Bulletin of the Psychonomic Society
1976, Vol. 7(1), 75-77

Word recognition as a function of retrieval processes

JAN C. RABINOWITZ and ARTHUR C. GRAESSER II
University of California, San Diego, La Jolla, California 92093

The influence of organization on recognition was assessed using three types of recognition tasks. A triple-recall procedure provided a criterion for classifying individual items according to different degrees of organization. Recognition performance was found to parallel that of recall, when tested 1 week later. There was poor, but above chance, discrimination for items that were not accessed during recall. For recalled items, recognition improved as the degree of retrievability increased.

In the standard free recall paradigm, subjects study a list of words and later recall the list in whatever order the words occur to them. Some of the words are accessed at recall, while others are available but inaccessible; that is, the items are stored but the retrieval mechanism is incapable of accessing them. It is clear that some of the nonrecalled words are stored; subjects are able to recognize them when given a recognition test. Moreover, a single recall protocol does not tap all of the items that can be accessed through recall. Recent studies (Mandler, Worden, & Graesser, 1974; Tulving, 1967) have used a triple-recall procedure, where subjects give three successive recall outputs after list presentation. In this procedure, subjects recall roughly the same number of words in each output, but the particular words recalled vary from output to output. These results suggest that the strategy of retrieving items varies from recall to recall. We can now speak of items that are organized either stably or variably for access at recall, along with items that are clearly inaccessible.

This study was designed to assess recognition performance for items that vary in their accessibility or degree of organization (cf. Mandler, 1972). In order to isolate items that differ in degree of organization, we used a triple-recall procedure. On the basis of this triple recall, items are classified as common (C) items (recalled in three outputs), Variable 2 (V2) items (recalled in two), Variable 1 (V1) items (recalled in one), or inaccessible (I) items (never recalled). We expect that recognition performance for C, V2, V1, and I items would be consistent with that of recall, namely C > V2 > V1 > I. It may be argued that this ordering can be attributed to the experience of items during the act of recall (Broadbent & Broadbent, 1975). However, a study by Rabinowitz, Mandler, and Patterson (Note 1) supports the position that it is the accessibility of items, rather than exposure to experience, that determines later recognition. Of additional interest are the I items. Given that these items cannot be retrieved during recall, is it possible to recognize them 1 week later?

Three methods of testing recognition were used in this study. a new/old recognition test provides d′ scores for C, V2, V1, and I items when compared to an overall false alarm rate for new (N) distractor items. However, when recognition performance is very good, even if not perfect, it is often difficult to detect differences among classes of items. To avoid such possible ceiling effects and indeterminate d′ scores, we also used a 2-alternative forced-choice procedure where words from each class were paired with words from every other class. In this case, subjects give a relative judgment, often comparing two well-organized items. In both the new/old and 2-alternative forced-choice recognition tasks, we infer differences between types of items from a series of absolute (binary) judgments. In order to obtain these differences more directly, we used a functional measurement task (Anderson, 1974) in which subjects give numerical confidence ratings for the C, V2, V1, I, and distractor items. Furthermore, these confidence ratings can be related on an interval scale.

METHOD

Subjects

Twelve University of California, San Diego undergraduates participated in the experiment as paid volunteers.

Materials and Apparatus

Six lists of 50 words were randomly generated from an original item pool of 428 high-frequency (Kučera & Francis, 1967) nouns of six letters or less.

All phases of the experiment were controlled by a PDP-12 computer. The computer generated random orderings of items for each subject, presented the items for acquisition and recognition on a remote video screen, and recorded the subjects' responses. The subjects' recall protocols were tape recorded.

Design

In the first session, each subject studied one practice list of male names and three of the six experimental lists. Each of the six lists

This paper is sponsored by George Mandler, who takes full editorial responsibility for its contents. This research was supported by the National Science Foundation Grant No. GB20798. Requests for reprints should be sent to Jan Rabinowitz, Department of Psychology, C-009, University of California, San Diego, La Jolla, California 92093.

was presented an equal number of times, counterbalanced across subjects and the three serial positions.

The subjects returned 1 week later for the second session, in which they expected more list learning and recall. Instead, they were given three recognition tasks: a functional measurement scaling task, a standard new/old recognition test, and a 2-alternative forced-choice (2AFC) reognition test, in that order.

One of the three presented lists and one of the three nonpresented lists were assigned to each recognition task. Old words came from only one of the presented lists for any given task. The subjects were told which one of the three presented lists was being tested before each recognition task. The assignment of lists to tasks was counterbalanced such that every list in each of the three presentation positions was used in each recognition task an equal number of times.

Each recognition task was preceded by a practice task of the same type and procedure, using the practice list of first names and new names. Subjects took a 5-min break between each recognition task. All subjects were run individually.

Procedure

In the first session, each list was presented twice in succession in two unique random orders. Words were presented for 3 sec with a 1-sec pause between presentations. Subjects pronounced each word aloud as it was presented. After the last word was presented, a three-digit number appeared on the screen and the subject counted backwards by 3s for 45 sec. This interpolated counting task served as a recency buffer. Subsequently, the course of the triple recall went as follows: recall, counting backwards (45 sec), recall, counting backwards, recall. The computer signaled the subject when to initiate each recall by displaying RECALL LAST LIST. After 30 sec in which no words were recalled, the computer signaled the subject to count backwards by displaying a three-digit number.

In the functional measurement (FM) recognition task, subjects made confidence judgments for pairs of words. On each trial, the subject judged a pair of words by placing a pin somewhere along a 200-cm response scale. The ends of the scale were labeled "Absolutely certain that both words were presented" and "Absolutely certain that both words were not presented." Subjects were instructed to judge each pair by averaging how confident they were for each individual word.

One C word and one N word were selected at random to serve as the context words for the scaling. Twenty-five words were randomly selected to be scaled, five words of each item type. In the event that there were not five words of a given item class, additional words were selected from the other item types. Thus a 2 by 25 factorial design was used; each to-be-scaled word being paired once with each of the context words. Subjects were run through three replications of the design. The first replication was treated as practice and was not analyzed. The presentation order of the 50 pairs was randomized anew for each replication.

All of the 50 old words from one of the presented lists and all 50 words from a nonpresented list were used in the new/old recognition test. Subjects decided whether each word was old (had been presented the week before) or new (had not been presented the week before) and responded by pusing one of two buttons. Items were presented individually in a random order.

Five words were selected at random from each of the five item classes for use in the 2AFC recognition test. In the event that there were not five words in a particular item class, additional words were selected from the other classes. Each of the 25 words was paired with each of the other 24 words yielding 300 pairs. These pairs were presented side by side in a random order. Subjects were instructed to choose the word they thought most likely to have been presented the week before and to indicate their choice by pressing one of two buttons.

RESULTS AND DISCUSSION

The major concern in this experiment is the recognition of the five categories of words: the C, V2, V1, and I items, along with the distractors (N items). However, not all subjects produced both V1 and V2 items, but all subjects did have some variable items. Therefore it was necessary to combine V1 and V2 items into a single V category for all three types of recognition tests. One subject had to be eliminated from the analyses due to an equipment failure.

In all three recognition tasks, recognition performance was compatible with that of recall, i.e., C > V > I. However, differences among these three types of items were not statistically significant in all three tasks. The FM procedure provides ratings of each word along an interval scale, given a nonsignificant Context Word by Item Word interaction (cf. Anderson, 1974). This was the case for all 11 subjects. For each subject, we computed a mean scale value for the items in each of the four categories. The mean of the means for the C, V, I, and N items were 156.0, 153.7, 126.9, and 118.5, respectively. A high scale value indicates a high confidence that the word was presented. An ANOVA showed a significant effect of category type, $F(3,30) = 32.44$. (A .05 level was used as a criterion for significance in this and all subsequent tests.) There were no significant differences in the scale values between N and I items, $t(10) = 1.42$[1], and V and C items, $t(10) = .56$. The only significant stepwise comparison on the C-V-I-N continuum was the V-I difference, $t(10) = 7.62$.

In analyzing the new/old recognition test, we obtained d′ scores for C, V, and I items for each subject. Hit rates for the C, V, and I items (.91, .89, and .55, respectively) were compared with the false-alarm rate for the N items (.35). The mean d′ scores for C, V, and I items were 1.97, 1.80, and .57, respectively. While there was a main effect of item type, $F(2,20) = 20.33$, the effect was due to the difference between V and I items, $t(10) = 5.30$; C and V items did not differ significantly, $t(10) = .56$. The new/old recognition test replicates the FM task in showing a significant difference between the V and I items and not between the C and V items. However, the new/old recognition task did show a difference between I and N items; the .57 d′ score for I items was significantly different from zero, $t(10) = 3.74$.

In analyzing the 2AFC recognition data, d' scores were computed for six pairs of item types: C-V, C-I, C-N, V-I, V-N, and I-N. Table 1 shows the d′ scores

Table 1
D' Values for the Forced Choice Recognition Task

Nonchosen Word	Chosen Word I	V	C
N	.43	1.33	2.02
I		.92	1.53
V			.58

Note. The table shows the d' values for choosing a word type shown in the columns over a word type shown in the rows. C = common, V = variable, I = inaccessible, and N = new.

for these six pairs. An overall ANOVA showed significant differences among the six pairs, $F(5,50) = 9.60$. The pattern of d′ scores perfectly supports our expected underlying continuum of C > V > I > N. For example, in keeping with perfect transivity, subjects chose C items over N items more often than C items over I items, and C items over I items more often than C items over V items. All stepwise differences in the d′ scores for the pairs shown in Table 1 were found to be statistically reliable. Of particular interest is the fact that the C items differed significantly from V items, which contrasts with the results from the new/old recognition test and the FM procedure. The 2AFC test was the only test that showed reliable differences among all classes of items.

Of further interest in the 2AFC recognition data is the additivity of the d′ scores along the C-V-I-N continuum. This can be demonstrated by using the three d′ values for the three single-step pairs, C-V, V-I, and I-N, to predict the other three pairwise differences. For example, the predicted d′ value for C-I is 1.50, the sum of V-I (.92) and C-V (.58). This compares favorably with the obtained value of 1.53. Similarly, the predicted values for the V-N and C-N pairs are 1.35 and 1.93, respectively, which are comparable with the obtained values of 1.33 and 2.02. Thus d′ distances between adjacent items along a single underlying C-V-I-N scale accounts well for the observed d′ values for item pairs.

In conclusion, recognition performance closely parallels that of recall. This finding is not surprising if one adheres to the view that both recall and recognition are different measures of the same memory phenomena, i.e., the occurrence of an item in a specific context (e.g., Mandler, 1972). The large differences consistently observed between V and I items support the notion that the organization underlying recall is a critical determinant of recognition performance. The importance of organizational variables in recognition is further demonstrated by the C-V differences. Just as C items are accessed more often than V items in the recall protocols, in recognition the C items are also more likely to be accessed. However this difference is apparently small, reaching statistical significance only in the 2AFC test.

In adition to the organization underlying recall, the presence of a copy cue (presentation of the target word) provides some cues for access to the memory trace. This is supported by the discrimination of I items from N items as observed in the new/old and 2AFC recognition tests. This discrimination is rather poor in comparison with the items that were accessed in recall. Thus just as organizational processes are the critical determinants of recall, organizational processes are also the main determinants of recognition performance 1 week after presentation.

REFERENCE NOTE

1. Rabinowitz, J. C., Mandler, G., & Patterson, K. *The structure of recognition: Effects of recall and generation*, in preparation.

REFERENCES

ANDERSON, N. H. Information integration theory: A brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco: Freeman, 1974.

BROADBENT, D. E., & BROADBENT, M. H. P. The recognition of words which cannot be recalled. In P. M. A. Rabbitt and S. Dornic (Eds.), *Attention and performance V*. New York: Academic Press, 1975.

KUČERA, H., & FRANCIS, W. *Computational analysis of present-day American English*. Providence: Brown University Press, 1967.

MANDLER, G. Organization and recognition. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press, 1972.

MANDLER, G., WORDEN, P. E., & GRAESSER, A. C. II. Subjective disorganization: Search for the locus of list organization. *Journal of Verbal Learning and Verbal Behavior*, 1974, **13**, 220-235.

TULVING, E. The effects of presentation and recall of material in free recall learning. *Journal of Verbal Learning and Verbal Behavior*, 1967, **6**, 175-184.

NOTE

1. All t values reported in this study are planned specific comparisons based on an F with 1 df.

(Received for publication October 6, 1975.)