Memory & Cognition
1982, Vol. 10 (1), 33-42

The word-frequency paradox in recognition

GEORGE MANDLER, GEORGE O. GOODMAN, and DEANNA L. WILKES-GIBBS
University of California at San Diego, La Jolla, California 92093

High-frequency words are recalled better than are low-frequency words, but low-frequency words produce higher hit rates in a recognition test than do high-frequency words. Two experiments provided new data on the phenomenon and also evidence relevant to the dual process model of recognition, which postulates that recognition judgments are a function of increments in item familiarity and of item retrievability. First, recall and recognition by subjects who initially performed a single lexical decision task were compared with those of subjects who also gave definitions of high-, low-, and very low-frequency target words. In the second experiment, subjects initially performed either a semantic, elaborative task or an integrative task that focused attention on the physical, perceptual features of the same words. Both experiments showed that extensive elaborative processing results in higher recall and hit rates but lower false alarm rates, whereas word frequency has a monotonic, linear effect on recall and false alarm rates, but a paradoxical, curvilinear effect on hit rates. Elaboration is apparently more effective when the potential availability of meaningful connections with other structures is greater (as for high-frequency words). The results are consistent with the dual process model.

A major challenge to any theory of recognition of prior occurrences is the word-frequency effect. What is challenging is the paradoxical finding that high-frequency words are recalled better than low-frequency words but in episodic recognition, hit rates for low-frequency words are higher than those for high-frequency words. The earliest study to report the word-frequency effect in recognition appears to be that of Gorman (1961), which was later generalized by Schulman (1967). It should be noted, though, that the paradoxical reversal is not simply a function of some uniqueness of low-familiarity words or of the testing procedure. Glanzer and Bowles (1976), for example, have shown that false alarms demonstrate the dominance of high-frequency words; they are higher for high- than for low-frequency words.

The present paper is concerned with providing more evidence for the generality of the phenomenon across different kinds of processing conditions, and also with relating these to the dual process model of recognition (see Mandler, 1979, 1980, 1981).

The dual process model states that the recognition of prior occurrence is the result of two additive and separate processes: familiarity and retrievability. We have assumed that the familiarity of an event is determined by the integration, perceptual distinctiveness, and internal structure of that event. Familiarity is affected by the frequency of exposure of the event and by the amount of attention expended on the event or item itself. Retrievability, on the other hand, is determined by interevent relationships and the elaboration of the target event in the context of other events or items.

Retrievability could not account for the better recognition of low-frequency words, since their recall is worse than that of high-frequency words. We have proposed an incremental effect of presentation on familiarity, that is, that the original presentation produces a larger relative increment for low- than for high-frequency words (Mandler, 1980). We assume that each presentation and processing of an event adds some specified degree of familiarity to the target. The effective familiarity value of a word will be the ratio of that increment to the sum of the base familiarity value of the event plus the increment, and this will be larger for low-frequency than for high-frequency words. One of the consequences of this proposal is that words must have some perceivable baseline value of familiarity so that the ratio of increment to base familiarity can be evaluated. Thus, the paradox of the word recognition effect should be demonstrable for low- and high-frequency words, but not for nonwords. For the latter, recognition cannot be based on the ratio between the increment and the base familiarity, and it may well be based on sheer familiarity. In any case, the recognition of nonwords should not be better than that for high- or low-frequency words, and we have included such items in the experiments reported below.

Similar arguments involving the improved discriminability of low-frequency words after exposure have been suggested by Glanzer and Bowles (1976) and by Kinsbourne and George (1974). Glanzer and Bowles have offered a semantic analysis of the word-frequency effect. They assume that low-frequency words have

The research reported here and the preparation of this report were supported by National Science Foundation Grant BNS 79-15336. George Goodman is now at Bell Laboratories, Piscataway, New Jersey. Requests for reprints should be sent to George Mandler, Center for Human Information Processing, C-009, University of California, San Diego, La Jolla, California 92093.

0090-502X/82/010033-10$01.25/0

fewer semantic features (meanings) than do high-frequency words and that during presentation, some constant number of such meanings are marked. They further assume that in recognition, "the key factor is the proportion of marked meanings in the total number of meanings" (Glanzer & Bowles, 1976, p. 30). This proposal is, of course, also a version of a differential incremental model. However, Glanzer and Bowles' model requires the unlikely assumption that both recall and recognition are entirely a function of a semantic analysis.

The particular task used to present items is likely to interact with subsequent recognition performance. It is important that subjects not be given any information contributing to the items' discriminability other than the increment in familiarity. If, for example, subjects are told that the items will be tested later for recognition or recall, we know that different kinds of processing are likely to result (cf. Tversky, 1973).

False alarm rates for words of varying frequencies would be useful in examining word-frequency effects on recognition. Since no extensive analyses of such false alarm rates have been presented in previous research, the present experiments were designed to permit such analyses.

We decided to require a lexical decision ("Is the item a word?") for the input task in Experiment 1. Lexical decisions are likely to have a high probability of "yes" responses for both high- and low-frequency words; they need relatively few processing resources (as compared with a definitional task, for example), and they are unlikely to form the basis of subsequent discriminations. Thus, deciding whether or not an item is a word should not make it easier to differentiate old and new items in a subsequent recognition test. And finally, at least at the intuitive, phenomenal level, it seems likely that deciding whether or not a string of letters is a word does not necessarily involve knowing what the word means. Thus, we expect that lexical decisions can and will be made to a large extent on the basis of the familiarity of the words and that access to meanings will be minimized.

In Experiment 2, we explored some of the consequences of the suggestion that intraitem integration contributes to the familiarity value of an item, whereas extraitem elaboration is the basis for retrievability. In light of the work by Craik and his associates, it is generally reasonable to assume that elaboration involves the relations between the target item and other events stored in memory (cf. Craik & Lockhart, 1972; Craik & Tulving, 1975). The integration of an event, which deals primarily with the perceptual characteristics of the item, involves operations similar to those suggested by Craik's notion of shallow, superficial (e.g., phonetic) processing. We have previously suggested that while elaboration also involves integration of the event, if for no reason other than the mere activation of the representation of its perceptual features, it is not at all clear how integrative activity might affect elaboration (and, consequently, retrievability).

Finally, the two experiments provide some evidence of generality for the word-frequency effect, in that the recognition of words of varying frequencies was tested under four different processing conditions.

EXPERIMENT 1

Experiment 1 was designed primarily to replicate the word-frequency effect, with respect to both recall and recognition. However, it included three important variations. First, in addition to high- and low-frequency words, we included a group of very low-frequency items. These were English words, but they were selected so that they would not be recognized as such (i.e., they had a low or zero level of baseline familiarity). Second, in contrast to previous studies that used two-alternative forced choices (e.g., Glanzer & Bowles, 1976), we obtained separate hit and false alarm rates for words in the three frequency groups. Finally, we used two different tasks in the initial exposure of the words. After the lexical decision trials, one half of the subjects were also given a definition (meaning retrieval) task. This manipulation not only should show differential effects on recall and recognition but also provides differential attention to the items, which, as we have indicated above, should affect the increment in familiarity.

It should be emphasized that we are concerned only with the difference between the two conditions as it is generated by the additional and increased elaboration produced by the definitional requirement. That difference is likely to be due to both the repetition of the items and the increased elaboration in the definition condition. The intent here is to produce a difference, not to locate its source.

Method

Design. Four groups of eight subjects each participated in the experiment. Presence or absence of a meaning retrieval (definitional) task and order of recall and recognition tests were between-subjects variables, and three levels of word frequency were the within-subjects variable. All 32 subjects first performed a lexical decision task requiring judgments of the word/nonword distinction. In this task, each subject was presented 20 words from each of the high- (H), low- (L), and very low- (VL) frequency lists. Each subject received a unique, random selection of items in the lexical decision task, and this individualized list was presented in a new random order for subsequent tasks (meaning retrieval and recognition). In the lexical decision task, reaction times and errors were recorded for each of the 60 (H, L, and VL) items.

Following the lexical decision trials, 16 of the subjects were dismissed for the day, and the other 16 subjects performed a meaning retrieval task. Subjects were given a new random ordering of the 60 items they had seen previously in the lexical decision task, and they were asked to give a short definition of each word. Reaction times, as well as accuracy (correct, incorrect, no definition), were recorded. These subjects were then dismissed for the day.

Twenty-four hours after the initial task(s), all subjects returned to the laboratory. Half were given a recognition test followed by recall, and the test order was reversed for the other 16 subjects. In the recognition task, subjects were given the 60 old items plus 60 new items (all randomly arranged).

Subjects and Materials. Thirty-two undergraduate students at the University of California, San Diego, participated in the experiment as a part of a course requirement. Three pools of 40 words were created for each frequency level. H words ranged from 200 to 787 occurrences/million, with a mean of 323 occurrences/million. All of the L words had frequencies of 1 occurrence/million (Kučera & Francis, 1967). A VL list was selected by random sampling from the unabridged *Oxford English Dictionary*. Only words that did not appear in Kučera and Francis, that were English, and that were unknown to the experimenters were selected. All the words in the three lists were from 4 to 10 letters in length, with mean lengths of 6.12, 6.00, and 6.02 for H, L, and VL words, respectively.

Procedure. All aspects of the experiment were controlled by a PDP-12/30 computer, and the items were displayed on a DEC VR-12 CRT display. Responses in the lexical decision and recognition tasks were made by pressing one of two buttons separated by 4 in. The buttons were labeled WORD and NONWORD for the lexical decision task and OLD and NEW for the recognition test. Midway between the two response buttons was a READY button, used by the subject to initiate trials. Response buttons were counterbalanced; even-numbered subjects responded WORD and OLD on the left key, and odd-numbered subjects made those responses on the right key. For the meaning retrieval task, verbal responses were monitored by the experimenter and manual responses were made only on a key marked RECALL. All three buttons carried the appropriate labels during the experimental session.

Upon arriving at the laboratory, subjects were told that the experiment was designed to study the way people retrieve information about words and language. The subject was told that he/she would first be asked to make decisions whether or not a particular letter string was a word. Those subjects who later performed the meaning retrieval task were also informed that they would subsequently be asked to give the meanings of a series of words.

For the lexical decision task, subjects were told that they must decide as quickly and accurately as possible whether the letter strings they would see were or were not words in the English language. It was explained that the word READY would first appear on the screen and that pressing the READY button would make the signal disappear and be replaced 2 sec later by a string of letters. If this string was an English word, they were to press the button marked WORD; if not, they were to press the button marked NONWORD. Test stimuli remained on the screen until a response was made or until 5 sec had elapsed. Then the word READY reappeared to signal the next trial.

In the meaning retrieval task (performed by half the subjects), the temporal relation between the READY signal and the word was the same as that in the previous task. Upon presentation, the subject was to try to retrieve the meaning of the word. When sure of the word's meaning, he or she was to press a button marked RECALL and to give a brief definition. The word remained on the screen until a response was given or until 10 sec had elapsed. Following a postresponse interval of approximately 20 sec, the READY signal reappeared on the screen. It was emphasized that no speed stress was involved in this task and that a reasonable amount of time would be given to make a response. If a subject was unable to give the meaning of an item, he/she pressed the RECALL button and said, "I don't know that word." The experimenter monitored all responses in an adjacent room.

At the end of the first session, the subjects were asked to return in 24 h for more tasks of the same kind, "designed to increase the reliability of our measures." When they returned, half the subjects were given a recognition test, followed by recall; for the other half of the subjects, the order of these tests was reversed.

For the recognition task, subjects were told that they would again be presented a series of words. They were to decide whether a particular letter string had been included in the lexical decision task the day before. If they thought the letter string was an item from that task, they were to press the button marked OLD; if not, they were to press the button marked NEW. Both speed and accuracy were emphasized in the instructions, and subjects were told that all the items they had seen in the lexical decision task would be presented, as well as some they had not seen previously in the context of this experiment. Otherwise, the procedure was identical to that of the lexical decision task. Each subject was presented all the items in the H, L, and VL pools (20 old and 20 new items each).

In the recall task, each subject was given paper and pencil, and was asked to recall all of the items from the lexical decision task. Approximately 5 min were allowed for this task.

Results

Since we assumed that a familiarity judgment is an important determiner of a lexical decision, all analyses of H and L words were conditionalized on items' having been called words in the lexical decision task. Thus, the data are relevant only for items that were actually recognized as "words." In fact, a large percentage of both H and L items were called "words," the proportions being .98 and .87, respectively. There were no significant variations in these proportions among subgroups, nor did analyses including all items show any significant variation from the data presented here.[1]

Of the VL items, only 7% were called "words." Since one of the reasons for including this set of words in the experiment was to determine the effect of nonword presentation on recognition and recall, all analyses using VL words were based on the 93% that were not called words in the lexical decision task.

We shall first discuss the overall effects of the experimental manipulations on hit rate, false alarm rate, and recall, shown in Figure 1.[2] An analysis of variance was performed on all three measures, with the following

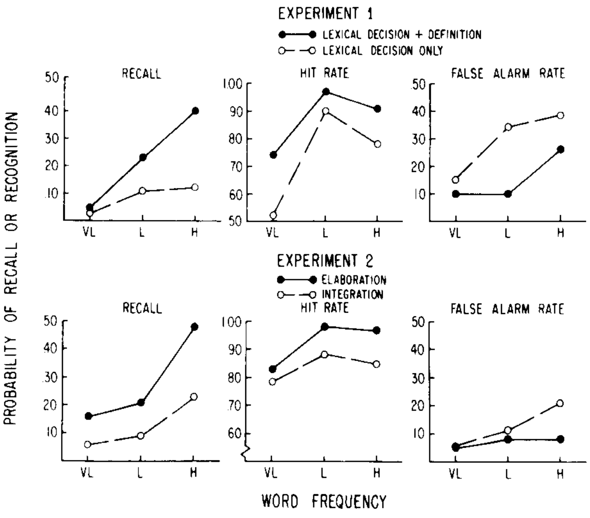

Figure 1. The effect of word frequency and processing instructions on recall, hit rates, and false alarm rates in Experiments 1 and 2. V, L, and H refer to very low-, low-, and high-frequency words.

design: presence (the definition group) and absence (the regular group) of the meaning retrieval task, two levels of test order (recall first vs. recognition first), and three levels of word frequency (VL, L, and H).

For the recall test, all three variables were significant.[3] The order effect showed that recall prior to recognition was at a mean proportion of .10; after recognition, it was .20 [F(1,28) = 34.39, MSe = .008]. Test order did not interact with either of the two other variables. Group (regular vs. definition) and word frequency were significant [F(1,28) = 58.95, MSe = .008, and F(2,56) = 50.00, MSe = .008], but so was their interaction [F(2,56) = 16.29, MSe = .008]. The data in Figure 1 show that, while the additional definition task produced much better recall and recall increased with word frequency, the differential effect of definition increased as a function of word frequency.

For hit rate, only group and word frequency showed significant effects [F(1,28) = 22.09, MSe = .02, and F(2,56) = 53.85, MSe = .014]. For the regular group, the mean hit rate was .73, but for the definition group it was .87. The hit rates for the three word classes were .63, .94, and .84 for VL, L, and H words, respectively. Even though these two variables did not interact significantly, we have shown the hit rate breakdown for them in Figure 1 for comparison purposes. The most important finding here is that the VL items did not behave like the L words; they did not show higher recognition accuracy than the H words.

For the false alarm data, group and word frequency were significant [F(1,28) = 13.89, MSe = .033, and F(2,56) = 26.72, MSe = .012], as was their interaction [F(2,56) = 7.04, MSe = .012]. These data are shown in Figure 1. False alarm rates increased with word frequency and were higher in the regular than in the definition group, but the difference between these two groups was greatest for the L words and smaller for the VL and H words.

Discussion

Overall, the free recall data are consistent with general expectations. A preceding recognition test increases recall significantly, and the additional definitional task produces more recall than just a prior lexical decision task.

These recall data must, however, be supplemented by intrusion rates (i.e., incorrect recalls), particularly in light of a previous study in which we studied the effect of prior recognition tests on the recall of categorized lists. We found that the increase in category-related intrusions was at least as great as the increase in recall (Mandler & Rabinowitz, 1981). In the present study, there was a mean number of 1.97 intrusions/subject, or 17% of the mean total recall. The intrusions were about evenly divided between internal intrusions (i.e., distractors from the recognition test) and external intrusions (i.e., new words), with .91 from the former and 1.06 from the latter source. However, the presence of the prior lexical definition task affected primarily the external intrusions (means of 1.88 with and .24 without the prior task); for the internal intrusions, the two means were 1.00 and .82, respectively. On the other hand, the intervening (prior) recognition task affected primarily the internal intrusions, with means of 1.69 and .13, respectively; for the external intrusions, these means were .94 and 1.18, respectively. The significant interaction between word frequency and the order effect is shown in Figure 2. The interaction [F(2,56) = 4.854, MSe = .388] shows that internal intrusions increase with word frequency but that the effect is primarily due to the intervening recognition test. We shall return to these results later, but for the time being, we can conclude that the increase in recall as a result of the intervening recognition test occurs at the cost of a large increase in intrusions.

That the definitional task does produce better recall than the lexical decision task alone is to be expected on the basis of the greater elaboration required by the former (cf. Craik & Tulving, 1975). However, this extra processing was much more effective as word frequency increased. This suggests that increasing the elaboration of a target item depends on the number of other events, related to the target item, that are available in semantic storage. High-frequency words have more "meanings" than do low-frequency words (see Glanzer & Bowles, 1976). As a result, the definitional instructions provide more of an opportunity to activate potential retrieval cues and interitem relations for high-frequency words than for low- and very low-frequency items. We note

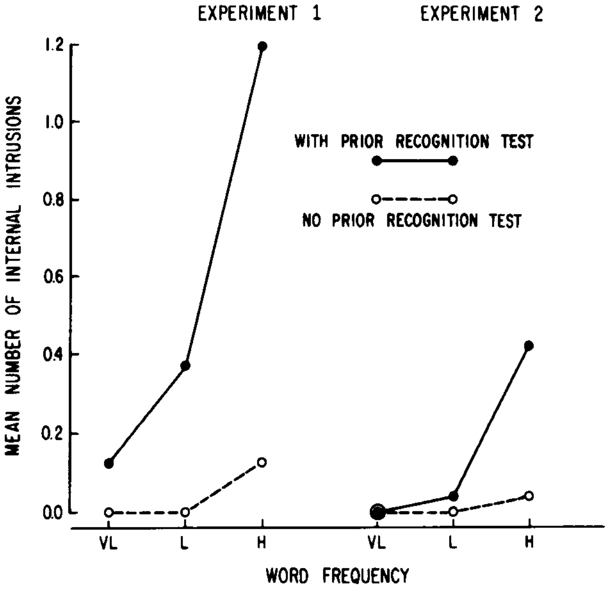

Figure 2. The interactive effects of word frequency and presence or absence of a prior recognition test on internal intrusions in recall for Experiments 1 and 2.

that the definition group not only was required to process the items more extensively but also had more total experience (time) with the target words. The effect of time is, of course, dependent on what is done with it, but we shall show similar effects in Experiment 2, in which total time with the target items was held constant.

The definitional task requires more processing and therefore produces greater retrievability, as well as greater incremental familiarity of the items. However, the effect of word frequency on hit rates is quite different from its effect on free recall. The highest hit rate is produced by low-frequency words, demonstrating, of course, the word-frequency effect. Given the absence of an interaction between word frequency and group (regular vs. definition), this result contrasts sharply with that from the free recall test. Increased processing does not interact linearly with greater meaningfulness of the items, which suggests that the effect of such processing differs for the retrieval required in recognition from that required in recall. That, of course, is exactly the argument that motivates the incremental model, and the notion that familiarity, or, specifically, incremental familiarity, is needed for an explanation of the effect of word frequency on recognition performance. In other words, the increased processing of the definitional task, as contrasted with the lexical decision task alone, provides both better retrievability and differential familiarity increments. The former is most important for recall, and the latter, for recognition.

False alarm rates increase monotonically with word frequency and behave more like recall than like hit rate data, but, in contrast to both of these, they show higher rates for the regular than for the definition group. These data indicate that false alarm rates reflect the conceptual status of word classes. When subjects engage in a complex task, such as defining the target items prior to recognition, items that have undergone such processing can be distinguished from items (distractors) that have not. In contrast, we assume that the lexical decision task is based primarily on familiarity and, therefore, provides little basis for such a conceptual decision. As a result, the false alarm rate for the regular group reflects only the increasing familiarity of the word-frequency groups, whereas the lower false alarm rates for the definition group incorporate the discrimination based on the prior definitional task.

If the incremental judgment is based on some prior evaluation of a word's familiarity (or frequency) value, then the absence of such knowledge should inhibit application of an incremental rule. The fact that the VL words do not conform to the word-frequency effect offers support for this argument.

Underwood and Freund (1970) have shown an interesting reversal of the word-frequency effect using a two-alternative forced-choice recognition test. When high- and low-frequency words were tested with high- and low-frequency distractors, respectively, the low-low combination produced the usual better level of recognition than the high-high combination. However, when high-frequency words were tested with low-frequency distractors and low-frequency words with high-frequency distractors, the effect was reversed. High (old)-low (new) pairs were superior to low (old)-high (new) pairs. Given that in the two-alternative forced-choice paradigm, individuals probably respond to the absolute difference between the recognition probabilities of the two words in the pair, the current theory also predicts this interesting result. In the high-high combination, the difference between the recognition probability for high (old) items (given their relatively high initial familiarity values and higher retrievability) and high (new) items (also with high familiarity values) should be smaller than the difference between low (old) and low (new) items. In the latter case, the incremental ratio for low (old) items is high, the basic familiarity of the new items is low, and the resulting difference is relatively large. Similar arguments apply to the reversal phenomenon, in which the high-low pairs produce better performance than the low-high pairs. Here, the difference between the high (old) and the low (new) items is relatively great, whereas the difference between low (old) and high (new) items will be relatively small. However, it is not necessary to use the two-alternative forced-choice paradigm to demonstrate this effect. Our present data show a similar result, using hit rates and false alarm rates for high- and low-frequency words. The relevant group is the definition group, which is treated somewhat similarly to a learning-test group used by Underwood and Freund. The d′ for the low (old)-low (new) comparison is 3.28; for the high (old)-high (new) comparison, it is much lower, 2.29. But the d′ value for the high (old)-low (new) comparison is 2.95, whereas it is only 2.62 for the low (old)-high (new) group. The effect is primarily a function of the relative false alarm rates (and therefore, presumably, the base familiarity values) of high- and low-frequency words. Glanzer and Bowles (1976) provided direct evidence for this argument: In a forced-choice test, subjects selected high (new) over low (new) items. These data are also inconsistent with Underwood's (1971, p. 330) statement that for "low frequency . . . words, the frequency of the distractors is of little consequence."

The finding that a meaning analysis following the lexical decision task produced higher recognition probabilities than did the lexical decision task alone seems to be in direct contradiction to a counterintuitive result reported by Kinsbourne and George (1974; see also Eysenck, 1979). These investigators found that both low- and high-frequency words produced lower recognition probabilities when subjects made concreteness ratings followed by a short (2-sec/word) exposure of a memory list, as compared with the memory task alone. Thus, two exposures of the words produced lower recognition probabilities than one exposure did. Our data show the opposite result. If a set of words is given two successive presentations, then the effective familiarity as a result of the first presentation will interact with the effect of the second presentation. If the increment in familiarity on the first presentation is large (e.g., because of extensive attention to the target words), then the first presentation will result in a greater degree of effective familiarity than will a small increment. Thus, following the first presentation, effective familiarities will differ and will, in turn, be affected by the size of the increment on the second presentation. In brief, for words with equal initial familiarity, the effective familiarity value after two presentations will be smaller if the first increment is large and the second small than if the first increment is small and the second large. In the Kinsbourne and George study, the first exposure consisted of a concreteness judgment, which presumably involved longer examination and manipulation of the items than did the subsequent brief memory study. In our experiment, the second definitional task clearly involved more attention to the item than did the initial lexical decision.

EXPERIMENT 2

In order to explore the generality of the word-frequency effect and the applicability of our model, we required subjects in Experiment 2 to process words of different frequency levels in an integrative or elaborative manner for relatively extensive time periods. Although it is clear how to manipulate elaborative activity, appropriate integrative activities are far less obvious. We asked subjects to manipulate and examine the physical, perceptual aspects of the items. However, these instructions were explorative, and whether such an analytic procedure would in fact produce greater integration (and familiarity) of the item as a whole was not clear. It might well be the case that integration requires attention to the event in a truly integrative, holistic manner in order to produce significant changes in familiarity.

Method

Experiment 2 was run in two sections. In Section 1, after item presentation a recognition test was given, followed by a recall test. In Section 2, the order of the tests was reversed; recall was given first and recognition second. In addition, a buffer task in Section 1 consisted of 5 min of conversation with the experimenter, whereas in Section 2, a 5-min space relations test was used. In all other aspects, the two sections of this experiment were identical.

Design. Four groups of six subjects each participated in each section. The between-subjects variables were delay between presentation and recognition test (5 min vs. 48 h) and the type of instructions given for the presentation task ("integration" vs. "elaboration"). Both of the instruction groups (integration and elaboration) were given 60 items (20 each of high, low, and very low frequency). Five minutes or 48 h later, they were given a recognition task of 120 items followed by a written recall test, or vice versa.

Subjects and Materials. The subjects were 48 undergraduate students at the University of California, San Diego, who were fulfilling a course requirement. The word pools used were identical to those in Experiment 1.

Procedure. All aspects of the experiment were identical to Experiment 1, except for the instructions to the subjects. Subjects in the integration condition were given the following instructions: "We are interested in how well people can describe and pay attention to the purely physical characteristics of words. Your task is to describe the physical, 'internal' characteristics of the words which will appear on the screen. Do not talk about the word's meaning, how or where it is used, its definition, or anything like that. Imagine you are on the phone to someone who wants to know exactly how a particular word looks, is spelled, and sounds. You will want to describe the letters–their actual shapes, sizes, type font etc., combinations of letters, the shape of the word, the sounds which various constituents make, and so on. Imagine that you wanted the other person to be sure to be able to recognize the word as a physical object. Describe it in that sense–concentrating only on the physical characteristics of the word."

Subjects in the elaboration condition were given the following instructions: "We are interested in the types of information which people extract from or consider important about words. What types of things are noticed about a word's meaning, its relation to other words, the way it is used, and so on. Your task is to describe in great detail the meaning of words. We do not want you to talk about its physical characteristics (for example, do not talk about its spelling or the way it sounds). Imagine that you are trying to tell somebody what the word means, that they do not know its meaning. You would want to give its definition, what particular category (e.g., 'Arm' is a 'part of the body') it might belong to, give them some examples of sentences in which it might be used, what some synonyms are etc. If you are not sure of a particular word's meaning, then talk about what you think it might mean."

Following these instructions, any questions were answered. Then, 15 practice items were presented on the video display. Each item appeared for 20 sec, during which time the subject performed the description task as instructed. The experimenter remained with the subject during these 15 practice trials, providing feedback on how well he/she was following the instructions. The experimenter then left the room, and the subject was presented 60 test items for 20 sec each. Descriptions were monitored over headphones by the experimenter, located in an adjoining room.

At the conclusion of the description task, all the subjects were engaged in conversation by the experimenter for 5 min (in Section 1) or performed a space relations test for 5 min (in Section 2). Following this buffer interlude, half of the subjects were told to return 2 days hence for more, "similar" tasks. The remaining 24 subjects continued with the experiment. Thus, either 5 min or 48 h after list presentation, subjects in Section 1 were instructed for the recognition task; instructions for recall were given at this point in Section 2.

In the recognition task, the procedure and instructions were the same as those in Experiment 1. For the recall task, subjects were asked to write down as many of the items from the original list as they could recall. Following the recognition and recall tests, the subjects were debriefed and dismissed.

Results

The data for all three dependent variables, shown in Figure 1, were subjected to an analysis of variance with two levels of test order (recognition preceding recall, and vice versa), two levels of test delay (5 min and 48 h), two levels of instructions (integration and elaboration), and three levels (within subjects) of word frequency (VL, L, and H).[4]

For the recall data, main effects of order, instructions, and word frequency were significant [F(1,40) = 11.86, MSe = .022; F(1,40) = 39.38, MSe = .022, and F(2,80) = 77.93, MSe = .01, respectively]. The order variable showed a significant interaction with delay [F(1,40) = 15.39, MSe = .022], and this interaction is shown in Figure 3, Panel A. As long as there was only a short delay between item presentation and test, recall was not affected by a preceding recognition test with the same items. However, 48 h after initial presentation, a recognition test prior to recall significantly increased recall compared with a test without prior recognition. To look at it in a different way, recall declined over 48 h when tested prior to recognition, but it improved over the level seen in the 5-min condition if it was preceded by a recognition test.

The instruction and word-frequency variables also showed a significant interaction [F(2,80) = 8.29, MSe = .01], as can be seen in Figure 1. Although recall increased with word frequency and was greater following elaboration than following integration, the facilitative effect of elaboration on recall was much greater for the H than for the L words.

There was also a weak triple interaction of instruction, word frequency, and delay [F(2,80) = 3.89, MSe = .01, $p < .05$]. It showed that the double interaction seen in Figure 1 was modified, so that recall after 48 h was lower than it was after 5 min, particularly for H words.

For the hit rate measure, only the instruction and word-frequency variables showed significant effects [F(1,40) = 11.67, MSe = .024, and F(2,80) = 21.18, MSe = .011]. The mean rate after integration was .84, whereas for elaboration it was .93. For the three levels of word frequency, mean hit rates were VL = .80, L = .93, and H = .91. Although the interaction was not significant, the relevant data are shown in Figure 1 for comparative purposes.

The data for the false alarm rates showed significant main effects for delay, instructions, and word frequency [F(1,40) = 21.78, MSe = .012; F(1,40) = 8.31, MSe = .012; and F(2,80) = 17.30, MSe = .006, respectively].

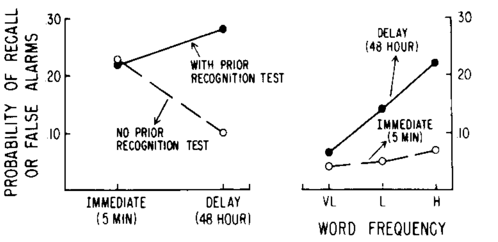

Figure 3. Significant interactions with test delay in Experiment 2. Panel A shows the interaction between delay and test order for recall. Panel B shows the interaction between delay and word frequency for false alarms.

The significant interaction between instructions and word frequency [F(2,80) = 8.75, MSe = .006] is shown in Figure 1. Integration produced higher false alarm rates than elaboration did, and false alarm rates increased overall with word frequency. However, the difference in false alarms between the two instruction conditions also increased with word frequency. In other words, in this comparison, the effect of word frequency on false alarms was primarily due to the integration condition.

The significant interaction between delay and word frequency [F(2,80) = 7.12, MSe = .006] is shown in Figure 2, Panel B. In this case, the word-frequency effect on false alarm rates was almost entirely due to the 48-h delay condition; it was practically absent when testing occurred immediately after presentation.

We note that, once again, the VL words did not behave like the L words. Hit rate for the VL words was at all times less than the hit rate for the L and H words. Thus, just as in Experiment 1, we assume that these VL words have no effective base familiarity against which subjects could evaluate an increment due to presentation.

Discussion

The effect of processing on recall is very similar to that found in Experiment 1. More extensive processing has a disproportionate effect on the high-frequency words. Note that the effect is similar despite the fact that in Experiment 1, the two conditions also differed in frequency and length of exposure to the items, whereas in Experiment 2, the exposure time was the same for both instruction conditions. Again, it is the availability of potential elaborative extensions that interacts with the processing variable. If an item is already richly interconnected in semantic storage, then elaborative processing will find and provide a greater variety of connections, which can then become available as cues in free recall. The triple interaction mentioned above amends this conclusion further by showing that the interaction between instructions and word frequency is present only for the immediate test condition. After a 48-h delay, the two functions shown in Figure 1 are parallel. Thus, the greater accessibility of high-frequency words due to elaboration is lost over 48 h.

The temporal effects on accessibility for recall are also shown in Figure 3. Although recall decays over 48 h, a preceding recognition test serves as an effective reminder for subsequent recall. It is interesting to note that the additional recognition test immediately after presentation does not affect recall; it is effective only if there has been some potential or actual loss of recallable items.

The intrusion data were similar to those found in Experiment 1. The mean number of intrusions was .81 items/subject. The contribution of external intrusions was .56; of internal intrusions, it was .25. Given the much smaller absolute level of intrusions in Experiment 2 than in Experiment 1, we report only the significant effects here, all for internal intrusions. The instructions variable was significant, with the mean for integration being .42; for elaboration, it was .08 [F(1,40) = 6.154, MSe = .072]. Figure 2 shows the interaction between word frequency and the presence and absence of the prior recognition test [F(2,80) = 5.214, MSe = .097]. Again, the effect of presenting the distractors increased with increasing word frequency. Thus, at a very low level, these results are similar to those found with categorized list; that is, the increase in recall following recognition is paralleled by an increase in intrusions from the distractor set (Mandler & Rabinowitz, 1981).

Hit rate data show the paradoxical interaction with word frequency, although not as pronounced as in Experiment 1, whereas the false alarm data again are more similar to free recall than to the hit rates in terms of major effects. The interaction between instructions and word frequency is, however, different in shape from that found between definition and word frequency in Experiment 1. The largest effect of word frequency on false alarms occurs for the high-frequency words, rather than for the low-frequency words as in Experiment 1. This difference seems to be due to the effect of elaboration (in Experiment 2) compared with definition (Experiment 1). The former has similar effects on high- and low-frequency words, whereas the latter produces an increase in false alarms for high-frequency words over low-frequency words by a factor of 2.5. This suggests that, in a recognition task, rejection of high-frequency distractors is easier after a simple definitional retrieval of the targets than after extensive elaborative retrievals. In addition, the strong interaction shown in Figure 3 between false alarm rates after 48 h and word frequency (as well as presumed meaningfulness) further underlines the sensitivity of false alarm rates to semantic factors. These comparisons between Experiments 1 and 2 are advanced with some caution because of the difference in instruction variables and timing, with all subjects in Experiment 1 receiving the lexical decision task and half of them receiving the definition task in addition. In Experiment 2, the two groups received only one of the two instructional tasks each.

We noted earlier that recognition can be affected by conceptual analyses of the words, in this case by recalling whether or not they have undergone some specific prior processing. This effect is apparent, particularly in the false alarm rates. Figure 3 shows that the false alarm values for high- and low-frequency words are always greater after the 48-h delay. This indicates that, after the delay, subjects cannot make the conceptual judgment (i.e., that old words belong to a category defined as having undergone a specific process, whereas new words do not). As a result, a word cannot be called old or new on this categorical basis, and the actual familiarity increment becomes a more powerful determinant of the recognition judgment.

GENERAL DISCUSSION

One of the most striking conclusions is illustrated by the six panels of Figure 1. They show the effects of two experimental variables, processing instruction and word frequency, on three different measures of memory. These effects are consistent across the two experiments and quite disparate across the three different measures. Elaborative processing, whether requiring a brief definition or more extensive meaning analyses, produces better recall and higher hit rates, but lower false alarm rates. Word frequency shows a linear, monotonic effect on recall and false alarm rates, but a curvilinear effect on hit rates. The two variables interactively affect recall and false alarm rates, but not hit rates. All of these effects hold for both experiments.

These configurations show that identical experimental variables can have greatly dissimilar effects on three different modes of retrieval. Such a conclusion is consistent with the notion that manipulations at input affect different aspects of the representation of an event. Evidence for this effect at output depends on which aspects of the representation are required by the memorial task. For example, in contrast to recognition of old events as indexed by hit rates, retrievability of the target event and test delay affect the retrieval operations that are indexed by both false alarms and free recall.

The fact that categorization of events (in this case, by type of prior processing) affects false alarms is consistent with previous research. We have suggested that categorization acts directly on the familiarity value of an item (see Mandler & Rabinowitz, 1981; Rabinowitz, 1978). Similar findings that a variety of attributes may contribute to the familiarity of an event have been reported by Herrmann, Frisina, and Conti (1978) and Macht and O'Brien (1980).

The exposure to a recognition test prior to recall significantly increases recall probabilities, but only if these memory tests are delayed. In the immediate condition of Experiment 2, recall is unaffected by the preceding recognition test. Thus, there is no automatic improvement of accessibility as a result of the recognition test. Rather, it seems that after access or retrieval cues have decayed or been lost over time, the recognition test serves to reinstate these cues. It "reminds" the individual how the list is structured and thereby improves retrievability.

The pattern of recall intrusions in the two experiments is quite similar. There are more than twice as many intrusions following recognition than prior to it, and there are also twice as many intrusions after the less elaborate processing tasks than following the more elaborate ones. While keeping in mind the differences between the two experiments, we can still note that in Experiment 1, with relatively less time for elaboration than in Experiment 2, there are also many fewer intrusions overall. This result is, of course, consistent with the data produced by the within-experiment analyses.

The ability to reject intrusions seems to increase with greater elaboration of the target items. A similar effect can be seen in the recall of categorized lists, in which one finds few if any extracategorical intrusions. In that case, elaboration of the items in terms of the specific categories that were used makes it possible to reject intrusions that do not belong to those categories. More important, we have been able to show here that the very large effect of a recognition test on intrusions with categorized lists (Mandler & Rabinowitz, 1981) can be found in a reduced fashion for lists of unrelated words. The intervening recognition test improves recall, but the cost is a significant increase in incorrect intrusions.

The dual process theory does not invoke any raw "meaning" as a variable in recognition. Retrievability depends on semantic factors, to be sure, but these factors by themselves do not determine recognition. Thus, critiques of "the role of meaning" in word recognition (e.g., Underwood & Humphreys, 1979) do not address the class of models advocated here. Nevertheless, it seems foolhardy to state that "word meaning is . . . infrequently used in making recognition decisions" (Underwood & Humphreys, 1979, p. 577). The present data suggest that false alarm rates are indeed affected by variables that could be classed as semantic, and we have shown previously that false alarm rates to categorically related words are four to five times greater than those to unrelated words in the recognition of categorized lists (Rabinowitz, Mandler, & Patterson, 1977).

The false alarm data support our general approach in two ways. First, taken together with the lexical decision data, the effect of word frequency on false alarms indicates that ordering of the three classes of word frequency is defensible. Frequency affects false alarm rates, and it is unlikely that this effect is mediated by the retrievability of these "new" words; familiarity is by far the more likely candidate for this effect. Second, the sensitivity of false alarm rates to conceptual, semantic variables supports the two-process model indirectly. We have assumed that recognition decisions are made on the basis of familiarity and retrievability and that the former takes place rapidly, whereas the latter is a slower process (see Mandler, 1980). Thus, if a decision on the basis of familiarity is impossible or difficult, then the semantic, conceptual factors that affect retrievability are likely to come into play. Just such a difficulty with sheer familiarity judgments should occur in the case of the new distractors, and as a consequence, the false alarm rates to these items are affected by semantic, conceptual variables. The recognition of old items, on the other hand, can be based on the more rapid familiarity judgments, and the conceptual variables play a less pronounced role.

We indicated earlier that the integration manipulation was speculative at best. We had hoped to show familiarity effects as great as those in the elaboration condition, but relatively fewer elaborative effects as indexed, for example, by recall. The results suggest that both familiarity and retrievability were lower for the integration condition than for the elaboration condition. In general, the data support the extensive literature on "depth of processing." We have extended it by showing that elaborative processing interacts with word frequency, or, as we have interpreted it, with the availability of elaborative structures. Previous studies (Jacoby, Bartz, & Evans, 1978; Seamon & Murray, 1976) have shown that elaborative instructions are more effective with material rated high in meaningfulness than with low-meaningfulness material. That effect is also likely to be due to the availability of more extensive interitem structures.

We began with a statement of the paradox between recognition and recall of high- and low-frequency words. We believe that the dual process theory, by stressing the different kinds of representations available for memorial processing, offers a reasonable solution to the apparent paradox. Traces of memorial events incorporate different kinds of processing products. These include elaborative structures, familiarity indexes, and incremental familiarity products. These various aspects of the memorial representation can be and are used differentially, depending on the requirements of the retrieval task. High-frequency words have a potential for extensive semantic elaboration. High- and low-frequency words both have available stored representations of their base familiarity values. And very low-frequency items (nonwords) have neither the potential for extensive elaboration nor discernible familiarity values. The result of these various attributes of stored events is a highly differentiated response to the requirements of different memorial tasks.

REFERENCES

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 671-684.

Craik, F. I. M., & Tulving, E. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 1975, **104**, 268-294.

Eysenck, M. W. Depth, elaboration, and distinctiveness. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory*. Hillsdale, N.J: Erlbaum, 1979.

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

Gorman, A. M. Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, 1961, **61**, 23-29.

Herrmann, D., Frisina, R. D., & Conti, G. Categorization and familiarity in recognition involving a well-memorized list. *Journal of Experimental Psychology: Human Learning and Memory*, 1978, **4**, 428-440.

Jacoby, L. L., Bartz, W. H., & Evans, J. D. A functional approach to levels of processing. *Journal of Experimental Psychology: Human Learning and Memory*, 1978, **4**, 331-346.

Kinsbourne, M., & George, J. The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 1974, **13**, 63-69.

Kučera, H., & Francis, W. *Computational analysis of present-day American English.* Providence, R.I: Brown University Press, 1967.

Macht, M. L., & O'Brien, E. J. Familiarity-based responding in item recognition: Evidence for the role of spreading activation. *Journal of Experimental Psychology: Human Learning and Memory,* 1980, **6**, 301-318.

Mandler, G. Organization and repetition: Organizational principles with special reference to rote learning. In L.-G. Nilsson (Ed.), *Perspectives on memory research.* Hillsdale, N.J: Erlbaum, 1979.

Mandler, G. Recognizing: The judgment of previous occurrence. *Psychological Review,* 1980, **87**, 252-271.

Mandler, G. The recognition of previous encounters. *American Scientist,* 1981, **69**, 211-218.

Mandler, G., & Rabinowitz, J. C. Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory,* 1981, **7**, 79-90.

Rabinowitz, J. C. *Recognition retrieval processes: The function of category size.* Unpublished doctoral dissertation, University of California, San Diego, 1978.

Rabinowitz, J. C., Mandler, G., & Patterson, K. E. Determinants of recognition and recall: Accessibility and generation. *Journal of Experimental Psychology: General,* 1977, **106**, 302-329.

Schulman, A. I. Word length and rarity in recognition memory. *Psychonomic Science,* 1967, **9**, 211-212.

Seamon, J. G., & Murray, P. Depth of processing in recall and recognition memory: Differential effects of stimulus meaningfulness and serial position. *Journal of Experimental Psychology: Human Learning and Memory,* 1976, **2**, 680-687.

Tversky, B. Encoding processes in recognition and recall. *Cognitive Psychology,* 1973, **5**, 275-287.

Underwood, B. J. Recognition memory. In H. H. Kendler & J. T. Spence (Eds.), *Essays in neobehaviorism.* New York: Appleton-Century-Crofts, 1971.

Underwood, B. J., Freund, J. S. Word frequency and short-term recognition memory. *American Journal of Psychology,* 1970, **83**, 343-351.

Underwood, B. J., & Humphreys, M. Context change and the role of meaning in word recognition. *American Journal of Psychology,* 1979, **92**, 577-609.

NOTES

1. If one assumes that the familiarity values of words are normally distributed and underlie the lexical decision judgment, then these data also support the notion that the average familiarity value of high-frequency words is greater than that of low-frequency words.

2. The reaction time data will not be reported in detail, since the only significant effect was that of word frequency in both experiments. The effect paralleled the hit rate data: Reaction times for low-frequency hits were fastest, and those for very low-frequency hits were the slowest. In Experiment 1, the means for VL, L, and H were 1.112, .791, and .900 sec, respectively [F(2,56) = 25.38, MSe = .034]; in Experiment 2, the means were 1.274, 1.086, and 1.124, respectively [F(2,80) = 13.79, MSe = .034].

3. Unless otherwise stated, all reported effects were significant at the 1% level or better.

4. The order variable is confounded with sections of the experiment, since Section 2 was run subsequently to rather than concurrently with Section 1. However, examination of error variances between the two sections and comparison of the order effects with those found in Experiment 1 suggest that no important or systematic differences existed between the two sections. They were, therefore, combined in a single analysis. The difference between the 5-min buffer tasks in Sections 1 and 2 is unlikely to have any systematic effect.

(Received for publication January 9, 1981;
revision accepted June 24, 1981.)