were included in an effort to more directly compare response biases for words that differ in word frequency but not in list status. The pattern of results on these trials is, for the most part, consistent with the results shown in Table 1. When high- and low-frequency words were both new, subjects exhibited a slight preference for the high-frequency words. When they were both old, preference reversed in favor of low-frequency words. These findings are consistent with the idea that responses were based on item familiarity, and they conform to predictions based on the model shown in Figure 1. By contrast, subjects displayed a consistent preference for rare words relative to high-frequency words whether they were both old or both new. This result is clearly inconsistent with the notion that responses are based on item familiarity considering that a new rare word is surely less familiar than a new high-frequency word. Furthermore, the choice proportions suggest that the mirror effect does not necessarily extend to rare words.

Response scaling. The aforementioned results suggest that familiarity, as that word is ordinarily construed, may not always be the dimension along which subjects base their recognition decisions. Nevertheless, as detailed later, the data are sufficiently orderly to warrant the assumption that the obtained response probabilities were determined by each item's position along some unidimensional psychological scale. For lack of a better term, that scale might be generically labeled the "subjective sense of prior occurrence" (cf. Mandler, 1980). Indeed, the main conclusions of this experiment can be most clearly illustrated by calculating where each of the various word categories fall along this psychological scale.

The simplest technique that may be used to this end is Thurstone scaling. This procedure assumes that forced-choice decisions are determined by the absolute difference between two items along some underlying psychological scale and that the scale is unidimensional in nature. If these assumptions were true, then the obtained data should exhibit a property that has been termed *strong stochastic transitivity* (Coombs, 1964; Coombs, Dawes, & Tversky, 1970). That is, if A is preferred to B by a probability of p, and B is preferred to C by probability q, then A should be preferred to C by a probability that exceeds both p and q. To take one example from the present experiment, the probability of choosing L+ over H– is .819 (Table 1) and the probability of choosing H– over L– is .521 (Table 2). For strong stochastic transitivity to hold, we should expect to find that the probability of choosing L+ over L– exceeds both of these values. Indeed, from Table 1, the actual probability is .861 and strong stochastic transitivity is satisfied. Of the 20 possible tests of this kind, strong stochastic transitivity was satisfied on 19 occasions. The only exception was L+ versus R+ (.539), R+ versus L– (.879), and L+ versus L– (.861). Thus, the assumption that responding was based on differences along a unidimensional psychological scale is a plausible one.

The Thurstone scaling procedure is straightforward and consists of the following three steps: (a) entering the forced-choice response probabilities into a 6 × 6 matrix with rows and columns represented by H+, L+, R+, H–, L–, and R–; (b) converting the response probabilities into z scores; and (c) calculating a scale value for each word category based on the mean z score for each row of data (Baird & Noma, 1978). Following these steps yields interval scale values of 1.34, 1.66, 1.60, 0.46, 0.34, and 0.60 for H+, L+, R+, H–, L–, and R–, respectively (the 0 point of the scale was established arbitrarily by adding 1.0 to the average z scores). Figure 2 presents a graphical illustration of the location of each word category on the underlying dimension according to the Thurstone scaling procedure. As might be expected, the distance between old and new words is relatively large, whereas the distance between words in different frequency categories (e.g., H– and L–) is comparatively small.

The scale values shown in Figure 2 reveal that rare lures (R–) produce a higher subjective sense of prior occurrence than both high- and low-frequency lures (H– and L–, respectively). This result would not be expected if subjects were responding on the basis of an item's familiarity. In addition, although high- and low-frequency words exhibit a mirror effect, rare words clearly do not fall into the same pattern. The obvious question concerns why that might be. However, before pursuing an answer to that question, the generality of these results was examined using a yes/no recognition paradigm. Indeed, a replication seemed essential in light of the study by Mandler et al. (1982), which found that rare words were associated with a lower false alarm rate than both high- and low-frequency words.

Table 2
Recognition Choice Proportions on "Null" Trials in Experiment 1

Condition	New	Old
High/Low	.521	.381
High/Rare	.426	.438
Low/Rare	.404	.539

Note. The values represent the proportion of trials in which the first alternative was chosen over the second.

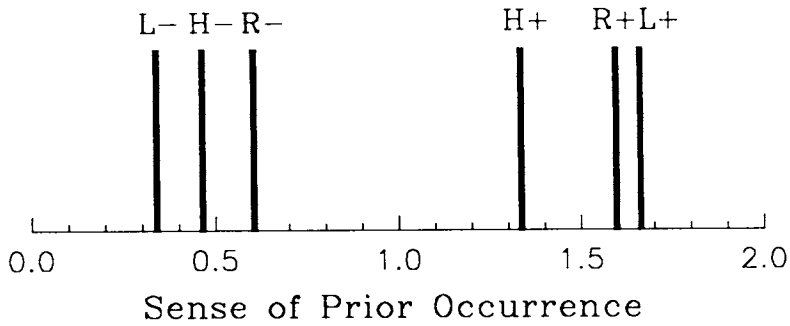

Figure 2. Scale values for each word category based on the Thurstone scaling procedure. (High-frequency, low-frequency, and rare words are denoted by H, L, and R, respectively, and list status, target versus lure, is denoted by + or −, respectively.)

Experiment 2

Method

Subjects. Seventy-two undergraduates at the University of California, San Diego, participated as subjects in the experiment to satisfy an introductory psychology course requirement.

Materials and design. The same words used in the previous experiment were used again here. For each subject, a single list of 150 words was constructed by randomly selecting 50 words from each of the three word pools (high frequency, low frequency, and rare). A different random order was used for every subject. Half of the words from each category on the list were rerandomized and presented again on a yes/no recognition test. An additional 25 words from each pool were randomly selected and intermixed with these test items to serve as distracters.

Procedure. All subjects were tested individually. After signing a consent form, subjects were informed that they would be viewing a long list of words on the screen and that the list would be followed by a recognition test. Following an instruction screen that introduced the list, the 150 words were presented one at a time at the center of a computer screen. Each word remained on the screen for 2.5 s and was followed by a 0.5-s interstimulus interval. After all 150 items were presented another instruction screen appeared informing the subject of the nature of the yes/no recognition test that would follow. On each of the 150 recognition trials, a single word appeared on the screen along with two boxes (a "yes" box and a "no" box) directly below and to either side of the word. The subject selected yes or no by moving the cursor to the appropriate box (using a mouse) and clicking once with the left button. After each decision, the word disappeared from the screen and a new word was presented for a decision. This process was repeated until all 150 items (75 targets and 75 lures) were tested.

Results and Discussion

Table 3 shows the number of hits and false alarms for the high-frequency, low-frequency, and rare words (the maximum value for each entry is 25). The third column shows the average of the d' scores calculated for individual subjects.

An overall ANOVA performed on the obtained d' scores was significant, $F(2, 142) = 8.12$, $MS_e = 0.31$. Subsequent t tests revealed that memory for low-frequency words exceeded that for high-frequency words, $t(71) = 20.87$, but the differences between low-frequency words and rare words and between rare words and high-frequency words were not quite statistically significant with the Bonferonni correction, $t(71) = 4.15$ and $t(71) = 3.36$, $p < .10$, respectively.

The hit and false alarm data shown in Table 3 were remarkably consistent with the obtained scale values from Experiment 1 (shown in Figure 2). More specifically, hits for low-frequency words and rare words exceeded that for high-frequency words, $t(71) = 2.76$ and $t(71) = 3.04$, respectively, whereas the small difference in hits for low-frequency words and rare words did not approach significance. Also in agreement with the previous experiment, false alarm rates for high-frequency words and rare words exceeded that for low-frequency words, although only the latter difference reached statistical significance, $t(71) = 2.42$.

In most respects, the present results are in agreement with those of Rao and Proctor (1984), who also used a yes/no recognition procedure involving high-frequency, low-frequency, and rare words. In general, they found relatively high false alarm rates for rare words (in some cases exceeding the false alarm rate for high-frequency words), and the obtained d' score for rare words fell midway between that for high- and low-frequency words. Furthermore, across five learning conditions in two experiments, high-frequency words exhibited a mirror effect with respect to low-frequency words. The relationship between high-frequency words and rare words was less clear cut, however. In two conditions, a mirror effect was obtained. In two other conditions, no mirror effect was obtained. However, because these authors were not concerned with an analysis of the mirror effect per se, the significance of

Table 3
Hit and False Alarm Rates in Experiment 2

Condition	Hits	False alarms	d'
High	15.97	4.35	1.50
Low	18.51	3.78	1.88
Rare	18.03	4.57	1.70

the relatively small differences in hits and false alarms was not evaluated in any condition. Therefore, a definitive statement regarding rare words and the mirror effect on the basis of that experiment is not possible.

Taken together, the results of Experiments 1 and 2 make two important points. First, in agreement with Glanzer and Adams (1990), the data do not easily conform to a simple strength model based on item familiarity. If so, we would expect false alarm rates to be highest for high-frequency words, second highest for low-frequency words, and lowest for rare words (cf. Mandler et al., 1982). Instead, rare words exhibit the highest false alarm rate of all. Second, for reasons that are as yet unclear, the results for rare words clearly do not conform to the mirror effect. Although the hit rate for rare words was consistently high (and comparable to low-frequency words), the false alarm rate was also high (and comparable to high-frequency words).

An answer to the question of why rare words appear to violate the mirror effect would be facilitated by a clearer understanding of the mirror effect itself. As indicated earlier, one possible explanation for the mirror effect, first proposed by Brown (1976) and recently developed in detail by Glanzer and Adams (1990), is based on the notion of subjective memorability. According to this account, subjects are aware of the fact that low-frequency words are more memorable than high-frequency words. This knowledge facilitates what Brown, Lewis, and Monk (1977) referred to as *negative recognition,* namely, the ability to determine that an item has not been seen before. Thus, instead of performing an exhaustive memory search to determine that a low-frequency lure was not on the list, subjects simply reject the lure on the grounds that it is the kind of word they would have remembered had it actually appeared before. From this point of view, the psychological scale represented in Figure 2 is simply the difference between an item's familiarity and its expected level of familiarity.

If this account were true, then a violation of the mirror effect might be expected if subjects were badly mistaken about the memorability of a particular class of words (such as rare words). Thus, although the results of Experiment 2 suggest that the rare words used here are actually more memorable than high-frequency words, subjects might nevertheless be under the mistaken impression that the opposite is true. If so, we might expect relatively high false alarm rates for rare words (because subjects assume they may have simply forgotten these unusual words) and relatively high hit rates as well (because these words are quite memorable after all).

A direct analysis of subjects' awareness of item memorability for high-frequency, low-frequency, and rare words has never been performed. However, as a general rule, the metamemory literature suggests that subjective estimates of memorability are usually reasonably accurate (e.g., Groninger, 1976; Thompson, 1982). The following experiments pursued the issue of subjective memorability for words of differing frequencies using several different procedures. The hypothesized results were as follows: subjects would regard low-frequency words as the most memorable, rare words as the least memorable, and high-frequency words as intermediate.

Subjective Memorability

Subjective memorability estimates were obtained using several different procedures. In Experiment 3, subjects were simply asked to rate item memorability on a 10-point scale for high-frequency, low-frequency, and rare words. In Experiment 4, subjects were instructed to imagine they had just seen a list of words and were then presented with a series of word pairs that constituted a kind of "recognition" test. For each pair, the subject was instructed to choose the item they would be more likely to recognize had it actually appeared on the imaginary list. Each word in a pair was drawn from a different frequency category (e.g., High+ vs. Low+). Experiment 5 was similar except that subjects were asked to choose the pair of words that would involve the easier recognition decision (e.g., High+/High− vs. Low+/Low−).

Experiment 3

Method

Subjects. Thirty-six undergraduates at the University of California, San Diego, participated as subjects in the experiment to satisfy an introductory psychology course requirement.

Materials and design. The same words used in the previous experiments were used again here. For each subject, a single list of 150 words was constructed by randomly selecting 50 words from each of the three word pools (high frequency, low frequency, and rare). A different random order was used for every subject.

Procedure. All subjects were tested individually. After signing a consent form, subjects were informed that they would be viewing a long list of words on the screen and that their task would involve rating each word for memorability. Subjects were instructed to rate each word assuming that their memory would be tested using a recognition procedure in which each word would be presented again for a yes/no recognition decision. Following an instruction screen that introduced the list and described the nature of a recognition test, the first word was presented along with a 10-point memorability scale, ranging from 1 (*very unlikely to recognize*) to 10 (*very likely to recognize*), with *unsure* corresponding to a rating of 5.5. The subject rated an individual word by moving a cursor to a point on the scale (using a mouse) and clicking once. That word then disappeared and the next word was presented for a rating. This procedure repeated until all 150 words were presented.

Results and Discussion

The results of this experiment were unexpectedly straightforward: predicted memorability varied directly with frequency of usage. The median memorability ratings for high-frequency, low-frequency, and rare words were 8.0, 6.4, and 4.6, respectively. A median test was performed on these data by tabulating the number of occasions in which a rating fell above versus below (or equal to) the grand median for each frequency category. The overall median test was highly significant, $\chi^2(2, N = 72) = 18$. Individual contrasts using the Bonferonni protection against Type I error verified the ordinal pattern apparent in the data. High-frequency words were rated as being more memorable than low-frequency words, $\chi^2(1, N = 72) = 6.72$, and rare words, $\chi^2(1, N = 72) = 16.10$. Low-frequency words, in turn, were rated as being more memora-

ble than rare words, $\chi^2(1, N = 72) = 9.38$. Exactly the same conclusions were reached using an ANOVA performed on the mean ratings (7.4, 6.4, and 4.7 for high-frequency, low-frequency, and rare words, respectively).

These findings weigh against the idea that subjects are aware of the fact that low-frequency words are more likely to be recognized than high-frequency words. The hypothesis that subjective memorability underlies the mirror effect is therefore weakened. One potential concern about the present design is that subjects may have misinterpreted the nature of the hypothetical memory test and responded as if memory would be tested by recall (despite instructions to the contrary). Indeed, postsession interviews occasionally revealed some confusion about the difference between recall and recognition. If memory were tested by free recall, then the memorability ratings obtained in this experiment would be on target. The next experiment was designed to evaluate subjective memorability in a way that was less likely to introduce such confusion. Specifically, subjects were asked to make hypothetical "recognition" decisions after imagining that they had just been exposed to a list of words.

Experiment 4

Method

Subjects. Thirty-six undergraduates at the University of California, San Diego, participated as subjects in the experiment to satisfy an introductory psychology course requirement.

Materials and design. The same words used in the previous experiments were used again here. For each subject, a single list of 75 word pairs was constructed by randomly selecting 50 words from each of the three word pools (high frequency, low frequency, and rare). The pairs consisted of one high-frequency word and one low-frequency word (H-L), one high-frequency word and one rare word (H-R), or one low-frequency word and one rare word (L-R). The list consisted of 25 repetitions of each pair type arranged randomly, and a different random order was used for every subject.

Procedure. All subjects were tested individually. The instructions to each subject asked them to imagine they had just seen a list of words and that all of the words to follow (to be presented in pairs) had appeared on that list. For each pair, subjects were instructed to select the word they would be more likely to recognize if this were a real recognition test. The 75 word pairs were presented side-by-side, one at a time, on the center of the screen. On each trial, the computer randomly assigned left/right positions to the two test words. The subject selected the word judged to be more memorable by moving a cursor to that word and clicking once. After a selection was made, the screen cleared and the next pair was presented for a decision. This process was repeated until all 75 pairs were presented.

Results and Discussion

The results of this experiment are summarized in the first column of Table 4. Each entry represents the proportion of trials on which the first alternative was chosen over the second. Thus, for example, when faced with a choice involving a high-frequency and a low-frequency word, the high-frequency alternative was chosen on 64.2% of the trials. When the choice involved a high-frequency word and a rare word, preference for the high-frequency alternative increased to 84.2%. Finally, when the choice involved a low-frequency word and a rare word, the low-frequency alternative was chosen on 79.7% of the trials. These values were tested against indifference (i.e., 50%) using the Bonferonni t-statistic and, in each case, the result was highly significant, $t(35) = 3.74$, 8.27, and 11.14, respectively. Thus, as in the previous experiment, subjective memorability estimates varied directly with frequency of usage.

Table 4
Predicted Memory Performance in Experiments 4 and 5

Condition	Experiment 4	Experiment 5
High/Low	.642	.526
High/Rare	.842	.738
Low/Rare	.797	.715

Note. The values represent the proportion of trials in which the first alternative was chosen over the second.

It should also be noted that, as in Experiment 1, these data exhibit strong stochastic transitivity. As indicated earlier, such a pattern is consistent with the idea that the forced-choice comparisons were determined by the difference between each item's location on a unidimensional psychological scale of subjective memorability. Thurstone scaling yielded actual scale values of 1.68, 1.23, and 0.09 for high-frequency, low-frequency, and rare words, respectively (the 0 point of the scale was arbitrarily established by adding 1.0 to the mean z scores).

The findings of this experiment again suggest that subjects do not regard low-frequency words as being more memorable than high-frequency words. On the contrary, they seem to mistakenly regard the reverse as being true. As before, the hypothesis that the mirror effect is a reflection of accurate subjective estimates of memorability remains unsubstantiated. However, subjects do seem unduly pessimistic about the memorability of rare words. Although these words were found to be more memorable than high-frequency words in Experiments 1 and 2, they were chosen on less than 16% of the trials involving a choice between a high-frequency word and a rare word.

Neither of the two previous experiments on subjective memorability distinguished between the identification of targets and lures, which is the crucial task in recognition. Instead, subjects were simply asked to rate item memorability (Experiment 3) or to choose the item that would be easier to recognize had it appeared on a list (Experiment 4). Perhaps more accurate memorability judgments would emerge if the procedure emphasized the discrimination between targets and lures of different word frequencies. The final experiment tested this idea.

Experiment 5

Method

Subjects. Thirty-four undergraduates at the University of California, San Diego, participated as subjects in the experiment to satisfy an introductory psychology course requirement.

Materials and design. The same words used in the previous experiments were used again here. For each subject, word pairs were created by randomly selecting 40 words from each of the three word pools (high frequency, low frequency, and rare). The pairs consisted of two high-frequency words (High+/High−), two low-frequency words (Low+/Low−), or two rare words (Rare+/Rare−). From these, three trial types were assembled: High+/High− versus Low+/Low−, High+/High− versus Rare+/Rare−, and Low+/Low− versus Rare+/Rare−. The final list consisted of 10 repetitions of each trial type arranged randomly, and a different random order was used for every subject.

Procedure. All subjects were tested individually. The instructions to each subject again asked them to imagine they had just seen a list of words. However, this time subjects were presented with two pairs of words (e.g., High+/High− and Low+/Low−) on every trial. For each pair, one word was designated as the target (High+ and Low+) and one word was designated as the lure (High− and Low−). Subjects were asked to imagine that the targets had appeared on the imaginary list and that the lures had not. Their task was to select the pair they believed they would be more likely to get right if this were an actual recognition test. Subjects made their selection by moving a cursor to the appropriate pair and clicking once. Once a selection was made, the screen cleared and the next set of pairs was presented.

Results and Discussion

The results of this experiment are shown in the second column of Table 4. The labels (e.g., High/Low) now represent trials involving pairs of words (e.g., High+/High− vs. Low+/Low−). The values in the table represent the proportion of trials in which the subjects chose the first of the two alternatives. On High/Low trials, for example, subjects chose the high-frequency pair on 52.6% of the trials, a value that did not differ significantly from indifference. However, on High/Rare trials the high-frequency pair was judged as the easier recognition discrimination on 73.8% of the trials, which did exceed indifference, $t(33) = 5.70$. Similarly, on Low/Rare trials, subjects exhibited a significant preference for the low-frequency pairs (71.5%), $t(33) = 5.65$. Thurstone scaling performed in a manner similar to that of the previous experiments yielded subjective memorability scale values of 1.36, 1.25, and 0.39 for high-frequency, low-frequency, and rare words, respectively.

Once again the results offer no suggestion that subjects are aware of the fact that low-frequency words are more memorable than high-frequency words. Subjective memorability instead tracks word frequency, although, in this experiment, the small advantage for high-frequency words over low-frequency words was not significant. These findings contrast to some extent with other findings from the metamemory literature showing that subjects, in general, can predict what they are likely to learn and remember (e.g., Groninger, 1976; Lovelace, 1984; Underwood, 1966). However, most of these studies show that predictive accuracy, although exceeding chance, is not extremely accurate. Furthermore, analyses are typically performed on an item-by-item basis rather than across classes. Subjects in the present series of experiments presumably would have predicted, with above-chance accuracy, their likelihood of recognizing individual high-frequency, low-frequency, and, perhaps, rare words. Across classes, however, their predictions are inaccurate.

At the very least, the hypothesis that accurate memorability estimates underlie the mirror effect must be amended to account for the fact that subjects do not correctly classify the relative memorability of different classes of words. It is possible, for example, that subjects learn about item memorability only during the memory test itself and that this experience is sufficient to offset their preexisting notions. However, until evidence of this kind is adduced, the present results would appear to suggest that subjective memorability and the mirror effect are unrelated.

General Discussion

The mirror effect refers to the common finding that words associated with high hit rates also tend to be associated with low false alarm rates (Glanzer & Adams, 1985). The present research investigated the generality of this phenomenon by testing memory for very unfamiliar words. The results of Experiments 1 and 2 showed that, although the mirror effect was obtained for high-and low-frequency words, it was not obtained for rare words. Furthermore, the false alarm rate for rare words exceeded that for low-frequency words and essentially matched that of high-frequency words. This result is consistent with findings reported by Rao and Proctor (1984) and weighs against the notion that recognition choices are governed exclusively by item familiarity (cf. Glanzer & Adams, 1990).

In agreement with all previous research on the subject, the present results suggest that the relationship between linguistic frequency and recognition performance can be accurately characterized by an inverted U (e.g., Zechmeister et al., 1978). However, at least for the large group of rare words considered in the present series of experiments, the decline in performance associated with rare words relative to low-frequency words results primarily from an increase in the rate of false alarms. By contrast, high-frequency words are associated with lower hit rates and higher false alarm rates relative to low-frequency words.

In two experiments reported by Mandler et al. (1982), and in contrast to the data reported here, the false alarm rate for rare words was well below that of both high-and low-frequency words. In two experiments reported by Rao and Proctor (1984), and in agreement with the data reported here, the false alarm rate for rare words was quite high across five conditions (and always higher than that for low-frequency words). The principal difference between the two sets of experiments was the size of the retention interval. The present Experiments 1 and 2, as well as those reported by Rao and Proctor (1984), involved immediate tests of recognition memory. By contrast, Mandler et al. (1982) used a retention interval of 24 hr in their first experiment and either 48 hr or 5 min in their second. Their most important finding for purposes of the present discussion was that the false alarm rate for rare words was very similar to that for high- and low-frequency words at the short retention interval and was considerably lower only at the longer retention intervals. Thus, it would seem that the least familiar words reliably produce the lowest false alarm rate only after a long retention interval. On an immediate memory test, the reverse may be true.

Experiments 3 through 5 tested the hypothesis that the mirror effect, when it occurs, is based on accurate subjective memorability estimates. According to this notion, lures judged to be memorable are easily rejected on the grounds that they would have been remembered had they actually appeared on the list. Inaccurate memorability estimates, on the other hand, might undermine the mirror effect (e.g., for rare words). The results of these experiments instead suggested that subjective memorability estimates mistakenly track word frequency, with high-frequency words consistently judged as being the most memorable. The hypothesis that the mirror effect is rooted in accurate estimates of memorability is therefore weakened.

That subjective memorability may sometimes play a role in negative recognition (i.e., the correct rejection of lures) seems incontrovertible. Brown et al. (1977), for example, pointed out that a subject's surname presented as a lure on a recognition test could easily be rejected on the grounds that it would have been remembered had it appeared on a preceding list. Similarly, most of us can confidently (and correctly) report that we have not recently had lunch with the President of the United States. On the other hand, whether subjective memorability underlies the mirror effect is less obvious, and the present findings appear to suggest that the effect emerges despite the mistaken impression that high-frequency words are more memorable than low-frequency words.

If subjective memorability does not underlie the mirror effect, what does? In their recent accounts of the mirror effect, Glanzer and Adams (1990) and Glanzer, Adams, and Iverson (1991) proposed an "attention/likelihood" model of recognition memory that is similar to but much more detailed than the strategy envisioned by Brown (1976). The model conceptualizes word representations as sets of features and assumes that, during list study, a subset of those features is marked as having been seen before. Because low-frequency words command more attention than high-frequency words, a greater number of their features are marked. The model further assumes that subjects are aware of the average number of features marked for high- and low-frequency words and that this information is used when making recognition judgments. Thus, for example, a low-frequency lure is easily rejected because it is found to have fewer marked features than would be expected had the word actually appeared on the list. A high-frequency lure is less easily rejected because the difference in the number of marked features associated with lures and targets is smaller.

Although consistent with earlier work on the mirror effect, the attention/likelihood model does not seem to offer an obvious explanation for the overall pattern of results obtained here. The model basically assumes that the high hit rate associated with low-frequency words results from the extra attention those words receive at encoding. Presumably, the relatively high hit rate associated with rare words in the present experiment could be explained in the same way. However, with regard to false alarms, the explanation is less clear. Experiments 3 through 5 suggest that false alarm rates, even when they do mirror hit rates, are not based on accurate estimates of subjective memorability. Therefore, to the extent that subjective memorability and "knowledge of feature marking" are equated, the present findings would appear to pose some difficulty for the attention/likelihood account.

On the other hand, perhaps subjective memorability and awareness of feature marking ought not to be equated. That is, although their preexisting beliefs about word memorability may be mistaken, perhaps subjects nevertheless discover during the course of encoding that more features are marked when studying low-frequency words than when studying high-frequency words. That information could then be used during the test phase to accurately reject low-frequency lures. The problem with this interpretation of the attention/likelihood model is that it fails to explain why rare words produce such a high rate of false alarms. If subjects learned during the course of encoding that a high proportion of rare word features were marked, why was that information of little use in rejecting rare lures? The absence of a mirror effect for rare words suggests that the assumptions of the model must somehow be modified to deal with this special case.

An alternative explanation for the false alarm pattern observed in Experiment 2, which does not assume any knowledge of memorability (or of item marking), holds that lures may be falsely recognized because of their perceptual or semantic similarity to encoded targets. Consider, for example, the case of rare lures, which produce a subjective sense of prior occurrence well above that of low-frequency lures (Figure 2). In most cases, subjects do not know the meaning of rare words and may therefore encode them in terms of their orthographic or phonemic properties. In the former case, having seen a word like *dative*, subjects might be easily lured by the visually similar word *davit*. In the latter case, having seen a word such as *nubbin*, they might be easily lured by the phonetically similar word *numen*. To the extent that such generalization occurred, the false alarm rate for rare words would be increased along with the hit rate.

A somewhat similar account was offered by Glanzer and Bowles (1976) and by Schulman (1976) in an effort to explain the higher false alarm rate for high-frequency words relative to low-frequency words. As with the aforementioned account, this phenomenon was attributed to a kind of generalization from high-frequency targets to high-frequency lures. More specifically, because low-frequency words have fewer and more exact meanings than high-frequency words, the semantic overlap between low-frequency lures and targets is small relative to that between high-frequency lures and targets (cf. Earhard, 1982; Glanzer & Bowles, 1976).

From both points of view, low-frequency words enjoy a significant false alarm advantage. In most cases, subjects know the meaning of low-frequency words and presumably encode them semantically. Thus, in contrast to the rare word case, subjects should be less vulnerable to orthographically or phonetically similar low-frequency lures. Moreover, the semantic encoding of low-frequency words is more precise and specific than that of their high-frequency counterparts. As a result, low-frequency lures are correspondingly less likely to semantically match encoded targets. Because they are the least susceptible to orthographic, phonemic, and semantic generalization, low-frequency lures should produce the lowest sense of prior occurrence and, therefore, the lowest rate of false alarms.

The preceding analysis is somewhat speculative, but it does suggest a potential line of inquiry for future research. Presumably, it should be possible to manipulate orthographic, phonemic, and semantic overlap between targets and lures to obtain more direct evidence of generalization. However, for now, the major conclusions to be drawn from the present research are that rare words appear to represent an exception to the otherwise ubiquitous mirror effect and that when the mirror effect does occur, it does so for reasons other than accurate subjective estimates of memorability.

References

Baird, J. C., & Noma, E. (1978). *Fundamentals of scaling and psychophysics.* New York: Wiley.

Brown, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition* (pp. 1–35). New York: Wiley.

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

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction.* Englewood Cliffs, NJ: Prentice-Hall.

Earhard, B. (1982). Determinants of the word-frequency effect in recognition memory. *Memory & Cognition, 10,* 115–124.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Gentner, D., & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory & Cognition, 9,* 434–443.

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: Theory and data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 5–16.

Glanzer, M., Adams, J., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 81–93.

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.

Groninger, L. D. (1976). Predicting recognition during storage: The capacity of the memory system to evaluate itself. *Bulletin of the Psychonomic Society, 7,* 425–428.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 756–766.

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

Mandler, G., Goodman, G. O., & Wilkes-Gibbs, D. L. (1982). The word-frequency paradox in recognition. *Memory & Cognition, 10,* 33–42.

Nelson, T. O. (1988). Predictive accuracy of the feeling of knowing across different criterion tasks and across different subject populations and individuals. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 190–196). New York: Wiley.

Rao, K. V., & Proctor, R. W. (1984). Study-phase processing and the word frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 386–394.

Schulman, A. I. (1976). Memory for rare words previously rated for familiarity. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 301–307.

Thompson, C. P. (1982). Memory for unique personal events: The roommate study. *Memory & Cognition, 10,* 324–332.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Columbia University Press.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology, 71,* 673–679.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society, 11,* 371–373.

Received July 25, 1991
Revision received November 13, 1991
Accepted November 26, 1991 ■