Journal of Experimental Psychology:
Learning, Memory, and Cognition
1990, Vol. 16, No. 1, 5-16

The Mirror Effect in Recognition Memory: Data and Theory

Murray Glanzer and John K. Adams
New York University

The mirror effect is a regularity in recognition memory that requires reexamination of current views of memory. Five experiments that further support and extend the generality of the mirror effect are reported. The first two experiments vary word frequency. The third and fourth vary both word frequency and concreteness. The fifth experiment varies word frequency, concreteness, and the subject's operations on the words. The experiments furnish data on the stability of the effect, its relation to response times, its extension to multiple mirror effects, and its extension beyond stimulus variables to operation variables. A theory of the effect and predictions that derive from the theory are presented.

The mirror effect (Glanzer & Adams, 1985) is a strong regularity in recognition memory. It is summarized as follows. If there are two classes of stimuli, and one is more accurately recognized than the other, then the superior class is *both* more accurately recognized as old when old *and also* more accurately recognized as new when new. For example, low-frequency words are better recognized than high-frequency words. The mirror effect means that the greater efficiency in recognizing is always twofold. Old low-frequency words are better recognized as old than are old high-frequency words, and new low-frequency words are better recognized as new than are new high-frequency words.

In the discussion that follows, recognition performance is viewed as based on subjects' responses to underlying distributions of some measure for new and old items. These distributions are not, of course, directly observed. They are deduced from recognition data. The relation of the underlying distributions to the data obtained from standard recognition tests—yes/no, confidence rating, forced choice—is given in detail by Glanzer and Adams (1985) and others (Egan, 1975; Green & Swets, 1966; McNicol, 1972).

Some possible distributions for two classes of stimuli, when one is more accurately recognized than the other, are shown in Figure 1. Panel 1 represents the distributions that underlie the mirror effect. The panel shows the distributions for two classes of stimuli, A and B. Class A is recognized with greater accuracy. This is represented by the relatively large distance between the underlying A old (AO) and A new (AN) distributions. Class B is recognized with less accuracy. This is represented by the relatively small distance between the B old (BO) and B new (BN) distributions. The mirror effect means that the difference in accuracy of recognition of A and B determines *two* more differences in distance. A old is *higher* on the decision axis than B old, *and* A new is *lower* than B new, as shown in the panel. These differences will be the focus of the statistical analyses of the experiments reported here.

The mirror effect regularities do not follow from the simple fact that class A stimuli are handled more accurately than class B stimuli. Given the difference in accuracy, a variety of patterns of the underlying distributions could hold that violate the mirror effect. Two such patterns are shown in Panels 2 and 3 of Figure 1.

Each of the relations represented in the panels of Figure 1 implies a particular pattern of data for each of the standard recognition tests. The relations in Panel 1 imply for confidence rating data that

$$R(AN) < R(BN) < R(BO) < R(AO),$$

where R represents the mean confidence rating on a scale that has *very sure new* at its low end and *very sure old* at its high end.

For yes/no data the implied relations are

$$FA(AN) < FA(BN) < H(BO) < H(AO),$$

where FA is false alarm rate, and H is hit rate.

For forced choice data the relations are

$$P(BO, BN) < P(AO, BN), P(BO, AN) < P(AO, AN),$$

where P is the proportion of choices of the first argument over the second argument within the parentheses. The comma between the two middle terms signifies an indeterminate relation between those terms.

A meta-analysis of 80 recognition experiments supported the existence of the mirror effect (Glanzer & Adams, 1985) for all recognition paradigms: yes/no, confidence rating, and forced choice. The meta-analysis, moreover, demonstrated that the effect held for all stimulus variables that could be surveyed: word frequency, concreteness, meaningfulness, and others.

This research was supported by Grant 1 RO1 MH449 from the National Institute of Mental Health, Grant BNS 84-15904 from the National Science Foundation, and Contract F49620-86-C-0131 with the Air Force Office of Scientific Research.

We thank Leslie Sherman and Amy Wolff for assistance in collection and analysis of data, Jean-Claude Falmagne, Elliot Hirshman, Geoff Iverson, and Gay Snodgrass for their helpful comments on an earlier draft of this article. We also thank Douglas L. Hintzman, William E. Hockley, and two anonymous reviewers for their constructive comments.

Correspondence concerning the article should be addressed to Murray Glanzer, Department of Psychology, New York University, New York, New York 10003.

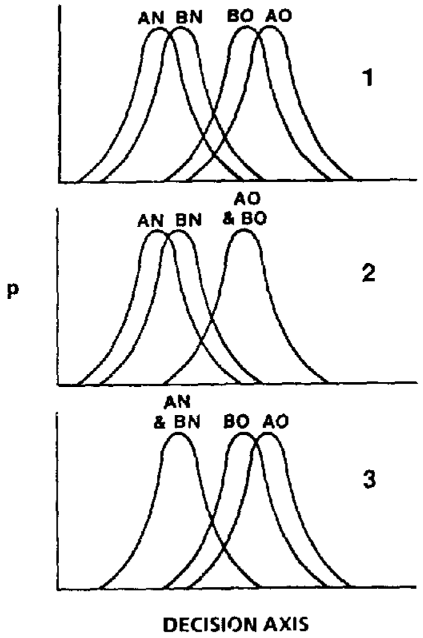

Figure 1. Three possible orders of underlying distributions when accuracy on stimulus class A is greater than accuracy on class B. (O = old, N = new.) Panel 1 shows the mirror effect.

Such a regularity in memory is a challenge to strength theories of recognition memory. This point was first noted by Brown (1976). According to strength theories, in a recognition test the subject decides on the basis of the strength of the items. Terms equivalent to strength are *the amount of marking* or *familiarity of the items*. These theories, therefore, label the decision axis in Figure 1 as strength, amount of marking, or familiarity. Such theories have problems in accounting for the mirror effect. They contain no inherent mechanism that arrays the underlying new and old distributions in the mirror order as depicted in Panel 1 of Figure 1. In this article we will consider a different theory of the effect: attention/likelihood theory.

Several experiments will now be presented. The first experiment will expand the data base of the mirror effect. It will also examine the mean confidence ratings for misses. This measure is of interest for the testing of theories of the effect.

Experiment 1: Word Frequency and Incidental Learning

In this experiment the subjects first carried out an incidental learning task, lexical decision. Then they were given a surprise recognition test in which the old words were the words presented during the lexical decision task. The times to make the lexical decision responses and the recognition responses were recorded.

Method

Procedure. In the lexical decision task, the subjects viewed words and nonwords on a monitor. The presentation was paced by the subjects who, for each item, pressed one of two response keys on a response board labeled *yes* (for word) and *no* (for nonword). The *yes* key was assigned to the subject's dominant hand. The subjects were told to be quick and accurate. The lexical decision task was preceded by eight practice items.

The recognition test was carried out on the computer keyboard. Only words were presented during this test. Onset of the test word started a timing period. When the subjects reached a decision, they pressed the space bar which ended the timing period. Then they pressed one of two keys indicating whether the item was old or new. The keys in this experiment and in Experiments 3, 4, and 5 were arranged so that old was assigned to the subject's dominant hand. Finally, they pressed one of four keys labeled *unsure, somewhat sure, moderately sure,* or *very sure.* This three-stage response was used to exclude from the subjects' recognition response times the additional time to move to the extreme rating scale positions. That time could produce an artifactual speed–accuracy trade-off.

All stimuli for both the incidental learning task and recognition test were presented centered, in uppercase letters. Following the subject's response, the screen went blank for 500 ms in the incidental learning task and for 2,000 ms in the recognition test. Then the next item appeared on the screen. The presentation of the items on the monitor was controlled by a computer which also recorded responses and response times. The program used for the computer is described in Adams (1985). Except when noted otherwise, the procedures and stimulus presentation used here were the same in Experiments 3, 4, and 5.

Materials. The words presented in the lexical decision task consisted of 124 high-frequency words (mean log Kučera-Francis frequency 4.8) and 124 low-frequency words (mean log frequency 2.4). The word groups both had a mean length of 5.0 letters. The 248 nonwords were constructed to be orthographically and phonologically legal. They had a mean length of 5.6 letters. The new words presented in the recognition test (124 high frequency and 124 low frequency) had the same mean frequency and length as the old words. The main list of lexical decision items was preceded by 12 initial filler items and followed by 12 final filler items (each consisting of six words and six nonwords) to eliminate serial position effects. Nonwords and filler words did not appear on the subsequent recognition test. The word sets were counterbalanced across subjects so that each of the experimental words was used an equal number of times as old and new in the recognition test. Again, this counterbalancing of word sets was used in Experiments 3, 4, and 5.

Subjects. Sixteen undergraduates participated in the experiment to fulfill an introductory psychology course requirement. All were native speakers of English. This description of the way subjects were recruited and selected holds also for Experiments 3, 4, and 5.

Results

The subjects were highly accurate on all classes of items in the lexical decision task. The proportions of correct responses are as follows: high-frequency words ($M = .99$), low-frequency words ($M = .97$), nonwords ($M = .95$). The effect of item class is statistically significant, $F(2, 30) = 13.43$, $p < .0001$, $MS_e = 0.022$. Analysis of proportions here and in the rest of this article was carried out on the arc sine transformation of the original proportions. (In this article, where scores are transformed either by arc sine or logarithm, the accompanying MS_es will be for the transformed scores.)

The lexical decision response times are, as expected, negatively correlated with the proportion correct. The geometric means (antilogs of mean logs) were 618, 671, 840 ms for high-

Table 1
Means for the Four Conditions of Experiment 1 (N = 16)

	New		Old	
Measure	Low	High	High	Low
Rating	3.34	3.76	5.09	5.56
P(yes)	.304	.359	.592	.661
RT[a]	1,213	1,192	1,170	1,166

Note. P(yes) = proportion of *yes* responses. RT = response time.
[a] Antilog of mean log response time (in milliseconds).

frequency words, low-frequency words, and nonwords, respectively. Analyses of response times in this experiment and in Experiment 3 were carried out on the logs of the response times. The effect of item class is, again, statistically significant, $F(2, 30) = 74.61$, $p < .0001$, $MS_e = 0.005$. In summary, the lexical decision task showed the usual pattern found in experiments with word frequency as the variable (see Glanzer & Ehrenreich, 1979).

Two related sets of recognition measures are of interest with respect to the mirror effect. One is the confidence ratings for the four stimulus conditions. The other is the proportions of hits and false alarms. In the case of confidence ratings, the mirror pattern is

$$R(\mathrm{LN}) < R(\mathrm{HN}) < R(\mathrm{HO}) < R(\mathrm{LO}),$$

where *R* signifies mean rating. The arguments *L* and *H* refer again to low- and high-frequency words; *N* and *O* refer to new and old. The ratings here and in the following experiments are placed on a single scale, with the highest value, 8, assigned to *very sure the item is old*, 7 to *moderately sure the item is old*, 6 to *somewhat sure the item is old*, 5 to *unsure the item is old*, 4 to *unsure the item is new*, and so on down to 1, assigned to *very sure the item is new*.

In case of hits and false alarms the mirror pattern is

$$FA(\mathrm{LN}) < FA(\mathrm{HN}) < H(\mathrm{HO}) < H(\mathrm{LO}).$$

Table 1 and the following tables are arranged so that the mirror effect is evidenced by a progression of increasing means going from left to right. The mean confidence ratings (row 1) and the hits and false alarms, the proportion of *yes* responses (row 2), both show the mirror effect.

The statistical analysis of the data sets—confidence ratings, hits and false alarms, response times—for this and the following experiments is carried out by first doing a preliminary one-way analysis of variance across the experimental conditions, in this case low-frequency new (LN), high-frequency new (HN), high-frequency old (HO), and low-frequency old (LO). This analysis is followed by two key comparisons: (a) high-frequency old versus low-frequency old and (b) high-frequency new versus low-frequency new.

These two comparisons are critical. If the differences are in the right direction and statistically significant, they support the stability of the mirror effect. One-tailed tests are used for these planned comparisons.

The overall evaluation (the one-way analysis of variance here of the four experimental conditions) of the confidence ratings shows $F(3, 45) = 120.92$, $p < .0001$, $MS_e = 0.147$. This overall evaluation gives highly significant effects because it includes the effect of new versus old items as well as the comparisons of interest. The overall evaluation, which in all experiments is highly significant, is reported here but not in the following experiments because it is not of interest. Only the key planned comparisons are presented.

These comparisons for the confidence ratings show both critical differences in the right direction and both statistically significant: high-frequency old versus low-frequency old, $t(45) = 3.46$, $p < .005$; high-frequency new versus low-frequency new, $t(45) = 3.10$, $p < .005$. The parallel analysis of proportion of the *yes* responses (hits and false alarms) shows the overall $F(3, 45) = 120.91$, $p < .0001$, $MS_e = 0.018$. Both critical differences are again in the right direction and statistically significant: low-frequency versus high-frequency hits, $t(45) = 3.14$, $p < .005$; low-frequency versus high-frequency false alarms, $t(45) = 2.57$, $p < .01$. The analysis of the proportion of *yes* responses is partially redundant with the analysis of confidence ratings. It therefore will be reported in only abbreviated form in the following experiments. The mean proportions will be included in the tables to underscore the regularity of the effect.

Also shown in Table 1 are the mean response times for each of the four conditions. There are two possible expectations concerning the pattern of response times. One, and of greater concern to us, is that there is a speed–accuracy trade-off, with response times for LO > HO and LN > HN. Such a trade-off would make the mirror effect trivial. Hockley and Murdock (1987) present evidence (Hockley, 1982) against a trade-off. The possibility of trade-off is, however, important enough to require full checking. The other possibility is a positive correlation of speed and accuracy: LO < HO and LN < HN. The Hockley (1982) data show such a positive correlation. Here, however, neither pattern holds: neither a speed–accuracy trade-off (negative correlation) nor a speed–accuracy positive correlation.

Analysis of variance of the log response times reveals only the difference between new and old as statistically significant, $t(45) = 2.11$, $p < .05$, $MS_e = 0.002$. Neither of the other relevant comparisons is large or statistically significant: high-frequency old versus low-frequency old; high-frequency new versus low-frequency new. There is no evidence here, therefore, that speed–accuracy correlation plays a role in the mirror effect. It could be argued that, for the response arrangement used in this experiment, the subjects' initial response may have preceded their actual decision and that this reduced the correlation of speed and accuracy. The similarity of the mean response times, all close to 1,200 ms, does not support such an argument. We will, however, examine this question again in Experiment 3 with a different response arrangement.

The mean ratings for misses have been singled out as important by Brown, Lewis, and Monk (1977). They note that missed highly memorable old items are rejected with greater confidence than missed low-memorable old items. This finding has theoretical importance because it contradicts expectations on the basis of strength theories. The data of this experiment replicate the finding. The mean confidence rating for low-frequency misses is 2.15 and for high-frequency misses is 2.33. The difference, though small, is in the right direction and is statistically significant, $F(1, 15) = 15.58$, $p < .002$, MS_e

= 0.016. We consider this finding and others that show the same relation in the final section.

In summary, the present experiment shows the following: (a) another replication of the mirror effect for word frequency on both confidence rating and *yes/no* data; (b) no evidence of a speed–accuracy correlation; (c) differences in the mean confidence ratings for misses. The next experiment was designed to examine Points a and c further.

Experiment 2: Word Frequency and Intentional Learning

This was a replication of Experiment 1, with several changes in procedure. It was carried out as a group experiment with intentional instead of incidental learning and auditory instead of visual presentation.

Method

Procedure. A group of subjects heard a single list of words read at a 1-s rate. They were told that they would be given a recognition test. The test consisted of a printed list of words, mixed old and new. Next to each word was a sequence of letters and numbers—*Y, N, 1, 2, 3,* and *4*. The subject indicated old by circling *Y*, new by circling *N*, and degree of confidence by circling the number (*1* for *unsure*, *4* for *very sure*).

Materials. A shorter list was constructed from the materials used in Experiment 1. The study list consisted of 50 high-frequency words (mean log frequency = 5.1) and 50 low-frequency words (mean log frequency = 2.5) plus 24 initial filler words and 24 final filler words. The filler words were evenly divided into high- and low-frequency words. The test list consisted of the 100 study list words plus matched (same mean log frequency) groups of 50 new high-frequency and 50 new low-frequency words as distractors. The subjects' responses on the test were self-paced.

Subjects. Thirty-five undergraduates in a memory course participated in the experiment as a class exercise.

Results

The main results are shown in Table 2. They parallel closely the results of Experiment 1. The mirror effect is present in both the mean confidence ratings and proportion of *yes* responses.

The tests ($MS_e = 0.436$) of the mean ratings again show the key differences to be statistically significant: high-frequency old versus low-frequency old, $t(102) = 3.59$, $p < .0005$; high-frequency new versus low-frequency new, $t(102) = 2.11$, $p < .025$. The mirror pattern also holds in the parallel analysis of the hits and false alarms (row 2), both $ps < .05$.

The mean ratings for the misses show the same pattern as in Experiment 1. The mean confidence ratings for misses are lower for low-frequency words (2.56) than for high-frequency words (2.68). This is based on 34 subjects because 1 subject did not have any misses. The difference again is slight but statistically significant, $F(1, 33) = 7.90$, $p < .01$, $MS_e = 0.030$.

Table 2
Means for the Four Conditions of Experiment 2 (N = 35)

	New		Old	
Measure	Low	High	High	Low
Rating	3.22	3.56	5.51	6.08
P(yes)	.228	.281	.613	.704

Note. P(yes) = proportion of *yes* responses.

In summary, the results of Experiment 2—with auditory presentation, intentional learning, and group testing—confirm the findings of Experiment 1.

Experiment 3: Multiple Mirror Effects and Partial Order—Frequency and Concreteness

The purpose of this experiment was to develop a multiple mirror effect by using two variables—normative frequency and concreteness—in a single set of items. Each of the two variables alone produces a mirror effect. We combined these two variables factorially in order to produce a more complex mirror effect involving more ordered terms than the four ordered terms seen in the preceeding experiments. If the two variables are equal in their effectiveness, then the mean confidence ratings should give the following partial order:

$$R(\mathrm{LCN}) < R(\mathrm{HCN}), R(\mathrm{LAN}) < R(\mathrm{HAN}) <$$
$$R(\mathrm{HAO}) < R(\mathrm{LAO}), R(\mathrm{HCO}) < R(\mathrm{LCO}),$$

where *C* represents concrete and *A* abstract words (for example, LCN = low frequency, concrete, new). Parallel to these inequalities for the ratings should be a partial order for the hits and false alarms:

$$FA(\mathrm{LCN}) < FA(\mathrm{HCN}), FA(\mathrm{LAN}) < FA(\mathrm{HAN}) <$$
$$H(\mathrm{HAO}) < H(\mathrm{LAO}), H(\mathrm{HCO}) < H(\mathrm{LCO}).$$

We had actually expected that frequency would be more effective than concreteness. In that case, where one variable is stronger, a full rather than a partial order is expected. A fully ordered set of terms will be produced in the next experiment.

Method

Procedure. The procedure was basically the same as that in Experiment 1, with lexical decision as the incidental learning task. The sequence of responses required of the subjects in the recognition test was, however, simplified. The subjects made a single response to each test word, pressing one of an array of eight keys with the rightmost key indicating *very sure old* and the leftmost key indicating *very sure new*. The eight keys were in the top row of the keyboard with labels indicating confidence levels. On the test the subjects saw a series of 280 words—half old, half new. The stimulus presentations in both study and test were self-paced. The interstimulus intervals on the study and the test lists were the same as in the Experiment 1 (500 and 2,000 ms, respectively).

Subjects. Sixteen undergraduates participated.

Materials. The composition of the lists differed from that in Experiment 1. In the lexical decision task 140 words and 140 nonwords were presented. The 140 words were drawn, 35 from each of four 70 word sets: low-frequency concrete (LC), high-frequency concrete (HC), low-frequency abstract (LA), and high-frequency abstract (HA). The two low-frequency sets both had mean log frequency of 1.5; the two high-frequency sets both had mean log frequency 3.9, based on the Kučera-Francis (1967) norms. The two concrete sets both had a mean concreteness rating of 6.8; the two abstract sets both had a mean rating of 2.6, based on the Paivio, Yuille, and Madigan (1968) norms. With both concreteness and frequency varied, it was not possible to match the word lengths across conditions as closely as in Experiments 1 and 2. The means for the four groups listed above were 7.2, 5.9, 7.8, and 6.9 letters, respectively. Because we were concerned that these differences might affect the pattern of results, we subsequently carried out a special analysis of the data to determine whether the differences had an effect. They did not. This analysis will be reported briefly later.

The main list of lexical decision items was preceded by 80 filler items and followed by 80 filler items (half words and half nonwords) which did not appear on the recognition test. The recognition test list consisted of the old words plus the remaining unpresented 140 words, 35 from each of the four word sets.

Results

In the lexical decision task, high-frequency words took less time to respond to (M = 592 ms) than did low-frequency words (M = 702 ms), $F(1, 15) = 65.38$, $p < .0001$, MS_e = 0.007. High-frequency words (M = .98) were responded to more accurately than low-frequency words (M = .92), $F(1, 15) = 37.34$, $p < .0001$, $MS_e = 0.062$. The concrete versus abstract words in the lexical decision test did not differ significantly in response time ($F < 1$). Accuracy was, however, somewhat higher for abstract (M = .96) than concrete words (M = .94), $F(1, 15) = 3.92$, $p = .07$, $MS_e = 0.038$.

The overall recognition test results are presented in Table 3. They are considered first with respect to frequency alone and concreteness alone. This simplification is justified because a factorial analysis of variance of the data showed that the two stimulus variables, frequency and concreteness, do not interact. After examining the main effects of frequency and concreteness separately, the results for the combination of the two variables will be examined.

Summing across concreteness conditions, the mirror effect for frequency is evident again in both the confidence ratings and the proportions of *yes* responses (hits and false alarms). The key tests ($MS_e = 0.272$) of the confidence ratings for old low-frequency (M = 5.79) versus old high-frequency words (M = 5.31) show $t(105) = 3.66$, $p < .0005$; and for new high-frequency (M = 3.63) versus new low-frequency words (M = 3.17) show $t(105) = 3.50$, $p < .0005$. Parallel tests on the proportion *yes* data give the same results (both $ps < .025$).

Summing across frequency conditions, the mirror pattern also appears for concreteness in both the confidence ratings and the hits and false alarms. Tests of the confidence ratings show concrete old (M = 5.74) higher than abstract old (M = 5.37), $t(105) = 2.88$, $p < .005$, and concrete new (M = 2.98) lower than abstract new (M = 3.82), $t(105) = 6.40$, $p < .0005$. Concrete hits (M = .679) are higher than abstract hits, (M = .654), but the difference is not statistically significant. Concrete false alarms (M = .177) are lower than abstract false alarms (M = .320), $t(105) = 6.01$, $p < .0005$.

The mean confidence ratings of misses again show the order noted by Brown et al. (1977). The order holds for both frequency and concreteness. The low-frequency misses are rated lower (2.81) than the high-frequency misses (3.03), $t(45) = 2.86$, $p < .005$, $MS_e = 0.096$. The concrete misses are also rated lower (2.89) than the abstract (2.95), but the difference is not statistically significant.

Before moving to the consideration of the accuracy scores for the combined conditions, two issues will be touched on. One concerns the effect of word length on the pattern of the results. We noted earlier that the word sets differed in mean length. Although those lengths did not correspond to the mirror effects observed, we decided to check on any possible effects of word length fully. We did this by removing words from the word sets so that the reduced word sets all had identical mean lengths while preserving the match of frequency and concreteness. This meant going from four sets of 70 words to four sets of 30 words. We then computed the mean ratings for the reduced sets and analyzed the pattern produced. The pattern and overall analysis of variance corresponded fully to those obtained for the complete sets of words. To convey the correspondence, the means for the reduced set corresponding to the means in the first row of Table 3 read from left to right as follows: 2.72, 3.32, 3.68, 4.01, 5.19, 5.46, 5.42, and 6.10. The means are only slightly different from those for the larger set of items. The results of statistical analysis based on the reduced set also differ only slightly and in no important way from the full analysis. The differences in word length, therefore, were not important.

The second issue concerns the response times. There is evidence of some differences: old are faster than new items, $F(1, 15) = 12.776$, $p < .003$, $MS_e = 0.013$; overall, high-frequency words are faster than low, $F(1, 15) = 20.168$, $p < .0005$, $MS_e = 0.004$; concrete are faster than abstract, $F(1, 15) = 5.716$, $p < .05$, $MS_e = 0.013$.

Our main concern was, however, the presence of a speed-accuracy trade-off. There is no evidence of this. There is no relation between the response times and either of the accuracy measures within either the new conditions or the old in Table 3. The rank order correlation of speed and accuracy is zero

Table 3
Means for the Eight Conditions of Experiment 3 (N = 16)

	New				Old			
Measure	LC	HC	LA	HA	HA	LA	HC	LC
Rating	2.73	3.23	3.61	4.02	5.19	5.54	5.44	6.04
P(yes)	.161	.193	.284	.357	.630	.677	.625	.732
RT[a]	2,011	1,885	2,051	1,970	1,877	1,927	1,728	1,844

Note. L = low frequency; H = high frequency; C = concrete; A = abstract; *P*(yes) = proportion of *yes* responses; RT = reaction time.
[a] Antilog of mean log response time (in milliseconds).

for both the new and old conditions. There is, then, no evidence in this experiment, or in Experiment 1, for any general relation between the mirror effect and response times.

We return now to the accuracy measures in Table 3. Contrary to our expectations, concreteness and frequency were approximately equal in their effectiveness. This can be seen by comparing the highest mean rating for frequency, which for the low-frequency old words (LC plus LA) is 5.79, and the highest mean rating for concreteness, which for contrete old words (LC plus HC) is 5.74. With two variables of equal strength, only partial orders are expected for the confidence ratings. A partial order is what is obtained:

$$R(\mathrm{LCN}) < R(\mathrm{HCN}), R(\mathrm{LAN}) < R(\mathrm{HAN}) < R(\mathrm{HAO}) < R(\mathrm{LAO}), R(\mathrm{HCO}) < R(\mathrm{LCO}).$$

A deviation from the partial order is obtained, however, in the hits and false alarms because $H(\mathrm{HCO})$, .625, is slightly lower than $H(\mathrm{HAO})$, .630. This we consider to reflect the relative weakness of the proportion *yes* data, which contain less information than do the ratings.

If a partial rather than full order is due to the absence of strong differences in the effectiveness of the two stimulus variables, frequency and concreteness, then a number of changes can be introduced to produce the full order. One possible change is to select the sets of words used so that either the differences in frequency or the differences in word concreteness would be greater than in the word sets used in this experiment. We could, for example, select only the very highest and very lowest frequency words. This, however, would reduce further an already limited pool of words. The other way would be to weaken one of the variables, for example, by adding middle range items in either the high-concreteness set or the low-frequency set. This could, however, weaken the effectiveness of the variable sufficiently to lose the mirror regularity. We decided not to change the word sets but to introduce an encoding task that would differentially affect the word sets. We therefore repeated the experiment, making the concreteness variable stronger by a concreteness encoding task.

Experiment 4: Multiple Mirror Effects and Full Order—Frequency and Concreteness Plus Concreteness Encoding Task

This was a replication of Experiment 3, with a change in the encoding task. We hoped that a concreteness encoding task would strengthen the concreteness variable and thus give a full order of the eight means produced by the combination of two variables—word frequency and concreteness. The eight means should display a higher order mirror effect.

The words were the same as those in Experiment 3. Instead of lexical decision, however, a concreteness encoding task was given. No nonwords were shown. During the initial list presentation, the subjects carried out, as an incidental learning task, a concreteness judgment on the words. During the recognition test the words, both new and old, were each judged first for concreteness before the recognition judgment was made. This was done in order to have the encoding operation affect new as well as old items.

Method

Materials. The study list consisted of the 140 words in four categories used in Experiment 3 (LC, HC, LA, HA) plus 4 practice items, 40 initial filler words, and 40 final filler words. The test list consisted of those 140 words plus 140 matched words.

Procedure. The subject was instructed that items that could be sensed (seen, heard, touched, tasted, or smelled) were concrete. During the encoding task the subjects pressed a key on a keyboard labeled "+" if the word on the screen was judged *concrete* or a key labeled "–" if it was judged *not concrete.* During the initial encoding trials, the subject received feedback on the correctness of the judgment. The feedback consisted of the word *right* or *wrong* appearing on the screen for 750 ms. During the recognition test, the subject made a concreteness judgment first for each word, but no feedback was given. Immediately after the concreteness judgment, the subject made a confidence judgment on whether the word was old or new, on an eight-key array as in Experiment 3.

Subjects. Sixteen undergraduates participated.

Results

On the initial encoding task, the subjects were more accurate on low-frequency (M = .96) than high-frequency words (M = .94), $F(1, 15) = 6.66$, $p < .03$, $MS_e = 0.025$, and on concrete (M = .97) than abstract words (M = .93), $F(1, 15) = 16.36$, $p < .002$, $MS_e = 0.054$. Items encoded incorrectly on test trials (3.5%) were not included in the scoring of recognition performance. Examination of the data shows, however, that even if these items are included, they do not change the pattern of results.

The results for the recognition test are given in Table 4. First, the encoding task was successful in making the concreteness variable stronger in the recognition task. The mean rating for old concrete words here is 6.72 as compared with 6.47 for old low-frequency words. (The corresponding means in Experiment 3 were 5.74 and 5.79.) We can expect, then, that a full order of inequalities will be found for these data.

The results will be examined again, first with respect to frequency alone and concreteness alone. As in Experiment 3, a factorial analysis of variance of the data showed that frequency and concreteness did not interact.

The means show the mirror effect for both word frequency alone and concreteness alone, and for both the mean confidence ratings and the proportion of *yes* responses (hits and false alarms) in each. The tests ($MS_e = 0.300$) of the ratings of high-frequency old ($M = 6.17$) versus low-frequency old ($M = 6.47$) give $t(105) = 2.25$, $p < .025$; and high-frequency new ($M = 3.29$) versus low-frequency new (M = 2.85), $t(105) = 3.21$, $p < .001$. Parallel tests on the proportion *yes* data show both comparisons with $p < .05$.

For concreteness the confidence ratings ($MS_e = 0.300$) of the concrete old ($M = 6.72$) versus abstract old ($M = 5.92$) give $t(105) = 5.87$, $p < .0005$. The ratings for the abstract new ($M = 3.30$) versus concrete new ($M = 2.85$) give $t(105) = 3.32$, $p < .001$. Parallel tests on the proportion *yes* data show both comparisons with $p < .001$.

The order of the mean confidence ratings for misses noted before holds again ($MS_e = 0.380$). Low-frequency misses ($M = 2.48$) have lower ratings than do high-frequency misses

Table 4
Means for the Eight Conditions of Experiment 4 ($N = 16$)

	New				Old			
Measure	LC	HC	LA	HA	HA	LA	HC	LC
Rating	2.67	3.02	3.03	3.57	5.85	5.99	6.48	6.96
P(yes)	.148	.200	.201	.300	.747	.758	.813	.882

Note. L = low frequency; H = high frequency; C = concrete; A = abstract; *P*(yes) = proportion of *yes* responses.

($M = 2.77$), $t(45) = 1.92$, $p < .05$. Concrete misses ($M = 2.47$) have lower ratings than do abstract misses ($M = 2.78$), $t(45) = 2.06$, $p < .025$.

Of particular interest here is whether there is an extended eight-category mirror effect, in full order, now that one of the experimental variables, concreteness, is stronger than the other. Table 4 displays the data for all eight combined conditions. Both mean confidence ratings and proportion of *yes* responses now show the expected full order:

$$R(\mathrm{LCN}) < R(\mathrm{HCN}) < R(\mathrm{LAN}) < R(\mathrm{HAN}) <$$
$$R(\mathrm{HAO}) < R(\mathrm{LAO}) < R(\mathrm{HCO}) < R(\mathrm{LCO})$$

and

$$FA(\mathrm{LCN}) < FA(\mathrm{HCN}) < FA(\mathrm{LAN}) < FA(\mathrm{HAN}) <$$
$$H(\mathrm{HAO}) < H(\mathrm{LAO}) < H(\mathrm{HCO}) < H(\mathrm{LCO}).$$

To examine the strength of the orderings, we carried out an analysis that parallels the tests carried out in the preceding experiments. In the earlier tests there were two old and two new means. Here there are four of each. The mirror effect will be evidenced by the strength of the linear component in each set of four means. We therefore evaluated the linear component of each set of four related recognition measures in Table 4, for example, the confidence ratings of LCO, HCO, LAO, and HAO in row 1. When this is done, we find the following: The linear component for confidence rating means of the four old conditions gives $F(1, 105) = 39.13$, $p < .0005$; for the four new conditions, $F(1, 105) = 19.45$, $p < .0005$; for hits, $F(1, 105) = 19.06$, $p < .0005$; and for false alarms, $F(1, 105) = 17.89$, $p < .0005$. The extent of order in the means can be fully conveyed by evaluating the proportion of variance accounted for by the mirror ordering in each array of means. For confidence ratings, the proportion of variance accounted for by these two linear components, after the effect of old versus new items is taken out, is .93; it is .85 for the hits and false alarms.

The results of the experiment strengthen the empirical basis of the mirror effect. The effect is shown, moreover, to produce an extended order when two variables that differ in effectiveness are used. The extended order is an eight-position mirror effect.

Experiment 5: Frequency, Concreteness, and Transformation

The purpose of this experiment was to examine multiple mirror effects with a third, new type of variable added—transformation of the list words. Kolers (1973, 1974, 1975a, 1975b), Kolers and Ostry (1974), and Graf (1982) have shown that recognition memory is better for transformed text (for example, text in which the letters are inverted or reversed) than for standard text. Of the seven separate experiments reported, however, only one shows the mirror effect.

The effect of transformation is of importance for establishing the generality of the mirror effect. Almost all of the demonstrations of the mirror effect are for stimulus variables, such as word frequency and concreteness. Those variables are produced by the selection of sets of items. Transformation falls outside the class of stimulus variables. Transformation can be applied to any item, and it is, therefore, independent of any item set. If transformation can be shown to produce the mirror effect, then a more general statement concerning the effect may be made: *Any* variable (not just classes of stimuli) that affects efficiency of recognition will produce the mirror effect. If the effect cannot be demonstrated for transformation, then the mirror effect may be limited to stimulus variables.

There are two reasons why the cited experiments on transformation may not have shown the mirror effect. One is that the testing procedure in those experiments was complex. In the Kolers experiments, the test items included not only old sentences in the same form as originally presented but also old sentences in a different form (for example, in standard form when originally presented inverted). The subjects classified the sentences as old same-form, old different-form, or new. Hits and false alarm rates that approximate those from ordinary recognition tests were derived from those classifications. The complexity of the procedure required of the subjects may have worked against clear demonstration of the effect. In the Graf (1982) experiment, the subjects viewed sentences during the study phase but were given word pairs during the test.

Another reason for the negative results may be floor effects on the false alarms. The subjects in the Kolers negative cases showed low false-alarm rates (.02 to .09) in both the standard and transformed conditions. This means that the possibility of a clear difference showing is slight. We therefore decided to examine the effect of transformation in a simpler arrangement and with material that we knew would not show floor effects.

This experiment was basically the same as Experiment 3 except for the addition of a transformation to half the words presented. This transformation, reversal of the order of letters in the word, required a decoding operation by the subject. Half the words were presented in standard order; half were presented in reverse order, for example, *emoh.*

Method

Procedure. The subjects were instructed to pronounce all words presented on the screen. Those presented in standard order were simply read. Those in reverse order had to be decoded and then spoken. The experimenter monitored the performance throughout to make sure that both tasks were performed correctly. This was done both in the list presentation and in the test. During the test, the subject said each word aloud and then responded as in Experiment 4 by pressing one of eight keys on the top row of the keyboard (with labels ranging from *NNNN*, *NNN*, . . . to *YYYY*) to indicate whether the word was new or old and the degree of confidence in the judgment.

Materials. The word lists were the same as in Experiment 3 and 4 except that two words were deleted from each of the four basic word sets (LC, HC, LA, and HA) to give a total of 272. This permitted the counterbalancing of word lists with the additional transformation variable. The mean log frequencies and mean concreteness measures for the basic word sets were the same as in Experiments 3 and 4.

The study list consisted of 136 words in four categories plus 4 practice items, 40 initial filler words, and 40 final filler words. The test list consisted of the 136 old words, plus 136 new words from the same four categories. Old words were presented with letters in the same order as in their initial presentation. For example, if a word was reversed initially, it was presented reversed during test.

Subjects. Thirty-two undergraduates participated.

Results

The overall means for each variable separately are shown in Table 5. The mirror effect appears in each row of the table. Preliminary analysis of the data indicated, however, that frequency and transformation interacted. The data for those variables are, therefore, separated out in Table 6, which shows the transformation conditions at both levels of word frequency. It can be seen that the mirror effect holds for the transformation at both high and low frequency. The test of the critical conditions for the means in Table 6 shows all the differences for the mean ratings ($MS_e = 0.425$) as statistically significant at the .01 level or better except the difference for new reversed versus new standard in the low-frequency condition ($p < .10$). The same pattern holds for the proportion *yes* data ($MS_e = 0.115$), in which all key comparisons are significant at the .025 level or better except, again, for the new reversed versus the new standard in the low-frequency condition ($p < .20$). The comparisons for frequency are all statistically significant for both confidence ratings and proportion *yes* at the .0005 level except for a nonsignificant and slightly reversed effect of low old versus high old in the reversed condition ($p > .20$).

Table 5
Means for the Transformation, Frequency, and Concreteness Conditions of Experiment 5 (N = 32)

Measure	New		Old	
	Transformation			
	Reversed	Standard	Standard	Reversed
Rating	2.74	2.96	5.66	7.25
P(yes)	.182	.211	.687	.922
	Frequency			
	Low	High	High	Low
Rating	2.55	3.15	6.30	6.61
P(yes)	.157	.237	.780	.829
	Concreteness			
	Concrete	Abstract	Abstract	Concrete
Rating	2.46	3.24	6.35	6.56
P(yes)	.136	.257	.799	.810

Note. *P*(yes) = proportion of *yes* responses.

Table 6
Means for the Transformation Condition, High and Low Frequency Separate

	New		Old	
	Transformation			
Condition	Reversed	Standard	Standard	Reversed
High frequency				
Rating	3.01	3.29	5.32	7.28
P(yes)	.215	.259	.630	.929
Low frequency				
Rating	2.47	2.62	5.99	7.23
P(yes)	.150	.164	.742	.916

Note. *P*(yes) = proportion of *yes* responses.

The reason for the interaction between frequency and transformation may be that the reversed old condition brings the performance close to the ceiling in both the ratings (greater than 7.2 on a scale of 8) and the proportion *yes* greater than .90). The mirror effect holds, however, for the transformation variable at both levels of frequency. This variable was our main concern. The deviation in word frequency is not of major concern because the meta-analysis (Glanzer & Adams, 1985) showed the mirror effect for word frequency in 23 out of 24 published experiments, and Experiments 1, 2, 3, and 4 above all show it.

The confidence ratings for misses have the same pattern as before on each of the variables. Low-frequency misses (M = 2.49) are lower than high-frequency misses (M = 2.79), $F(1, 31) = 19.80$, $p < .0001$, $MS_e = 0.070$; concrete misses (M = 2.52) are lower than abstract (M = 2.76), $F(1, 31) = 9.59$, $p < .005$, $MS_e = 0.096$; and reversed misses (M = 2.49) are lower than standard (2.69), $F(1, 31) = 2.86$, $p = .10$, $MS_e = 0.224$.

The results of this experiment support further the points made in the preceding experiments. The main new finding is that the mirror effect can be produced by variables other than stimulus variables such as word frequency and word concreteness. It is produced by transformations on a single set of stimulus words. The transformations induce subjects to carry out operations on the words that affect the accuracy of recognition. There is support, therefore, for the more general statement concerning the mirror effect. Any variable that affects recognition accuracy, not just stimulus variables, will produce the effect.

General Discussion

Brown (1976) and Brown et al. (1977) were the first to argue that the mirror effect required a change in the theoretical approach to recognition memory. They argued that the subjects took account of more than the strength of the items being evaluated. They took account also of the memorability of the items. This more complex basis of decision is incorporated in the theory that will be presented next, attention/likelihood theory.

Attention/likelihood theory is a sampling theory with two special mechanisms—an attention mechanism and a decision mechanism. The decision mechanism proposed differentiates it from current theories of recognition. The key idea concerning the decision mechanism in recognition is that the subjects evaluate a complex of information related to an item. The complex includes information about the relation of the given item to both a model new item and a model old item. This information is realized in a likelihood ratio (see Assertion 5 below).

The assertions of the theory are the following:

1. Stimuli are sets of features. The number of such features is N. This will be assumed constant for all stimuli. Because N refers to features, there is no reason to assume, at this point, that one stimulus has more or fewer features than another.

2. Some proportion of those features is marked in new stimuli. This proportion is p(new). The p(new) represents the noise level. This again, here, will be assumed constant for all stimuli. There is no reason, at this point, to assume that one new stimulus enters with greater noise marking than another.

3. Different classes of stimuli or different situations evoke different amounts of attention by the subject. This is translated into differences in the number of features, n(i), examined (sampled) during a trial. The sampling is random.

4. When features are examined, they are marked. The proportion of features marked is $\alpha(\mathrm{i}) = n(\mathrm{i})/N$. Therefore, the state of stimuli after they have been experienced is given by the following equation:

$$p(\mathrm{i, old}) = p(\mathrm{new}) + \alpha(\mathrm{i})\cdot(1 - p(\mathrm{new})). \qquad (1)$$

Conditions that evoke examination of a larger proportion of features will result in the marking of a larger proportion of features. The learning constant α(i) will be larger, and the learning rate faster.

5. During a recognition test, the subject uses the standard mechanisms of signal detection theory in making responses. Specifically, likelihood ratios are computed and decisions are made on the basis of those likelihood ratios.

Assertions 2 and 3 set up the underlying distributions for new items—binomials with the parameters n(i) and p(new) for a particular condition. Assertion 4 sets up binomial distributions for the old items with parameters n(i) and p(old). The subject uses information related to those distributions to generate likelihood ratios and responds on the basis of those likelihood ratios. This distinguishes this theory from strength theories in which the subject responds on the basis of strength or its equivalent: amount of marking, familiarity. The likelihood ratio is a key mechanism in the production of the mirror effect. For the binomial distributions we consider here, the log likelihood ratio for a single presented item is the following:

$$\ln L = x\cdot\ln\left(\frac{p(\mathrm{i, old})}{p(\mathrm{new})}\right) + [n(\mathrm{i}) - x]\cdot\ln\left(\frac{q(\mathrm{i, old})}{q(\mathrm{new})}\right). \qquad (2)$$

The n(i) is the number of features the subject observes. The x is the number of those marked. They are presented by the stimulus and are available to the subject. The logarithmic terms reflect the subject's model of the situation.

The process is the following. A test item is presented. The subject examines a number of features (n(i)) and notes the number of those that are marked (x). The subject then brings in two items of information—the proportion of marked features an old item of this type is expected to have and the proportion of marked features a new item is expected to have. On the basis of this information, likelihood ratios are computed. The likelihood is used in the final decision. For example, in a *yes/no* test if the likelihood ratio is greater than a preset likelihood criterion, the subject says "yes." The logarithmic terms in Equation 2 are the subjects' model of the situation. They play the same role as Brown's (1977) memorability evaluation.

The theory permits us to specify key statistics of the process. It also permits us by computation, to simulate the regularities that make up the mirror effect. Two key statistics are the mean and variance of the log likelihood ($\ln L$) distributions:

$$\mathrm{M}\ln L(\mathrm{i, j}) = n(\mathrm{i})\cdot p(\mathrm{i, j})\cdot\ln\left(\frac{p(\mathrm{i, old})}{p(\mathrm{new})}\right) + n(\mathrm{i})\cdot q(\mathrm{i, j})\cdot\ln\left(\frac{q(\mathrm{i, old})}{q(\mathrm{new})}\right) \qquad (3)$$

$$\mathrm{Var}\ln L(\mathrm{i, j}) = n(\mathrm{i})\cdot p(\mathrm{i, j})\cdot q(\mathrm{i, j})\cdot\left[\ln\left(\frac{p(\mathrm{i, old})\cdot q(\mathrm{new})}{p(\mathrm{new})\cdot q(\mathrm{i, old})}\right)\right]^2, \qquad (4)$$

where i is the experimental condition, such as stimulus set A or B, and j is the stimulus state, either new or old. The variance will be used later to test the theory.

One possible objection to the theory, as stated earlier, is that it has the subject hold in mind several different p(old)s. From one point of view, however, the subject has to have only some idea of the average p(new), the n(i), and the number of features. The p(i, old) can then be estimated, or at least ordered. We can simplify the theory further and assume that the subject works with a single p(old), for example, the average p(old) for several stimulus classes, not two or more as implied above. In that case, p(i, old) and q(i, old)—the logarithmic terms in the mean and variance—reduce to a single p(old) and q(old). The terms outside the logarithmic terms are not affected. They reflect the contribution of the actual stimuli, not the subject's model of the situation. It can be shown that the main effect considered so far—the mirror order—still holds under this simplification. Moreover, the derivations concerning the variances which will be tested later also hold. We cannot, however, handle the ratings for misses with this assumption.

Using the theory as presented above, we have carried out hundreds of computations with a large range of Ns, $n(i)$s, and p(news)s, and therefore also for a large range of α(i)s and p(i,old)s. Our only restriction on the ps has been that they stay under .50. Our computations show that the theory produces the mirror pattern for the standard recognition measures:

1. hits and false alarms: $FA(\mathrm{AN}) < FA(\mathrm{BN}) < H(\mathrm{BO}) < H(\mathrm{AO})$;

2. mean confidence ratings: $R(\mathrm{AN}) < R(\mathrm{BN}) < R(\mathrm{BO}) < R(\mathrm{AO})$;

3. two-alternative forced choice: $P(\mathrm{BO,BN}) < P(\mathrm{AO,BN})$, $P(\mathrm{BO,AN}) < P(\mathrm{AO,AN})$.

It also produces the order of confidence ratings for misses found in the data.

Some general tests of the theory are possible. We will not do conventional fitting of the values for the five experiments. Although the theory has only four basic parameters—N, two n(i)s, and p(new)—it cannot be used to fit the data of the present experiments. For example, the *yes/no* data of Experiment 1 give only four means with four parameters to be estimated. Therefore, instead of conventional fitting, we will apply some general tests of the theory. The tests will be concerned with the slopes of the receiver operating characteristic (ROC) for conditions in the Experiments 1 through 5, using Equation 4, the equation for the variance.

The theory permits us to derive some critical information about the ratio of the variances for pairs of conditions relevant to ROCs. We will do this for one case, that involving low (L) and high (H) frequency, first. For that case we consider four variance ratios: (a) Var ln L(LO)/Var ln L(HN); (b) Var ln L(LO)/Var ln L(LN); (c) Var ln L(HO)/Var ln L(HN); (d) Var ln L(HO)/Var ln L(LN). These variance ratios yield prediction concerning the slopes of the ROCs.

On the basis of Equation 4 we can derive two statements. Their derivation is given in the Appendix.

1. The four ratios above are listed in order of size, with the highest ratio first.

2. The first three ratios are all greater than 1.0. The fourth, for HO and LN, is indeterminate. It may be greater, less than, or equal to 1.0. Its size relative to the other ratios is, however, known. This is asserted in the first statement.

The four ratios above are related to four ROCs: (a) low-frequency hits against high-frequency false alarms (LO/HN); (b) low-frequency hits against low-frequency false alarms (LO/LN); (c) high-frequency hits against high-frequency false alarms (HO/HN); (d) high-frequency hits against low-frequency false alarms (HO/LN). ROCs 2 and 3 (standard ROCs) are the two that would ordinarily be plotted. ROCs 1 and 4 (crossed ROCs) will be considered here in order to test the theory fully.

There is a known relation (Green & Swets, 1966, p. 62) between the variances of the signal and noise distributions and the slope of the normalized (z score) ROC. The ratio of the signal variance to the noise variance is the inverse of the slope of the ROC. On the basis of this relation, the four ratios above imply the following two statements for the ROC's.

1. Because the ratios of variances are listed in order from highest to lowest, the slopes of the normalized ROCs should show the inverse order. The slope of the ROC for low-frequency hits against high-frequency false alarms (LO/HN) corresponding to the first ratio should be the lowest, and the slope of the ROC (HO/LN) corresponding to the last ratio should be the highest.

2. The first three normalized ROCs should all give slopes less than 1.0.

We will now examine the slopes of the four normalized ROCs obtained for each of the three variables in the five reported experiments. Because Experiments 3, 4, and 5 contain several variables, they give a total of nine sets of ROCs. All five experiments varied frequency. Experiments 3, 4, and 5 varied concreteness. Only Experiment 5 had transformation as a variable.

The slopes for these ROCs are presented in Table 7. The frequency variable gives the entries in the left part of the table. The concreteness and transformation variables give the entries in the right part of the table. The ordering of the ROCs has been set so that equivalent ROCs appear on the same row. For example, L is the strong, H the weak frequency condition; C the strong, A the weak concreteness condition; R the strong, S the weak transformation condition. Therefore, LO/HN, CO/AN, and RO/SN are on the same row—the ROCs for strong condition hits against weak condition false alarms.

1. With respect to the value of the slopes, every one of the 27 (nine sets of three each) predicted to be less than 1.0 is indeed less than 1.0 (LO/HN, LO/LN, HO/HN; CO/AN, CO/CN, AO/AN; RO/SN, RO/RN, SO/SN). The probability of 27 such results occurring by chance, using the binomial with $p = .5$, is 7.5×10^{-9}. If only the standard ROCs are considered—LO/LN and HO/HN and their parallels—then the probability of 18 such results occurring by chance is 3.81×10^{-6}.

2. With respect to the ordering of the sets of four slopes for each variable, we find that all but one corresponds fully to the predicted order. Applying the binomal, the probability of eight out of nine cases giving the predicted order by chance, with $p = 1/24$ and $n = 9$, is 3.78×10^{-12}. We can, again, restrict our attention to the standard ROCs—LO/LN and HO/HN and their parallels. We find then that eight out of the nine show the predicted order. The probability of this number or more occurring by chance, using the binomial with $p = .5$, is .0195.

Attention/likelihood theory does the following:

1. It handles the known regularities of the mirror effect—the ordering of hits and false alarms, the ordering of confidence ratings, and the ordering of choices in the two-alternative forced choice.

2. It handles new regularities—the size and order of ROC slopes.

Table 7
Slopes of Normalized ROCs for Each Variable in the Five Experiments (1, 2, 3, 4, and 5)

Freq.	1	2	3	4	5	Conc.	3	4	5	Transf.	5
LO/HN	.66	.56	.69	.64	.61	CO/AN	.68	.56	.61	RO/SN	.56
LO/LN	.74	.61	.81	.74	.72	CO/CN	.74	.65	.67	RO/RN	.66
HO/HN	.91	.70	.84	.71	.76	AO/AN	.89	.77	.82	SO/SN	.85
HO/LN	1.03	.75	.98	.82	.91	AO/CN	.96	.89	.90	SO/RN	1.00

Note. Experiments 1, 2, 3, 4, and 5 all varied frequency. Only Experiments 3, 4, and 5 varied concreteness. Experiment 5 alone varied transformation. Freq. = frequency; Conc. = concreteness; Transf. = transformation; L = low frequency; H = high frequency; C = concrete; A = abstract; R = reversed; S = standard; O = old; N = new.

There are three other approaches to either the general mirror effect or special cases of it. Two of these (Glanzer & Bowles, 1976; Gillund & Shiffrin, 1984) were concerned with a special case—word frequency effects. Both are strength theory approaches.

The first approach was based on work in our laboratory (Bowles & Glanzer, 1983; Glanzer & Bowles, 1976). One problem with the approach is that it was specific to the case of word frequency. It could not be generalized to other stimulus variables and would have further difficulties with variables such as transformation. Our dissatisfaction with that theory led to the formulation of attention/likelihood theory.

A second approach is part of the comprehensive memory theory of Gillund and Shiffrin (1984, p. 46). It also focuses on the specific case of word frequency. The approach assumes that the subject rescales the underlying distributions on the basis of distance from separate criteria and their standard deviations. The subject then aligns the distributions by placing the different criteria in a single location. The rescaling and alignment produce the mirror effect. The specific characteristics of the process that necessarily produce the effect are not given.

A third approach that concerns itself with the mirror effect is that of Hockley and Murdock (1987, p. 355). In that approach the order of underlying distributions depicted in Panel 1 of Figure 1 is assumed. The problem, however, is to explain why those underlying distributions are ordered as they are.

All three approaches handle the mirror effect as a special puzzle. We believe, however, that the generality of the mirror effect requires a new view of the process underlying recognition memory. Subjects make their recognition decisions by using a complex of information about each stimulus. In the theory outlined here, the complex is a likelihood ratio. Whatever the approach, some equivalent complex and an appropriate decision mechanism will have to be postulated to handle this regularity in recognition memory.

References

Adams, J. K. (1985). Visually presented verbal stimuli by assembly language on the Apple II computer. *Behavior Research Methods, Instruments & Computers, 17*, 489-502.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in memory. *Memory & Cognition, 11*, 307-315.

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.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* New York: Academic Press.

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

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

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.

Glanzer, M., & Ehrenreich, S. L. (1979). Structure and search of the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 18*, 381-398.

Graf, P. (1982). The memorial consequences of generation and transformation. *Journal of Verbal Learning and Verbal Behavior, 21*, 539-548.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Hockley, W. E. (1982). Retrieval processes in continuous recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 497-512.

Hockley, W. E., & Murdock, B. B., Jr. (1987). A decision model for accuracy and response latency in recognition memory. *Psychological Review, 94*, 341-358.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition, 1*, 347-355.

Kolers, P. A. (1974). Two kinds of recognition. *Canadian Journal of Psychology, 28*, 51-61.

Kolers, P. A. (1975a). Addendum to "Remembering operations." *Memory & Cognition, 3*, 29-30.

Kolers, P. A. (1975b). Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 689-701.

Kolers, P. A., & Ostry, D. (1974). Time course of loss of information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior, 13*, 599-612.

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

McNicol, D. (1972). *A primer of signal detection theory.* London: George Allen & Unwin.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement, 76*(No. 1).

(Appendix follows on next page)

Appendix

The order and values of the variance ratios can be determined by examining the terms that make up each ratio and taking account of the relative sizes of corresponding terms. All that is assumed is that one condition (for example, low frequency) is more effective than the other (for example, high frequency). For low- and high-frequency words, we have the following relations: $n(\mathrm{L}) > n(\mathrm{H})$; $p(\mathrm{LO}) > p\ (\mathrm{HO})$; $p(\mathrm{LN}) = p(\mathrm{HN}) = p(\mathrm{N})$, where L = low, H = high, O = old, N = new. To simplify the comparisons, let the logarithmic terms in the varaince equation be written as R(H) and R(L), where

$$R(\mathrm{H}) = \left[\ln\left(\frac{p(\mathrm{HO})\cdot q(\mathrm{N})}{q(\mathrm{HO})\cdot p(\mathrm{N})}\right)\right]^2$$

$$R(\mathrm{L}) = \left[\ln\left(\frac{p(\mathrm{LO})\cdot q(\mathrm{N})}{q(\mathrm{LO})\cdot p(\mathrm{N})}\right)\right]^2.$$

Because $p(\mathrm{LO}) > p(\mathrm{HO})$, then $R(\mathrm{L}) > R(\mathrm{H})$.

Let us look first at two ratios of variance:

$$\mathrm{Var\ ln}\ L(\mathrm{LO})/\mathrm{Var\ ln}\ L(\mathrm{HN}) = \frac{n(\mathrm{L})\cdot p(\mathrm{LO})\cdot q(\mathrm{LO})\cdot R(\mathrm{L})}{n(\mathrm{H})\cdot p(\mathrm{N})\cdot q(\mathrm{N})\cdot R(\mathrm{H})}. \qquad (\mathrm{A1})$$

$$\mathrm{Var\ ln}\ L(\mathrm{LO})/\mathrm{Var\ ln}\ L(\mathrm{LN}) = \frac{n(\mathrm{L})\cdot p(\mathrm{LO})\cdot q(\mathrm{LO})\cdot R(\mathrm{L})}{n(\mathrm{L})\cdot p(\mathrm{N})\cdot q(\mathrm{N})\cdot R(\mathrm{L})}. \qquad (\mathrm{A2})$$

Because $n(\mathrm{L}) > n(\mathrm{H})$ and $R(\mathrm{L}) > R(\mathrm{H})$, the first ratio has to be higher than the second. For the third ratio,

$$\mathrm{Var\ ln}\ L(\mathrm{HO})/\mathrm{Var\ ln}\ L(\mathrm{HN}) = \frac{n(\mathrm{H})\cdot p(\mathrm{HO})\cdot q(\mathrm{HO})\cdot R(\mathrm{H})}{n(\mathrm{H})\cdot p(\mathrm{N})\cdot q(\mathrm{N})\cdot R(\mathrm{H})}. \qquad (\mathrm{A3})$$

Because $p(\mathrm{LO})\cdot q(\mathrm{LO}) > p(\mathrm{HO})\cdot q(\mathrm{HO})$, the second ratio has to be higher than the third. (The inequality holds when $p(\mathrm{old}) \leq .50$. We assumed this boundary initially and used it in all of our exploratory computations.) Finally, we look at the fourth ratio:

$$\mathrm{Var\ ln}\ L(\mathrm{HO})/\mathrm{Var\ ln}\ L(\mathrm{LN}) = \frac{n(\mathrm{H})\cdot p(\mathrm{HO})\cdot q(\mathrm{HO})\cdot R(\mathrm{H})}{n(\mathrm{L})\cdot p(\mathrm{N})\cdot q(\mathrm{N})\cdot R(\mathrm{L})}. \qquad (\mathrm{A4})$$

Because $n(\mathrm{L}) > n(\mathrm{H})$ and $R(\mathrm{L}) > R(\mathrm{H})$, Ratio 4 has to be less than Ratio 3. These comparisons give the order of the four ratios.

The examination of the terms composing each ratio shows that the first three are all greater than 1.0. For example, every term—n(L), $p(\mathrm{LO})\cdot q(\mathrm{LO})$, R(L)—in the numerator of Ratio 1 is greater than the corresponding term in the denominator—n(H), $p(\mathrm{N})\cdot q(\mathrm{N})$, R(H). Ratio 4 is the only one that is indeterminate in size. Of the corresponding terms in the numerator and denominator, $n(\mathrm{H}) < n(\mathrm{L})$, and $R(\mathrm{H}) < R(\mathrm{L})$, but $p(\mathrm{HO})\cdot q(\mathrm{HO}) > p(\mathrm{N})\cdot q(\mathrm{N})$.

If it is assumed that the subject works with only a single p(old) in structuring the decision process, then $R(\mathrm{H}) = R(\mathrm{L})$. The logarithmic terms do not then affect the relations between the variances. However, the other parameters do, and the predictions above still hold.

Received February 28, 1989
Revision received May 19, 1989
Accepted May 19, 1989 ■