

# Recognition memory for words and pictures at short and long retention intervals

ROBERT E. GEHRING, MICHAEL P. TOGLIA, and GREGORY A. KIMBLE  
*University of Colorado, Boulder, Colorado 80302*

In two experiments, subjects studied a long series of words and pictures for recognition. Retention intervals varied from several minutes to a few months. The complicated testing procedures in Experiment I required the use of a traditional correction for guessing to obtain estimates of subjects' memory performance. A comparable, but simpler, design in Experiment II permitted the calculation of sensitivity and bias measures. In both studies, pictorial memory was superior to verbal memory at all retention intervals tested, and this advantage was essentially constant over time. In addition, the experiments identified an increasing tendency to call verbal test items "old" over time. Bias scores in Experiment II revealed that subjects adopted a more lenient criterion in responding to words than to pictures, and increased leniency was noted for both item types over time. Explanations of the results are offered in terms of differences in initial encoding and of a loss of discrimination between experimental and extraexperimental materials.

Recognition memory has been shown to be better for pictures than for words up to retention intervals of about 2 weeks (Bloom, 1971; Corsini, Jacobus, & Leonard, 1969; Davies, 1969; Jenkins, Neale, & Deno, 1967; Shepard, 1967). One purpose of the present studies was to obtain information concerning pictorial and verbal recognition memory over a longer period of time. To this end, picture and word recognition were compared for intervals ranging from 10 min to 3 months in Experiment I and from 15 min to 2 months in Experiment II.

A further purpose of Experiment I was to determine whether qualitative changes in memory occur at long retention intervals. Conceivably, changes might occur in the direction of a more highly generalized representation of the materials (Bartlett, 1932; Carmichael, Hogan, & Walters, 1932). As probes for such changes, recognition test items in Experiment I included: (a) verbal synonyms of study items, (b) pictorial synonyms, (c) pictorial representations of study words, and (d) nouns describing study pictures, as well as repeated items and unrelated filler items.

The use of the various types of test cue proved unnecessary as it turned out. The results showed that the retention curves for pictorial and verbal materials

Experiment I reports the results of a dissertation submitted by the first author in 1973 in partial fulfillment of the requirements for the PhD degree at the University of Colorado. The author gratefully acknowledges comments, help, and advice from Stephen J. Young and from the members of his dissertation committee: Gregory A. Kimble, Bruce R. Ekstrand, Walter Kintsch, Eugene S. Gollin, and Verne C. Keenan. This work was sponsored in part by NIMH Grant MG 19577 and by NSF Grant GB 34077 X. Requests for reprints should be sent to Robert E. Gehring, Department of Psychology, Indiana State University Evansville, 8600 University Boulevard, Evansville, Indiana 47712.

were essentially the same over time regardless of the type of test cue.

Experiment II replicates the first study, using simpler and more typical materials, and sensitive measures of retention and decision criteria. Whereas various types of test items were employed in Experiment I, only old items and distractors were used in the second study. Confidence ratings obtained in the recognition test permitted the construction of ROC curves and the calculation of separate measures of detectability and bias. These measures both confirm and clarify the conclusions reached in Experiment I.

## EXPERIMENT I

### Method

**Subjects.** Subjects were undergraduate psychology students receiving course credit for participation. In a preliminary part of the study, 33 male and female subjects judged the equivalence between various pairs of items in order to insure that: (a) verbal synonyms were really comparable to pictorial synonyms, and (b) study-word/test-picture items were comparable to study-picture/test-word items.

In the main experiment, 124 female subjects signed up for and completed two prescheduled experimental sessions each, with the following six intervals between sessions: (a) 19-3 months, (b) 21-1 month, (c) 22-1 week, (d) 20-1 day, (e) 20-1 h, and (f) 22-10 min. Temporal ordering of conditions was as follows: (a) 3-month subjects were the first to study and the sixth and last to be tested; (b) 1-month subjects were the second to study and the fifth to be tested; (c) 1-week subjects were the third to study and the fourth to be tested; (d) 1-day subjects were the fourth to study and the third to be tested; (e) 1-h subjects were the fifth to study and the second tested; and (f) 10-min subjects were the sixth to study and the first tested. The totals of subjects completing the experiment reflect a return rate of 100% for both 10-min and 1-h conditions, progressively diminishing to a minimum 79% rate at the 3-month interval. Had subjects been purely randomly assigned to conditions, it seemed on the basis of pilot study that return rates at the longer intervals would have been less than those actually realized, resulting







STUDY ITEMS	TYPES OF ITEMS	RELATED TEST ITEMS
ROPE	18- REPEATED WORDS	ROPE
SOFA	18- VERBAL SYNONYMS	COUCH
FISH	18- STUDY WORD- TEST PICTURE ITEMS	
	18- REPEATED PICTURES	
	18- PICTORIAL SYNONYMS	
	18- STUDY PICTURE- TEST WORD ITEMS	AIRPLANE

Figure 1. Types of items (16 verbal + 16 pictorial fillers not shown).

in more bias due to the uncontrolled loss of a disproportionately larger percentage of subjects from the longer conditions. For statistical purposes, assignment to retention conditions can be considered virtually random. Subjects had essentially random information for choosing one experimental number in preference to another of the approximately 20 available, and subjects were not given any choice of experimental conditions.

**Materials.** The study and test materials included two sets of 140 slides each. Both sets contained 70 drawings of common objects plus 70 nouns with common-object referents. Figure 1 illustrates the various possible relationships between study and test items. The study items appear on the left of the figure and the related test items appear on the right. There were 18 of each of the following types of pairs: (a) repeated words such as "ROPE," (b) verbal synonyms like "SOFA" and "COUCH," (c) pairs composed of words for study followed by pictures on the test (exemplified by "FISH"), (d) repeated pictures, (e) pictorial synonyms illustrated by the two different types of leaf, and (f) items which were study pictures followed by test words. In addition, there were 16 filler words and 16 filler pictures on both lists. Fillers were words and pictures representing common objects but not closely related in meaning to any other items.

On both study and test lists, there were 14 blocks of eight items each randomly intermixed with 4 blocks of seven items each. Every block of eight contained one item of each type (repeated word, study picture/test word, study word/test picture, etc.). Each block of seven contained one item of every type except a verbal or pictorial filler. Within each block, items appeared in random order, with the restriction that there be no more than four pictures or four words in succession. The restriction insured that two items of a particular type (repeated word, pictorial synonym, etc.) could not appear in succession unless the last item in one block were by chance to be of the same type as the first item in the next block. The block arrangement assured equal distribution of all types of items

throughout the lists, though each subject had no way of knowing about the existence of the blocks.

The plan of the experiment required that the verbal and pictorial synonym pairs be equally synonymous, and that the word-picture and picture-word pairs also be equally good representations of each other. Therefore, in a preliminary study, subjects rated members of pairs of these types of items for percentage of equivalence. The results revealed that verbal synonyms and pictorial synonyms were equally synonymous (median equivalence rating: 81.97% for verbal synonyms and 83.48% for pictorial synonyms; range: 73.8%-93.3% for words and 76.7%-92.6% for pictures). In addition, the pictures and words were approximately equal in the rated adequacy with which they represented each other in the pictures-of-words and names-of-pictures items (median equivalence ratings: pictures of words 95.38%, names of pictures 94.70%; range: pictures of words 82.3%-95.4%, names of pictures 87.4%-99.8%).

**Procedure.** Prior to study, the subjects were told the entire testing procedure, and questions regarding the procedure were answered. During study, every item was presented for 5 sec. Immediately after study, subjects were told that testing would involve only recognition and that, therefore, thinking about or rehearsing items should not be attempted. At the test session, subjects had 8 sec to respond to each item. The task was to classify every test item as one of the six types illustrated in Figure 1, or else as a verbal or pictorial filler, by checking the proper column on a prepared answer sheet.

Subjects in the 10-min retention group received the same instructions as those in the other groups except that they were told the test would take place during the initial session. As soon as the study list was presented, these subjects were told that they would have only a very short break while the slide trays were changed, but that they could ask any questions about the testing procedure. The time between the end of the study list and the beginning of the test list was less than 1 min. The retention interval for any given item averaged about 10 min.

### Results and Discussion

A surprising and interesting finding involved a change in the pattern of responding over time, rather than any change in memory. From the immediate retention test to 3 months after study, subjects showed a large increase in the probability of, in effect, calling any verbal item "old" (by checking the "repeated word" column) from 20% to 30%. There was no apparent shift in response bias over time for test pictures, the comparable percentage values being 19% and 17%. This differential shift in response bias is highly significant. Analysis of variance for the interaction between the word/picture variable and the number of "old"/"new" responses over time shows the progressive increase of "old" responses and decrease in "new" responses for words, relative to pictures, to be highly reliable:  $F(5,118) = 37.195$ ,  $p < .001$ . All analyses of variance herein are based upon the unweighted means method of coping with unequal n (Winer, 1971, p. 402ff, p. 445ff).

Because of the response bias, there were more correct identifications of repeated words than of repeated pictures at the 1- and 3-month intervals. This consideration of hit rates alone might lead to an erroneous conclusion that verbal recognition performance catches up with and surpasses pictorial recognition performance over time. The falsity of such a conclusion can be demonstrated by a transformation or correction of the hit rate data into unbiased

retention functions for verbal and pictorial information.

Two possible transformations are (a) signal-detection measures and (b) a traditional type of correction for guessing. The assumptions which would be needed for signal-detection analysis using  $d'$  are violated in the present design due to the complex assortment of types of items. Specifically, the distributions of data for the various item types fail to meet the requirements of normality and homogeneity of variances. The possibility of obtaining confidence ratings, which might have yielded ROC curves and a bias-free nonparametric recognition score ( $A$ ), was rejected, since the experimental design was already complex and the recognition decisions themselves were complicated and probably taxing for subjects.

An application of a traditional correction for guessing was therefore made (cf. Kintsch, 1970). For each subject, the percentage of filler words falsely identified as either (a) repeated words, (b) verbal synonyms, or (c) study-picture/test-word items was subtracted from the percentage of hits for each of these three item types, respectively. Thus, the actual hit rates were diminished to allow for an estimate of the contribution of guessing or response bias. The same procedure was used to correct the pictorial data.

After this correction for guessing, clear and consistent results were obtained. Since these results were comparable for all types of items, all corrected data pertaining to verbal and pictorial memory were collapsed across type of study item and plotted in Figure 2. From this figure, it is clear that pictorial recognition is superior to verbal recognition at all intervals studied. This seems due to a relatively more effective encoding of pictorial than of verbal items into memory, inasmuch as the pictorial superiority holds even at the shortest (10-min) interval. Also, Figure 2 suggests that there is no systematic effect upon the difference over time.

Analysis of variance performed on the data plotted in Figure 2 indicated that the main effect of temporal interval was significant,  $F(5,118) = 311.32$ ,  $p < .001$ , as was the superiority in recall for pictures over words,  $F(1,118) = 1,446.51$ ,  $p < .001$ . The interaction shown by the converging functions in Figure 2 also proved significant,  $F(5,118) = 29.15$ ,  $p < .001$ . This is probably not due to any intrinsic difference between verbal and pictorial recognition memory. Rather, at the 3-month interval, verbal, but not pictorial, performance seems to be closely approaching the level of chance performance.

## EXPERIMENT II

### Method

**Subjects and Design.** The subjects were 100 male and female University of Colorado undergraduates, serving to fulfill a course requirement. Assignment to conditions was as in Experiment I. Regarding sequencing, (a) 30 subjects in the 2-month condition studied first and were tested third, (b) 30 subjects in the 1-month

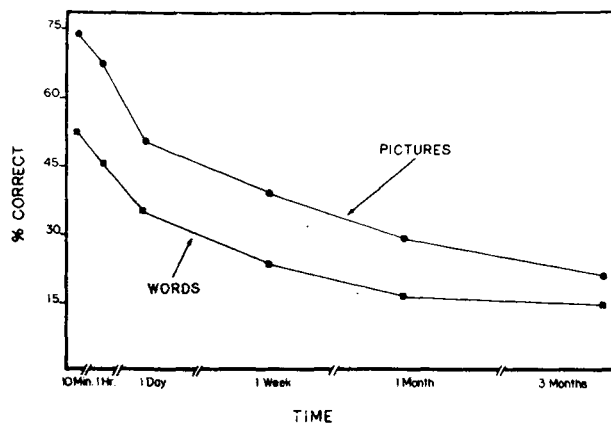


Figure 2. General superiority of pictorial performance over all retention intervals in Experiment I, after correction for guessing. The standard errors of the mean for the collapsed pictorial data over each increasing interval are 2.265, 2.815, 3.119, 2.978, 2.572, and 2.172, respectively; and for the verbal data, 2.466, 2.971, 2.625, 2.694, 1.890, and 1.910, respectively.

condition studied second and were tested second, and (c) 40 subjects in the 15-min condition studied third and were tested first. The above numbers of subjects reflect return rates of 100% for 15 min, 81% for 1 month, and 83% for 2 months.

**Materials.** A total of 210 experimental items, 105 pictures and 105 words presented on slides, were employed in the study. Most of these items were the same as those used in Experiment I. Substitutions were made, however, to equate for taxonomic category of items. The slides represented toys, animals, furniture, fruit, tools, clothes, and so forth. There were approximately equal numbers of old and new words and pictures of each of these categories in order to maximize comparability between pictures and words on both the study list and the test list.

Both training and test lists consisted of 140 slides, 70 pictures and 70 words. Of these, 35 pictures and 35 words were randomly chosen for inclusion on both lists, constituting a total of 70 old or target items. The 70 fillers on the study list and 70 distractors on the test list account for the remaining experimental items. Both study and test slides were randomly ordered, with the restriction that no more than four pictures or words appear consecutively.

Answer sheets were prepared prior to the test, providing spaces for 10 responses more than the 140 spaces that were used, in order that subjects would not change their response tendencies in anticipation of the end of the test. At the top of each answer sheet there was a 6-point rating scale with numeric values identified as representing the following confidence judgments: 0, absolutely certain new; 1, fairly certain new; 2, guess new; 3, guess old; 4, fairly certain old; 5, absolutely certain old.

**Procedure.** The subjects were instructed to study each slide carefully for a subsequent recognition test at their preassigned time. They were also informed that they would have to indicate their degree of confidence for each recognition judgment, but the details of the procedure were not explained until the time of the test. Following these instructions, the 140-item training list was presented once by a Kodak Carousel projector at a 5-sec rate.

Immediately following the study, the 15-min delay group received their answer sheets and listened to the testing instructions. The nature and use of the confidence rating scale was explained in detail. The subjects were told to use the scale fully and to indicate as accurately as possible their degree of confidence that each test slide was or was not one they had seen during training. They were also instructed to use the guess-old and guess-new categories when they were unsure of how to respond, since it was important to mark an answer for every test item. The other two groups returned to the laboratory either 1 or 2 months later and received the same test and instructions.

In the test, all the subjects were shown the second series of 140 slides. Each slide was presented singly for 8 sec, which was

adequate time for viewing the slide and marking a confidence response. To help subjects keep their places on the numbered answer sheet, the experimenter called the number of every fifth test slide.

**Results and Discussion**

As a first step in the analysis of the data obtained, separate ROC curves for words and pictures were plotted from the confidence ratings for each retention interval. These curves appear in Figure 3. For a quantitative comparison of pictorial and verbal recognition memory, areas (A) under the ROC curves were calculated following procedures outlined by Green and Swets (1966). This nonparametric measure was used in favor of the detection parameter *d'* because it provided a distribution-free measure of memory performance. A values were computed for each subject's two curves, and an analysis of variance was calculated on these scores. In Figure 4, mean area is plotted as a function of retention interval. It is clear from this figure (as well as Figure 3) that pictorial memory was superior at all intervals tested, an observation supported by a highly reliable main effect of experimental item,  $F(1,97) = 228.54, p < .001, MSE = .0003$ . The decrease in memory performance over time also proved highly significant,  $F(2,97) = 136.57, p < .001, MSE = .0001$ . The Word/Picture by Retention Interval interaction approached, but did not reach, significance ( $p < .10$ ), perhaps because the verbal retention at the maximum interval of 2 months was not approaching chance level

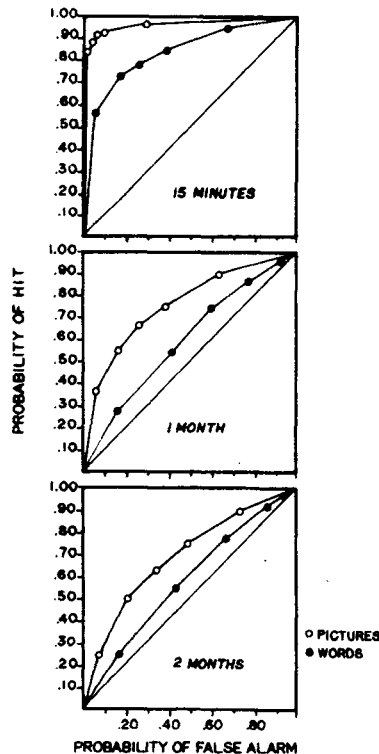


Figure 3. ROC curves for pictures (open circles) and words (filled circles) at the three retention intervals in Experiment II.

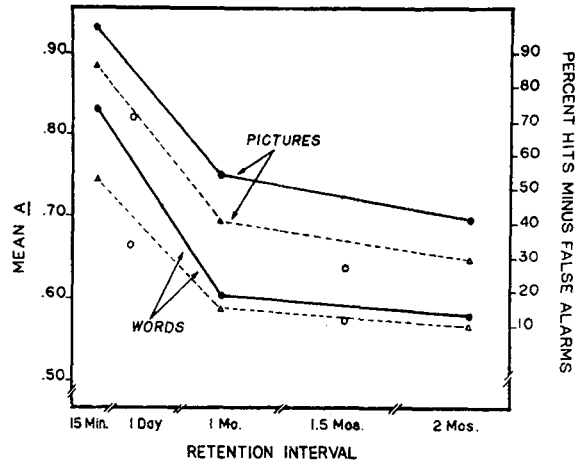


Figure 4. Retention curves for words and pictures, Experiment II. Solid lines represent mean area under the ROC curve. Broken lines represent retention with traditional correction for guessing. Points for 1 day and 1.5 months in the latter curves are from a separate study at Indiana State University Evansville. See text.

as it was at the 3-month interval in the first experiment.

In Experiment I, there was an increasing tendency over time to call any test word "old," but no such tendency occurred for the picture stimuli. A possible interpretation for such a response pattern is a shift in the criterion for responding "old" to words but not for pictures. For that reason we analyzed criterion changes for the results of this experiment.

As the first step for obtaining criterion measures, the hit rate and false alarm rate for both pictures and words were determined for each subject. This was achieved by collapsing the confidence rating scale into a yes-no scale. Then verbal and pictorial criterion scores were calculated for each subject using the percent bias formulas and methods described by Hodos (1970), wherein scores can range from -100% to +100%. On this scale, positive percentages indicate strict criteria and a more positive score indicates a stricter criterion. On the other hand, a more negative percentage indicates a more lenient criterion.

The mean bias scores for words and pictures appear in Table 1. The main effect of the word/picture variable was significant,  $F(1,97) = 30.20, p < .001, MSE = .205$ , as was the effect of retention interval,  $F(2,97) = 4.65, p < .025, MSE = .201$ . It is evident from Table 1 that subjects always set a stricter criterion for responding to pictures than to words.

Table 1  
Mean Bias Percentages at Each Retention Interval

Retention Interval	Type of Item	
	Words (Percentage)	Pictures (Percentage)
15 min	-7	+26
1 month	-26	+14
2 months	-30	+2

Table 2  
Percentage Distributions of Positive Responses  
as a Function of Time

Retention Interval	Word Item			Picture Item		
	Old (Hits)	New (FAs)	Total Called "Old"	Old (Hits)	New (FAs)	Total Called "Old"
15 min	79.2	26.8	106.0	93.0	6.9	99.9
1 month	74.6	57.1	131.7	66.4	26.1	92.5
2 months	77.7	67.0	144.7	63.7	34.6	98.3

Note—FA = false alarm

Over time, however, subjects' criteria became somewhat more lenient for both pictures and words.

The criterion measures in Table 1 obscure certain aspects of the subjects' performance. Table 2 presents percentages of "old" responses given to new words, new pictures, old words, and old pictures at each retention interval. The percentages were obtained by collapsing all ratings indicating any degree of confidence that the item was old.

These data are revealing in several ways. First, they are consistent with the previous finding that uncorrected measures of recognition memory for only the old words may surpass the same measures for pictures at long retention intervals. Second, these data show, as before, that there is an increasing tendency to identify verbal items as being "old," but no such tendency in the pictorial data. Third, by subtracting false alarms from hits, these data provide a traditional correction for response bias of the type employed in the first study. These measures appear as the open triangles in Figure 4. In order to provide a more complete picture of the retention functions, we have also included, as the open circles in Figure 4, some additional results obtained by the first author at Indiana State University. The materials used in this additional experiment were the same as those employed in the studies being reported here. There were 28 subjects in a 1-day retention group and 39 in a 1.5-month retention group. Clearly, the functions obtained by connecting the points obtained in the two experiments are similar to those obtained using the A measure. They also bear a striking resemblance to those presented in Experiment I.

### GENERAL DISCUSSION

The two experiments, though somewhat different in design and employing different indices of memory, support the same major conclusions: (1) recognition memory is superior for pictures at all intervals tested, (2) this seems to reflect a difference in initial encoding because the superiority of pictorial memory appears at the shortest interval tested, and (3) the difference in favor of pictures diminishes only slightly, if at all, over 3 months.

One interpretation of the overall pictorial

superiority is in terms of the extraexperimental environment of the subjects. All verbal items employed in the experiments were words that subjects had undoubtedly encountered many times prior to arrival at the laboratory. Many or most of the words would also be encountered during the longer retention intervals. While the pictures all depicted common objects and the subjects probably had seen pictures of most of these objects at one time or another, the particular style of the line drawing used would not have been encountered by subjects either prior to the experiments or during the retention intervals. Therefore, the pattern of false alarms in both experiments and criterion changes in the second study could be accounted for by a loss of discrimination between items in the experimental list and extraexperimental materials. This loss would be particularly true for the verbal items.

The most direct implication of the above interpretation is that frequency of usage of verbal and pictorial items would be an important variable. By comparison with the materials used in these experiments, more familiar pictures, such as naturalistic photographs, would be expected to be more similar to the types of pictures most commonly seen outside the laboratory and would produce a greater increase in false alarms. Less familiar words should produce a decrease in false alarms over time. These predictions can be tested by further studies.

### REFERENCES

- BARTLETT, F. C. *Remembering*. New York and London: Cambridge University Press, 1932.
- BLOOM, S. W. *Recognition memory for pictures and their word labels*. Unpublished doctoral dissertation, University of Rochester, 1971.
- CARMICHAEL, L. C., HOGAN, H. P., & WALTER, A. R. An experimental study of the effect of language on the reproduction of visually perceived form. *Journal of Experimental Psychology*, 1932, 15, 73-86.
- CORSINI, D. A., JACOBUS, K. A., & LEONARD, S. D. Recognition memory of preschool children for pictures and words. *Psychonomic Science*, 1969, 16, 192-193.
- DAVIES, G. M. Recognition memory for pictures and named objects. *Journal of Experimental Child Psychology*, 1969, 7, 448-458.
- GREEN, D. M., & SWETS, J. A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.
- HODOS, W. Nonparametric index of response bias for use in detection and recognition experiments. *Psychological Bulletin*, 1970, 74, 351-354.
- JENKINS, J. R., NEALE, D. C., & DENO, S. L. Differential memory for picture and word stimuli. *Journal of Educational Psychology*, 1967, 58, 303-307.
- KINTSCH, W. *Learning, memory, and conceptual processes*. New York: Wiley, 1970.
- SHEPARD, R. N. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 1967, 6, 156-163.
- WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.