# Hierarchical Retrieval Schemes in Recall
## of Categorized Word Lists[1]

GORDON H. BOWER, MICHAL C. CLARK, ALAN M. LESGOLD, AND DAVID WINZENZ

*Stanford University, Stanford, California 94305*

These experiments investigate the effects of hierarchic organization of word-lists upon their free recall. *S*s recalled nested category lists presented either randomly or in a hierarchically organized manner. Recall was 2–3 times better with the organized presentation. Later experiments showed this effect (*a*) was similar with associative as well as conceptual hierarchies, (*b*) was attenuated with recognition tests of memory, and (*c*) could not be accounted for by associative "guessing." Another experiment demonstrated retroactive facilitation in recall of List 1 when List 2 contained the hierarchic superordinates of the words on List 1. Analyses suggest that the hierarchic principle was used as a retrieval plan for cuing recall, with generated candidates monitored for their list membership before being overtly recalled.

he following experiments were undertaken demonstrate the influence of structural anization upon the free recall of conceptual d hierarchies. The studies were initiated he belief that previous free recall experiats have investigated only relatively weak nipulations of this structural variable, and further belief that it should be possible to ange for a much more potent demonstran.

his research is most relevant to previous dies of free recall of categorized word lists. he list consists of several instances of ral taxonomic categories (e.g., animals, ors, occupations, etc.), it will be better alled than a comparable list of unrelated rds; moreover, in his recall, *S* tends to up together words from the same category. umber of variables inherent in this situation e been studied (e.g., Cofer, Bruce, and cher, 1966; Cohen, 1966). One of immete relevance to our studies is the presenion-order of a categorized word list. In cked presentation, the instances of a given egory are presented in adjacent temporal

order, whereas in *random* presentation the several instances of each category are separated by many intervening items in the exposure list. Psychologically, blocking of instances at input is more likely to lead *S* to discover the superordinate category (cf. Wood and Underwood, 1967) and thus it should mediate better recall. Although blocked lists are often recalled better than random lists, the effect has often been surprisingly small. For example, in the Cofer *et al.* study, blocking produced an average recall advantage of only 13% over a randomly ordered list. In Cohen's more recent and extensive studies, with long lists of many categories, there has been no effect at all of blocking vs. random input upon total recall.

If one believes that structural information about the input list is an important component of *S*'s ability to recall it, then the small or non-existent effect of blocking in these prior studies is rather disheartening. However, the possibility remains that the structural-organization variable has been only weakly and ineffectively manipulated in the prior studies. In consequence, one searches for a more potent way to manipulate this variable, and this is what we have done in the following experiments.
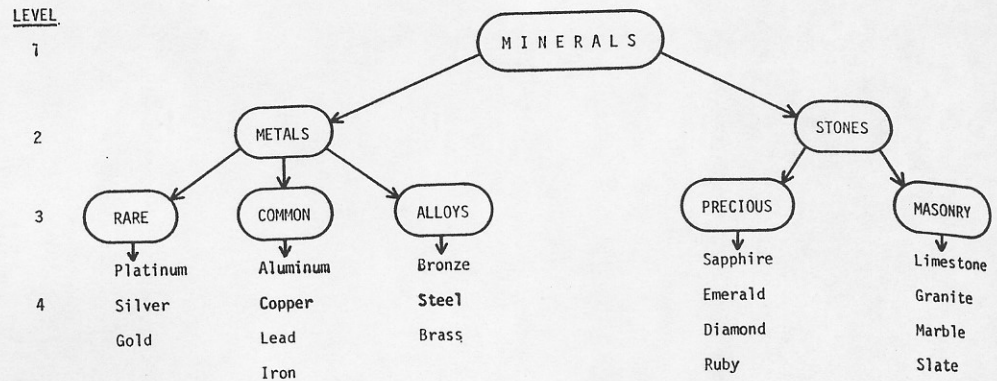
11

LEVEL



FIG. 1. The "minerals" conceptual hierarchy.

We have made three obvious changes from the prior studies of category blocking. First, the list to be recalled contains not only instances of a category, but also the category label itself. If recall of the category label serves as a cue for recall of its instances (cf. Tulving and Pearlstone, 1966), then the presence of the category name on the recall list should boost the recall of the structured material. Second, the word list is selected to be a hierarchically organized set of nested categories, so that the "instances" of a high-level category serve in turn as superordinates for still lower-level instances. We have constructed about eight of these conceptual hierarchies, one of which (for minerals) is shown in Fig. 1. The four different "levels" of the hierarchical tree are indexed by the left-hand column of numbers.

Such word trees have an obvious structure defined generally by the "class inclusion" relationship, and there would probably be fairly general agreement on the appropriateness of the words at different levels of the tree. Surprisingly, however, we have found that naive *S*s have considerable difficulty in trying to generate such trees given only the level-1 word and a general characterization of the target hierarchies to be generated (cf. our Exp. V later). The notion of "levels of class inclusion" turns out to be a terribly vague and imprecise concept when applied to natural

language, and in actuality there are a very large number of plausible options or branching points in generating such a hierarchical tree. Because of the open, nonexhaustive character of such conceptual word hierarchies, the one actually used can be considered as only one of many plausible trees that could have been developed from the level-1 word. In formal terms, the trees used in these experiments vary considerably from one another in the number of nodes at each level, and the use of nouns versus class-restricting adjectives at the higher levels. Many of the words have ambiguous meanings, so that the intended sense of the word would be established only by seeing it in relation to its hierarchical context.

The third change we have made from the prior studies is in the method of presentation. The prior studies have typically presented the lists one word at a time, whereas we present a complete set of words all at once for prolonged exposure. The one-at-a-time method has the advantage of operationally equating exposure time to each word, whereas the complete-presentation method can only equate total time for the entire set of words. However, the method of complete presentation makes it easier for *S* to discover more structural information about the input list than does the one-at-a-time method, and the former has been used for this reason.

## Experiment I

Our first experiment is in essentials a simple comparison of free recall of hierarchical word lists that are presented in a blocked as opposed to a randomized fashion. Four word hierarchies were learned concurrently. The $S$s in the Blocked condition were exposed to the four conceptual hierarchies organized in vertical trees as shown in Fig. 1. For $S$s in the Random condition, the same words were thoroughly scrambled, then assigned randomly to the nodes of four spatial trees. The spatial tree seen by Random $S$s had no apparent conceptual significance since the words located above and below a given node-word bore no obvious conceptual relation to the node.

A subsidiary factor that was varied orthogonal to the Blocked versus Random factor was Progressive parts versus Whole presentation of the hierarchies. In the Progressive parts condition, the level-1 and -2 words were presented on Trial 1, levels 1, 2 and 3 on Trial 2, then the full set (levels 1–4) on Trials 3 and 4. The thought here was that recall of the new level-4 words on Trial 3 might be appreciably aided by $S$ having already had practice in recalling the Blocked level 1–3 superordinate words. No such benefit would be expected for the Progressive parts $S$s receiving the Random list.

### Method

*Design and procedure.* There were six groups of $S$s, four learning by Progressive parts and two by the Whole presentation methods. Each $S$ had four input-output trials on a list composed eventually of four complete conceptual hierarchies comprising 112 words. Because of the much shorter times required for Trials 1 and 2 in the Progressive conditions, we were able to run a complete replication of the experiment with these $S$s within the experimental hour. The replication simply maintained the same progressive-parts conditions but with a different set of four hierarchies. The $S$s were run individually. A complete or partial hierarchy was shown printed on a large $5 \times 8$-in. card with a total study time calculated at two seconds per word on the card. The words on each card were arrayed in the form of a vertical tree but without the connecting lines and circles shown in Fig. 1. The same tree-form

was used even for the Random lists. After seeing four such cards, $S$ recalled the words orally in any order he preferred. Total time allowed for recall was five seconds per word on the input list. No $S$ ever needed this much time, and $S$ typically initiated the next input trial by indicating that he could recall no more.

The six experimental conditions were as follows: (*a*) Whole Blocked (WB): The four complete hierarchies (averaging 28 words per hierarchy) were presented in conceptually blocked fashion, one per card, for all four trials. (*b*) Whole Random (WR): The same set of 112 words, approximately 28 on each of four cards, were presented on each trial. The nodes in the tree on each card were filled by random selection without replacement from the 112 words, avoiding obvious conceptual relations amongst words in successive nodes of a tree. (*c*) Progressive Blocked (PB): On Trial 1, each exposure card contained only the levels 1 and 2 words (14 in total) in conceptually blocked fashion. On Trial 2, the appropriate level-3 words were added below the levels 1 and 2 words on each card (40 in total); and on Trials 3 and 4, the 72 level-4 words were added appropriately. Thus on Trials 3 and 4, the exposure conditions for PB $S$s were identical to those of WB $S$s. (*d*) Progressive Random 1 (PR1): This used the same trees as the WR condition, but with progressive exposure of levels 1 and 2 words on Trial 1, levels 1, 2, and 3 on Trial 2, then the full 4-level hierarchies on Trials 3 and 4. (*e*) Progressive Random 2 (PR2): This was similar to PR1 except the randomization of words was within, not across, levels of the conceptual trees. A level-$n$ word in a conceptual tree was used only at a level-$n$ node in a PR2 tree; the difference between such random trees and the conceptual trees is only in the vertical relationships among words above and below the nodal words. Thus the PR2 word sets being recalled over trials were identical to those of condition PB, except that the scrambling of words across cards did not allow the PR2 $S$s to easily discover the conceptual relationships. (*f*) Progressive Unrelated (PU): These $S$s learned 112 unrelated nouns by the progressive parts method, with 14 on Trial 1, 40 on Trial 2, and all 112 on Trials 3 and 4. The words were arrayed in vertical tree form as in the other cases. This condition was run for comparison to the PR conditions which had conceptual word lists. The 112 unrelated nouns were comparable in Thorndike–Lorge frequency to the words of the conceptual list, but because they were all nouns, tended to have higher "concreteness" ratings (Paivio, Yuille and Madigan, 1968) than did the words on the conceptual list.

As indicated before, there was time to complete four trials on a second list of four different hierarchies with $S$s in the Progressive conditions. The two sets of four hierarchies were animals, clothing, transportation, and occupations in set 1, and plants, instruments

body parts, and minerals in set 2. Half of the Ss in each progressive condition had set 1 first and set 2 second, while the remaining Ss had the reverse order. Similarly, four of the eight Ss in the WB and WR conditions learned set 1 and four learned set 2.

The Ss were 48 Stanford undergraduates fulfilling a service requirement for their introductory psychology course. There were eight Ss in each of the six experimental conditions.

## Results

For the four Progressive groups, there was no significant different in performance comparing their List 1 with List 2, so their performance on both lists was pooled to increase reliability of the following analyses.

TABLE 1

AVERAGE WORDS RECALLED OVER FOUR TRIALS

| Condition | Trials | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Whole: | | | | |
| Words presented | 112 | 112 | 112 | 112 |
| Blocked | 73.0 | 106.1 | 112 | 112 |
| Random | 20.6 | 38.9 | 52.8 | 70.1 |
| Progressive: | | | | |
| Words presented | 14 | 38.5 | 112 | 112 |
| Blocked | 13.0 | 34.5 | 93.2 | 108.4 |
| Random 1 | 8.6 | 16.0 | 34.9 | 46.3 |
| Random 2 | 8.4 | 15.5 | 33.9 | 52.6 |
| Unrelated | 9.9 | 18.7 | 32.4 | 51.1 |

The average recall scores for the four trials for the six conditions are shown in Table 1. Looking first at the two Whole-presentation groups, recall in condition WB is seen to be markedly superior to that of group WR. Mean recall is 3.5 times better in condition WB on Trial 1, and there is no overlap among the recall scores of the two groups of Ss on any trial. Recall of the 112 words in condition WB is almost perfect by Trial 2. Obviously, the structural organization of the blocked input list had a tremendously powerful effect on free recall in this situation.

Turning next to the four Progressive conditions in the lower portion of Table 1, recall in the Blocked condition is seen again to be much higher than in the comparable Random conditions. Recall scores in the two Random conditions, PR1 and PR2, do not differ significantly from each other, nor do they differ significantly from recall of Ss learning the unrelated word lists. All three of these conditions are significantly inferior to condition PB from Trial 2 onwards.

Performance on Trials 3 and 4 provides comparisons between the Whole versus Progressive parts methods, since on these trials Ss were exposed to the same full list. Within the Blocked conditions, WB significantly exceeded PB on Trial 3 simply because there was no variance for WB Ss on this trial (all recalled perfectly). Within the Random conditions, recall in WR exceeds that of either PR1 or PR2 on both Trials 3 and 4. Thus, in terms of overall recall in these conditions, it was clearly advantageous to present all words on all trials rather than progressively increasing the exposure set over trials.

One may ask the further question whether prior practice at recalling the level 1–3 words substantially improves S's ability to recall the level-4 words when they are presented for the first time on Trials 3 and 4. The level-4 recall data are presented for the relevant groups in Table 2.

Level-4 recall for group PB on Trials 3 and 4 significantly exceeds level-4 recall for group

TABLE 2

RECALL OF THE LEVEL-FOUR WORDS

| Condition | Trials | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| Whole: | | | | |
| Words presented | 73.5 | 73.5 | 73.5 | 73.5 |
| Blocked | 44.6 | 69.6 | 73.5 | 73.5 |
| Random | 13.2 | 22.8 | 35.2 | 44.8 |
| Progressive: | | | | |
| Words presented | 0 | 0 | 73.5 | 73.5 |
| Blocked | — | — | 56.6 | 70.4 |
| Random 2 | — | — | 16.3 | 30.0 |
| Unrelated | — | — | 13.8 | 27.6 |

B on Trials 1 and 2 ($t = 2.17$, $p < .05$), so at proactive facilitation exists when the new rds to be recalled fit as subordinates into e conceptual structure practiced on Trials 1 d 2. However, a similar facilitation appears mparing Trials 3 and 4 of group PR2 with ials 1 and 2 of group WR; a $t$ test for the el-4 recall over the respective trials for the o groups is significant, $t(14) = 2.63$, $p < .05$. r the unrelated words, comparing Trials 3 d 4 of group PU with Trials 1 and 2 of up WR, the difference in recall is in the me direction but not significant. From these veral comparisons, the overall conclusion is at recall of a new subset of words presented r the first time within a larger set is facilitated there has been prior practice at recalling e other words. This result may be specific ly to the complete presentation method used re, since that permits $S$ to spend relatively ore time studying the new words added to input list.

Further analyses of recall protocols demon- rated that WB $S$s were essentially using the ructural information in the blocked input a retrieval plan for generating their recall. rst, in condition WB, $S$s' recall was organ- d almost entirely according to conceptual tegories, whereas $S$s in condition WR tended show conflict between clustering according the underlying conceptual categories versus cording to the spatially contiguous words the randomized input. Second, all WB $S$s called their hierarchies from level-1 down rough level-4 words (from top down), and ver in the reverse order. In doing this, some in condition WB generated the conceptual es in breadth across levels (i.e., all the vel-$n$ words of a hierarchy before any of its vel-$n$+1 words), whereas other $S$s generated e hierarchy in depth, along a given (say) t-going branch before returning to higher des to generate in depth the other branches om it. Most WB and PB $S$s claimed that they uld have reproduced the hierarchical arrays actly as shown had they been permitted to ite their recall; given their perfect oral recall

and the facts above, this is a credible claim.

A question related to the results above is whether recall of a level-$n$ word for WB $S$s serves a cuing function, being correlated with recall of the level-$n$+1 instances nested within it. To answer this question, we considered the level-4 recall conditional upon level-3 recall by WB $S$s on Trial 1 where there were the most data and recall failures. The average prob- ability of recalling a level-4 word was .66 when its level-3 word was recalled, but only .30 when its level-3 word was not recalled. Moreover, when a level-3 word and at least one of its level-4 words were recalled, in 90% of these cases recall of the level-3 word preceded recall of the first level-4 word (not necessarily in immediate succession, however). Those cases where level-4 words were recalled without prior recall of the level-3 word were most often instances in which the level-4 words were cued (preceded) by a level-2 word, skipping the level-3 word. These results accord with the expectation that recall of a word served to cue recall of the corresponding words at lower levels. When a nodal word was not recalled, the entire "tree" developing out of that node was likely to be missing in recall.

In contrast to this strong dependence between recalls of conceptual levels 3 and 4 for the blocked hierarchies, recall of these same words was virtually independent for $S$s who viewed random hierarchies. Scoring Trial 1 recall of WR $S$s for the same words as above, the conditional probability of recalling a level-4 word was .23 when its level-3 word was recalled, and was .17 when its level-3 word was not recalled. These probabilities reveal no contingencies in excess of chance expectation on the null hypothesis of independent events, $\chi^2(1) = 1.5$, $p > .30$. Thus, the influence on recall of the moderately strong normative associations from a level-$n$ category word to a level-$n$+1 instance was considerably modu- lated by the method of presentation. If the associated words were presented in close temporal and spatial contiguity, so that $S$

might rehearse them together, they tended to be recalled together; if they appeared widely separated in a random list, then the same words were recalled almost independently in this case.

There were some small differences in the recallability of the eight conceptual hierarchies. Only the WB and WR conditions on Trials 1 and 2 were considered for these comparisons since, by the time the full hierarchies were presented, the PB $S$s were making very few errors. The proportions of words recalled for the various hierarchies averaged over Trials 1 and 2 for WB (and WR) $S$s were as follows: Plants, 87(26); minerals, 85(21); body parts, 84(34); transportation, 81(16); instruments, 79(27); occupations, 77(25); animals, 76(24); clothing, 72(29). The range of variation in recall of the word sets is about 15% for both groups, but the rank ordering of the sets by the two groups is quite discrepant (rank correlation $= -.14$). The poor correlation of the recall ranking of the eight hierarchies in conditions WB and WR may have been due to treatment effects or to low reliability since only four $S$s in a given condition learned a given hierarchy.

Perhaps the main fact to be remembered from Exp. 1 is that complete presentation of these conceptual hierarchies produced a tremendous facilitation of free recall relative to a random ordering of the same words. The effect is similar to what one finds in free recall of a sentence versus scrambled words (e.g., Miller and Selfridge, 1950). In both cases, recall depends upon the way the words are arranged, with a familiar structure in one case contrasted to unfathomable randomness in the other.

The notion of hierarchical organization as a recall aid is hardly original with us. Mandler (1967, 1968) has hypothesized that recall is largely a matter of subsuming list items under a hierarchical array of categories. Some earlier experimental work was done by Cohen and Bousfield (1956) who were interested in recall clustering produced by either (a) a 40-word list classifiable into eight independent categories of five instances each, versus (b) a dual-level list consisting of four superordinate categories each divided into two subordinate categories (e.g., *feline animals* and *canine animals*), with five instances in each of these eight subordinate categories. Cohen and Bousfield reported that total recall of these two lists was very similar (17.6 vs. 18.1 out of 40), but that recall clustering differed somewhat. The Cohen and Bousfield study differed in a number of respects from ours. For example, in their study the category names were neither presented nor recalled, and items were presented singly in random order. These procedures surely would reduce the probability that $S$ would notice and utilize the dual-level organization of the second list. In this regard, we may point out that our PR $S$s, who received a hierarchical word list but randomized, did not recall any better than our PU $S$s who had unrelated words. Obviously, the list-structure has to be discovered and utilized if $S$ is to derive any benefit from it in his recall.

Complete presentation of the Blocked hierarchies provides $S$ with a lot of structural information about the word list—he does not have to discover it for himself. This structural information in turn provides $S$ with a plan for retrieving the words from memory. The salient characteristics of a retrieval plan are that it tells $S$ where to begin his recall, how to proceed systematically from one unit to the next and to the words within each unit. The plan also helps $S$ to monitor the adequacy of his recall, helping him to identify where parts are missing and to identify when he has finished. It is plausible that a central ingredient in the present hierarchical retrieval plan is associative cuing of the words at level-$n+1$ by recall of the superordinate category at level-$n$. Having already recalled *masonry* (*stones*), $S$ is set to search for recency-tags on words in his associative hierarchy to this category (e.g., *granite*, *limestone*, *sandstone*, *flagstone*, etc.). It is further plausible to assume that the

implicit candidates generated to the category cue are recalled only if they pass a recognition criterion as having been on the list. Thus, of the four masonry-stone candidates, $S$ might overtly recall *granite* and *limestone* but not intrude *sandstone* and *flagstone* since the latter two words were not tagged as having been seen recently.

The detailed analyses of the recall protocols support this general view. When recalled, the superordinate category almost always preceded recall of its instances, and instance recall was considerably poorer if the superordinate was not recalled. Moreover, the intrusion data in group WB accord with the expectation from this category-cuing analysis. There were relatively few intrusions (20) on Trials 1 and 2 for group WB, but all were obviously intrusions of unpresented instances of a presented category, and these intrusions were "appropriately" placed in the recall order. Group WR had 26 intrusions of which 25 were clearly within the presented categories. In theory, these intrusions represent simple "false alarms" in recognition of a likely candidate which in fact was not on the list. The theory supposes that intrusions decrease and correct recalls increase with practice because of two factors: (*a*) The associations from category to presented instances become stronger, thus causing the latter to be more readily and reliably generated as candidates for recall to the former, and (*b*) the item-information that aids list discrimination improves, which information may be in the form either of an estimate of frequency or recency of experience or a trace-strength of the presented items.

## Experiment II

The next experiment investigated whether the large influence of Blocked structural-information upon recall could be replicated with a recognition test of memory. A frequent claim in the recent literature on memory is that recognition measures the amount of information stored independently of retrieval processes. In terms of the theory outlined above, the occurrence of an item on the input list causes information of some kind (a tag denoting recency, frequency, or trace strength) to be attached to the representation of that word in $S$'s semantic memory. In a test for recognition (or discrimination of list membership), the occurrence-information attached to the test-word in memory is consulted for a decision, yielding a judgment of recognition only if the information there exceeds a criterion (e.g., Bower, 1967; Parks, 1966). Such recognition tests, which directly provide the test word, clearly bypass the search and retrieval processes by which $S$ generates his recall. If the structural information provided by the blocked hierarchies has its main influence on the retrieval plans for recall, then one should find much less of an effect of structural information when memory is tested by recognition, which largely bypasses the retrieval aspects of the task. Accordingly, Exp. II was undertaken to see whether the large recall difference between the WB and WR conditions would be greatly attenuated in recognition tests.

### Method

*Design and procedure.* Each $S$ received two replications of a two-trial experiment, using the four-hierarchy sets 1 and 2 from Exp. I. There were eight conditions obtained by crossing Blocked vs. Random presentation with Recall vs. Recognition over the two trials of each learning task. Using the last-letter abbreviations L and N for recall and recognition, respectively, the four possible test sequences over two trials are LL, LN, NL, and NN. Subjects in conditions LN and NL were switched between Lists 1 and 2, as were $S$s in conditions LL and NN.

The four hierarchies of set 1 or set 2 were projected on four slides in tree form, at 56 sec per slide, and then $S$ either wrote his recall or took a written recognition test. He did not know during presentation how he would be tested. The recognition test used the "Yes-No" method: A sheet of paper given to $S$ contained the 112 list words scrambled amongst 112 distractors, and $S$ was told to check those words which he thought had been on the slides just studied. Half of the 112 distractors were unrelated nouns matched in Thorndike–Lorge frequency to the list words, while the other half of the distractors were conceptually

related words that could have fit into the conceptual hierarchies presented. The same set of words were used for both recognition tests in the NN condition, except they were listed in a different order on the two trials. Subjects were run in small groups of two-four, with test times sufficient for all *S*s to finish. A distinct instructional break was made between Lists 1 and 2. Half the *S*s were exposed to the Blocked lists and half to the Random lists of Exp. I for both lists and both trials. The *S*s were 64 Stanford undergraduates from the introductory psychology course with eight in each of the eight groups. Pooling across the two lists, there are 16 learning protocols for the eight experimental conditions.

*Results*

There were no significant differences between first versus second lists nor between sets

TABLE 3

PERCENTAGE OF WORDS RECALLED IN THE SIX CONDITIONS

| Condition | Trial | |
|---|---|---|
| | T₁ | T₂ |
| Blocked: | | |
| L L | .61 | .89 |
| L Nᵃ | .65 | — |
| N L | — | .88 |
| Random: | | |
| L L | .21 | .43 |
| L N | .29 | — |
| N L | — | .39 |

ᵃ L N denotes recall on T₁ and recognition on T₂.

1 and 2 of the hierarchies, so the data will be pooled over these variables. The first result of interest is shown in Table 3 which gives percentage correct recalls over Trials 1 and 2 for the *S*s that had recall tests. The recall differences between Blocked and Random *S*s are of similar magnitude to the differences observed in Exp. I, so that result is replicable. Further, recall on Trial 2 was approximately the same whether Trial 1 was a recall or a recognition test. Recall intrusions were less than 1% in all six conditions, but were categorical in nature. The Blocked *S*s gave 63 intrusions of which all were categorical;

the Random *S*s gave 31 of which 29 were categorical.

Recall of the eight conceptual hierarchies was separately scored, pooling all *S*s who were recalling them on either Trial 1 or 2. The recall proportions for the eight hierarchies averaged over Trials 1 and 2 for the Blocked condition (Random condition in parentheses) were as follows: Body parts, 81(34); plants, 75(35); minerals, 71(34); animals, 71(33); clothing, 71(35); transportation, 69(26); instruments, 69(32); and occupations, 61(34). The variation in recall of seven of the eight random hierarchies is too minuscule to give much of any correlation with recall of the eight blocked hierarchies (*r* = +.17). These scores may be compared to those obtained in Exp. I over the same trials but with fewer *S*s learning each hierarchy there (four *S*s instead of 16). For blocked presentation, rank orders of the recallability of the eight hierarchies in the two experiments correlate +.62; for random presentation, the two rank orders correlate +.48. It thus appears that the difference in recallability of the eight hierarchies is small but reasonably consistent within a given condition, but that the ordering differs consistently for blocked vs. random presentations. We shall look into this issue again in Exp. V in which several normative indices of the hierarchies are obtained.

TABLE 4

RECOGNITION HITS (H) AND FALSE ALARMS ON RELATED (R) AND UNRELATED (UR) DISTRACTORS

| Condition | Trial 1 | | | Trial 2 | | |
|---|---|---|---|---|---|---|
| | H | R | UR | H | R | UR |
| Blocked: | | | | | | |
| L N | — | — | — | .92 | .04 | 0 |
| N L | .88 | .05 | .01 | — | — | — |
| N N | .80 | .07 | 0 | .92 | .08 | 0 |
| Random: | | | | | | |
| L N | — | — | — | .78 | .06 | .01 |
| N L | .61 | .06 | .03 | — | — | — |
| N N | .60 | .10 | .02 | .80 | .19 | .07 |

The relevant results on recognition tests are shown in Table 4 for Trials 1 and 2. The three columns for each trial give proportions of (a) checking as "old" an old item (a hit), (b) falsely checking a conceptually related distractor, and (c) falsely checking an unrelated distractor. Hit proportions are based on 112 observations per $S$ per trial, whereas the false alarm proportions are each based on 56 observations per $S$ per trial.

Comparing entries within Table 4, the following conclusions are warranted: (a) False alarms were higher on related than on unrelated distractors; (b) hit rate on Trial 2 was unaffected by type of test on Trial 1, but the false alarm rate on Trial 2 was higher if Trial 1 was also a recognition test; and (c) $S$s having the Blocked input have better list discrimination than $S$s having Random input, as indicated by a higher hit rate and a lower false alarm rate.

The first conclusion, that related distractors are checked more often, is hardly surprising and can be handled theoretically in several ways. The second conclusion, that false alarms are highest in the NN conditions, is probably a result of using the same distractors on Tests 1 and 2. Thus, on Test 2 in the NN condition, a distractor arouses a familiar sense of having been seen before and is checked more often. These are ancillary findings.

The interesting result is the third, that recognition is better for Blocked than for Random $S$s. A $t$ test on recognition hits minus false alarms for all Blocked vs. Random $S$s on Trial 1 yields $t(62) = 5.94$, $p < .001$, and on Trial 2 yields $t(62) = 6.69$, $p < .001$. Comparing the averaged Blocked vs. Random results in Table 4 to those in Table 3 shows that the average recall differences were much larger than the recognition differences; recall differences were 38 and 48% on Trials 1 and 2 compared to recognition differences of 23% and 13% on the same trials.

We may ask whether the recognition test has attenuated the difference between Blocked and Random $S$s seen on the recall test.

However, an interpretative problem is that there is no atheoretical way to evaluate the significance of these differences due to Blocked vs. Random conditions in the recall vs. recognition measures. A model relating free recall to recognition is needed, but none can be stated now with any confidence. A simple threshold model, which supposes that $S$ recognizes all those items he could recall plus half of those he could not recall, leads to the formula $2N = 1 + L$, and this gives a fairly good prediction of the average hit rates (Table 4) from the average recall results (Table 3). However, it gives no prediction of false alarm rates, nor their ordering, nor does it provide any illuminating analysis of the retrieval processes in free recall.

The results of this experiment suggest that recognition of list membership of a word depends on the structure of the list as well as the words that were in it. In particular, recognition of a given word apparently depends upon the number and configuration of associations converging upon that word from other list words that have been recently primed. This view accounts for the effect of list structure upon hit rate, and it also explains why related distractors elicit more false recognitions than unrelated distractors, even in the Random condition. Underwood (1965) and Anisfield and Knapp (1968) have proposed a similar view of recognition memory.

An alternative interpretation of these recognition differences would attribute the effect to "unequal" exposure times to the individual words in the Blocked vs. Random hierarchies. Since only total time was controlled, $S$s in the Blocked conditions might have used the redundancy or predictability of the hierarchies to scan rapidly and more often over all the words. On the other hand, $S$s in the Random condition may have been much slower in reading the words, and for two reasons: (a) Recognition or reading times of words are known to be slower if the word sequence is unpredictable, so fewer words would be scanned in a fixed time, and (b) the

Random *S* might have used up exposure time trying to organize or systematize the random words he had already read, thus reducing the time he could spend processing later words. According to this view, then, the Blocked vs. Random difference in recognition resulted from less real exposure to the individual words in the latter condition. We are currently testing this interpretation with a method of presenting the hierarchy one word at a time, thus permitting control of *S*'s exposure time to each word. The *S* sees a complete hierarchical tree of unfilled, nodal circles with lines between nodes, and 28 successive slides show in systematic order one new nodal word for 2 sec, until each word in the conceptual hierarchy has been presented once. If differential exposure time to the individual words were the factor determining the results of Exp. II, then the Blocked vs. Random difference in recognition should vanish when exposure time is controlled in this manner.

## EXPERIMENT III

The next experiment investigates the retroactive effect upon free recall of a first list induced by the learning of a second list. Previous experiments by Tulving and Thornton (1959), and Postman and Keppel (1967) have shown retroactive decrements in free recall. For example, in the Postman and Keppel study, *S*s learned List 1, then had varying numbers of trials on List 2, then were asked to recall all the words in both lists. Recall of List-1 items was markedly poorer the greater the number of trials *S* had had on List 2. The authors discussed these results in terms of the unlearning (during trials on List 2) of the associations between the experimental context stimuli and List-1 responses.

These interference or unlearning effects have been produced with unrelated words in Lists 1 and 2. Our question here is whether we can produce just the opposite effect, retroactive facilitation, if the two word lists fit into the same conceptual hierarchies. In particular, retroactive facilitation of List-1 recall might occur if List 1 consists of level-4 instances of conceptual hierarchies, and List 2 consists of the levels 1–3 superordinates of these same hierarchies. In this case, when *S* is then asked to recall everything from Lists 1 and 2, his recall of the superordinates from List 2 should serve a cuing function in facilitating his recall of the List-1 instances.

## Method

*Design and procedure.* There were three experimental conditions, a rest control (C), relevant interpolation (RI), and irrelevant interpolation (II). All *S*s began with two trials on List 1 which consisted of 48 level-4 words from two conceptual hierarchies (e.g., *minerals* and *animals*). These hierarchies were so tailored that they each had six groups of four level-4 words. These four level-4 words (subsumed by one level-3 node) were presented together as a group for eight seconds by a slide projector. In List 1, there were 12 such slides arranged in a different random order for the two trials. Subjects gave their free recall by writing the list words, being permitted 3 min.

After two trials on List 1, the RI and II *S*s learned a second list while the C *S*s read *Peanuts* cartoons for a time equal to the longest interpolated learning interval. The interpolated list differed for RI and II *S*s. For RI *S*s, the interpolated list consisted of the levels 1, 2 and 3 superordinate words from the two hierarchies relevant for classifying the level-4 words which *S* had learned in his List 1 (e.g., *minerals* and *animals*). The relevance of List-2 to *S*'s List-1 words was not pointed out to him. For II *S*s, the interpolated list consisted of the levels 1, 2 and 3 superordinate words from two hierarchies which were irrelevant for classifying the prior List-1 words (e.g., *instruments* and *occupations*.) The two List-2 hierarchies for a given *S* were presented in tree form, one per slide for 18 sec (two seconds per nine words in each tree), and then *S* wrote his recall of List 2.

After *S* reached a criterion of one perfect recall of List 2, he was then asked to recall (in writing) everything he could from both Lists 1 and 2, starting with List 2 if he could. Rest control *S*s were also asked to recall List 1 at this time. Four minutes were allowed for this recall and protocols were collected. At this point, *S*s in groups C and II were given a sheet of paper containing the relevant superordinate categories for List 1 (i.e., levels 1–3 arrayed in tree form) and were asked to attempt a second recall, of the List-1 words only, since "the words on this sheet may help you to do better in your recall" of List 1. Three minutes were allowed for this cued recall by C and II *S*s.

After the procedures described above were completed, the entire experiment was replicated with

completely different word hierarchies partitioned to serve as List 1 and List 2 but with $S$ assigned to the same treatment condition as before. The order in which the various word hierarchies were used (first or second replication) was counter-balanced over $S$s.

The $S$s were 33 undergraduates fulfilling a service requirement for an introductory psychology course. They were assigned 10 to group C, 11 to group RI, and 12 to group II. The $S$s in a given condition were run in small groups of two to five at a time.

## Results

There was a significant improvement in $S$s' performance from replication-1 to replication-2. Considering List-1 performance within each replication, recall proportions averaged 52 and 68% on Trials 1 and 2 of replication-1, but were 59 and 79% on Trials 1 and 2 of replication-2. This general practice

TABLE 5

RECALL OF THE 48 LEVEL-FOUR WORDS OF LIST 1

| | Rest Control | List 2 | |
| --- | --- | --- | --- |
| | | Irrelevant | Relevant |
| 1. Trial 2 of List 1 | 35.3 | 34.5 | 35.6 |
| 2. Recall all | 34.9 | 34.5 | 42.2 |
| 3. Relevant probe | 40.3 | 40.1 | — |

effect, however, is orthogonal to the major factors of interest in this experiment, so results will be pooled over replications for the analyses which follow.

There were no significant differences among the three groups in their acquisition of List 1. The average List-1 recall scores on Trial 2 (the last acquisition trial) are shown in the first row of Table 5 and are obviously similar. Learning of the List-2 blocked hierarchies by RI and II $S$s was very rapid, with the median $S$ recalling his List-2 perfectly on Trial 2. These $S$s also recalled their List-2 perfectly on the "recall everything" test which immediately followed their criterion trial.

The results of major interest are the List-1 recall scores for the "recall everything" test given after List-2 learning; these are recorded in the second row of Table 5. Comparing this

recall to the corresponding recall scores on Trial 2 (row 1), we see that there was practically no change for the C and II $S$s, but a large increase for RI $S$s. The increase in recall for RI $S$s is highly significant: $t(21) = 7.95$, $p < .001$. On this later test, the gain in recall relative to what RI $S$s could have gained was 53%.

A similarly significant increase in List-1 recall occurred for groups C and II when the relevant superordinates hierarchies were given as recall probes on their last test (row 3 of Table 5). The II $S$s gained 42% and the C $S$s gained 39% of the responses they could have gained over Trial 2 with List 1. A comparison of the probe recall scores of C and II $S$s with that of RI $S$s on their "recall everything" trial yielded no significant differences, $F(2, 28) = 2.07, p > .10$.

The simplest explanation of these results would attribute the List-1 recall increments to the cuing function of the relevant superordinate words. That is, on Trial 2 with List 1, whole groups of level-4 words might not have been recalled because $S$ had not yet developed a scheme for cuing his recall of these. However, when the relevant superordinate categories are available on $S$'s recall sheet, either by $E$'s provision or by $S$ recalling them, they lead him to think about practically all the groups of level-4 words and to recall some words of these groups. This account is bolstered by analysis of the cluster-recall data. We scored a level-4 cluster as recalled if at least one of the four words in that group was recalled. For the clusters that were recalled, we also computed the mean words recalled per cluster. These measures for Trial 2 of List 1 and for the final recall trial are shown in Table 6 for the three treatment groups. The groups differ relatively little in the percentage of clusters recalled or in the mean words per recalled cluster. The effect of hierarchical cuing provided by $S$ or $E$ was to increase the percentage of clusters recalled but not the mean words per recalled cluster. Within each group separately, the increase in proportion

TABLE 6

PROPORTION OF CLUSTERS RECALLED AND MEAN
WORDS PER RECALLED CLUSTER ON TRIAL 2 AND FINAL
RECALL TRIAL

| Group | Proportion of clusters | | Words per cluster (of 4) | |
|---|---|---|---|---|
| | Trial 2 | Final | Trial 2 | Final |
| Relevant | .84 | .98 | 3.52 | 3.58 |
| Irrelevant | .89 | .99 | 3.20 | 3.35 |
| Rest Control | .86 | .98 | 3.40 | 3.42 |

of clusters recalled is significant ($t = 8.00$, 12.20, 5.80 for groups RI, II and C, respectively); but for no group was there a significant change in the mean words per recalled cluster.

We have attributed the retroactive facilitation of recall in group RI to the simple cuing effect of having the superordinates available at the time of recall. This is in accord with the results of Tulving and Pearlstone (1966) and Wood (1967b), who found more words recalled if *S*s were provided with relevant category cues at the recall test. An alternative interpretation might attribute the RI result to strengthening of List-1 words while *S* was studying the relevant List-2 superordinates. That is, while viewing the relevant superordinates, *S* might notice their relationship to List-1 responses and implicitly rehearse the List-1 responses, associating them with the superordinate context of List 2. This view would provide a "mediational" interpretation of our RI result similar to that provided for the retroactive facilitation observed with the A–B, A–B′ paradigm in paired-associates learning, where A is the cue, and B and B′ are associatively related response terms (cf. Postman, 1961). An implication of this interpretation, however, is that final recall for RI *S*s should have exceeded probed recall for the C and II *S*s, since the RI *S*s allegedly have had the benefit of extra rehearsal on List-1 words during their List-2 learning. In fact, however, there was no reliable difference between final recall for RI *S*s and probed recall for C and II

*S*s. Thus, this one implication of the mediational interpretation was not supported by the evidence.

EXPERIMENT IV

The first three experiments have examined the influence of structured input upon memory for conceptual hierarchies. The relationship between successive nodal words in a conceptual tree is roughly one of "class inclusion," or "instance of a superordinate category." In terms of associationistic psychology, however, these class inclusion relationships are only one of many bases upon which two words could become associated. Taking this broader view, then, a conceptual hierarchy is just a specific kind of associative hierarchy or network, one for which the associations between nodal words have approximately the same basis, namely, class inclusion. But this broader view immediately leads one to consider other kinds of associative trees—specifically, trees in which successive nodes are associated with one another, but in which the basis for the association may vary freely over different nodal pairs in the same tree.

The tree in Fig. 2, for *cheese*, will illustrate our points about associative hierarchies. We constructed such trees by starting with a level-1 nodal word, and recording three different but strong verbal associates to it as level-2 nodes. Each of these second-level nodes was then considered in turn and to each was recorded two different associates as level-3 nodes; finally two associates were recorded below each level-3 node, giving a four-level hierarchy of 22 words. We have constructed eight of these associative trees. They are based only on intuitive judgments of pairwise associations, since published association norms are too restricted to provide guidelines for constructing such depth hierarchies. There is obviously nothing unique or special about such associative hierarchies. There are doubtless thousands that could be constructed starting with almost any content word in node-1, with variation in the number and
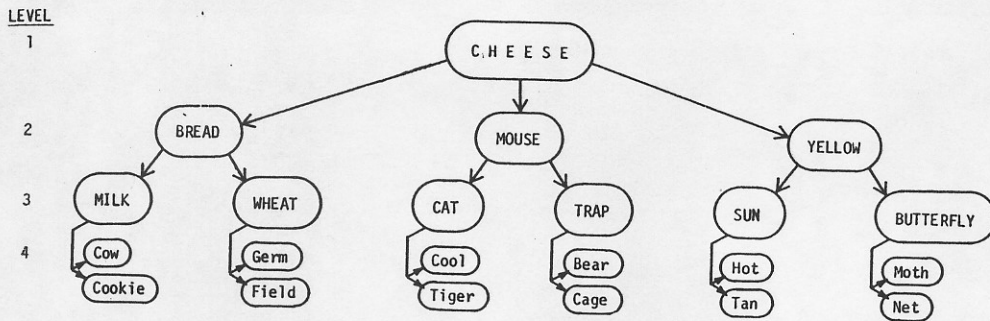
LEVEL



FIG. 2. The associative hierarchy for "cheese."

depth of branchings. The associative base for the hierarchies in Fig. 2 is not easily specified since the type of association relating node-$n$ to node-$n+1$ varies considerably. Associations between successive pairs are "intuitively sensible," but, like free-associative chains, the connections among more distant elements are sensible only by virtue of the intervening pairwise associations in the chain.

Given these associative hierarchies, one can again ask the same question as in Exp. I; namely, is free recall facilitated by presenting them in a structurally blocked fashion as opposed to a random fashion? Accordingly, a simple comparison was carried out between complete presentation of these blocked hierarchies and complete presentation of the same words mixed up randomly. The two conditions are similar to the WB and WR conditions of Exp. I.

### Method

The $S$s were 44 students in a high-school plane geometry class made available to us in a local public school system. The students were mainly tenth graders, with a few from grade 11. They were tested for 30 min on two consecutive days, having four input-output trials on different hierarchies on each of the two days. Their regular teacher cooperated as one of the $E$s and in proctoring their performance.[2]

*Design and procedure.* The $S$s received mimeographed booklets of the study materials and recall sheets. Each booklet consisted of 12 pages, four sets of three pages, consisting of two study pages with list

[2] Our thanks to Mrs. Sharon Lesgold, the teacher, for helping us with this experiment.

words followed by a blank recall sheet. Instructions were given orally by $E$.

There were four types of booklets distributed each day, each to one-fourth of the class. One type presented two hierarchies in an organized (blocked) tree form, and the second used these same words in two trees, but with total scrambling of the words across associative hierarchies and levels. The third and fourth types were the same as the first and second types, except that the order of presentation of the two hierarchies on each trial was reversed. On the first day, the *wish* and *hammer* hierarchies were used, and on the second day the *salt* and *cheese* hierarchies were used. Thus, the second day was a replication of the first, with new words, except that each $S$ received the condition opposite to what he had received on Day 1 (i.e., organized-random order or vice-versa).

The hierarchies of 22 words were printed in vertical tree form (with no circles or lines), one per study sheet. Each was studied for 45 sec as timed by a stop-watch by $E$, with instructions to turn the page to the next study sheet. After studying both hierarchies, at a signal $S$ turned to a blank recall sheet and tried to write all the words he had studied, in any order. Recall time was $3\frac{1}{2}$ min; its termination was indicated by $E$ and the next study trial began.

One $S$ was dropped because he was not available the second day, and one because she obviously misunderstood the instructions on the first day. This left 42 usable $S$s with two learning protocols each.

### Results

There were insignificant differences between recall on Days 1 and 2 within a given condition, so all protocols within a condition were pooled ignoring this Days factor. The mean words recalled (out of 44) for the Blocked condition on Trials 1–4 were 23, 34, 39, and 41; the comparable means for the Random

condition were 16, 23, 28, and 33. Mean recall in the Blocked condition significantly exceeded that of the Random condition on each trial (all $p$'s $< .01$). Each $S$ learned both a Blocked and a Random list; 40 of the 42 $S$s recalled better with their Blocked list.

Recall of the four different associative hierarchies was quite homogeneous. Percentages recalled averaged over the four trials were as follows for the Blocked (and Random) hierarchies: *Wish*, 81(61); *hammer*, 79(59); *cheese*, 77(59); and *salt*, 74(56). The differences in recallability of the four hierarchies are small and insignificant within both the Blocked and Random conditions.

An analysis of ordered recall similar to that of Exp. I (which used oral recall) was not done because of equivocation in deciding in which order many $S$s had written down the words they recalled. Some $S$s in the Blocked condition wrote their recall in the tree form they had studied; but the temporal order in which they had recalled the nodes could not be determined. It was possible to score the protocols for joint recall of the level-3 and its level-4 words. Considering Trial 1 with the Blocked list, the probability of recalling a level-4 word was .86 when its level-3 word was recalled, but only .36 when its level-3 word was not recalled. The comparable conditional recall proportions for the same words in the Random lists were .39 and .32, respectively. A similar pattern held for recall of level-3 words conditional upon recall of level-2 words. This pattern of results again points to the cuing role of the level-$n$ node in the recall of its level-$n+1$ words. This cuing effect is much more potent with the Blocked than with the Random hierarchies presumably because the $n$-to-$n+1$ association is immediately available for rehearsal in the Blocked list, but is buried among many competing responses in the Random list.

While the recall differences between the Blocked and Random conditions are large here, they are not as dramatic as the differences obtained with the conceptual hierarchies in Exp. I. There were, of course, many procedural differences between the two experiments— different $S$-populations, different exposure and recall methods, etc. Ignoring these procedural variations, however, it appears that the blocked associative hierarchies are more difficult than the blocked conceptual hierarchies, whereas the two random conditions (conceptual vs. associative words) would appear to be about comparable in difficulty. A direct experimental comparison of the two types of hierarchies is being done to check on this ordering.

The main conclusion to be gathered from the results of this experiment is that blocking of associative hierarchies considerably improves recall over a condition of random input of the same words. What is apparent to Blocked $S$s but not to Random $S$s is that many word pairs are associated and moreover that many lengthy associative chains exist in the word lists. The "structural principle" for the Blocked lists is that of *recursive rewriting* according to associative transitions, going from each level-$n$ node to its several level-$n+1$ nodes. Subjects notice and comment upon this construction principle, and they obviously use it for reconstructing the list from memory. Again however, the recursive principle is not allowed to run unchecked, since it alone would produce many associative intrusions of nonlist words. Rather the candidates produced by the associative recursion must be checked for list-membership information before they are overtly recalled. The few intrusions that do occur for the Blocked $S$s, however, appear predominantly to be "false recognitions" of candidates generated by the associative recursion principle.

## EXPERIMENT V

A possible objection to our experiments so far might be that they merely show that Blocked $S$s have learned the principle of list construction, and that knowledge of this principle is sufficient for them to produce

most of the words in the hierarchies without having any further specific information about the input list. For example, Ss told to write the numbers from 1 to 112 would produce more than Ss shown these same numbers on many slides in a randomly scrambled order and told to recall. The difference in "recall" would be utterly trivial in this instance because the list-construction principle suffices completely to generate all list items and no nonlist items. Are our Blocked vs. Random comparisons in the prior experiments simply showing this same trivial outcome in a slightly veiled disguise?

We think not. The difference between the numbers case and our hierarchies is in the range of acceptable alternatives that could be generated by the list-construction principle in the two cases. In the former case, the principle by itself generates all list items and no others; in the latter case, the principle generates only a few list items and very many nonlist items (intrusions?). We shall return later to a discussion of the psychological distinction between these two cases; but for our immediate purpose it is best to collect some evidence for our claim that the word-hierarchies we have used here are quite open-ended and non-exhaustive.

Experiment V essentially collected association norms on these hierarchies from a group of Ss. The S was given a general description of the list-construction principle and then tried to generate the word hierarchies. His generation was always set back on the "right track" when he erred. Given the complete tree up through the level-$n$ nodes, S was to generate the appropriate number of level-$n+1$ words under each node. After this, the "right" level-$n+1$ words were added to the tree, and S was asked to produce the appropriate number of level-$n+2$ words. Both the conceptual and the associative hierarchies of Exps. I and IV were used. The question was simply, how well will Ss do in generating such trees knowing only the principle by which the trees were constructed?

## Method

The Ss, ten Stanford undergraduate volunteers enrolled in Summer classes, were run individually. After general instruction illustrated with one example, S began work with the four associative hierarchies from Exp. IV, and then did the eight conceptual hierarchies from Exp. I. The order of working on the hierarchies within each set was randomized over Ss. All work was completed on one hierarchy before S went on to the next hierarchy. Work on a hierarchy began by giving S a slip of paper with the level-1 word printed in a box, and with 2 (or 3) lines radiating down to two (or three) empty boxes at level-2. The S wrote in what he thought were the appropriate words. He was then shown a second slip of paper with the "right" level-1 and level-2 words, with lines radiating down to the appropriate number of level-3 empty boxes which he was to fill. After filling these, he was given a paper with the level 1–3 words and filled in the appropriate number of boxes for level-4 words. The recursive "resetting" of S back on the "right" track gave him feedback on how well he was doing. Occasionally S could not think of enough appropriate words to fill the indicated number of empty boxes, and he was permitted to leave such spaces blank. Testing on the complete series required 30–45 min depending on S's speed of generating appropriate words. Because of an experimental error, responses of one S were lost on level-2 responses to one conceptual hierarchy and level-3 responses to another conceptual hierarchy.

*Scoring.* For notation, let $k$ denote the total number of responses required of S at a given level of the tree; $k$ is the product of the number of nodes at that level times the number of responses required at each node. With ten Ss, we shall have a distribution of $10k$ responses (minus a few omissions) at each level of each hierarchy. A variety of indices could be computed from such associative distributions. We shall report two. The first measure will be the proportion of the $10k$ responses which match the "correct" words at that level of the hierarchy. This will be relevant for interpreting S's ability to generate the particular hierarchies which E had constructed.

The second measure is one of dispersion, or conversely, how much Ss agreed on their responses at a given node. A convenient measure of agreement may be obtained as follows: the responses at a given level of a tree are rank ordered according to their frequency (with omissions considered as unique responses), and then we see how far down in the rank ordering one has to go to accumulate 50% of the response distribution. The more agreement of Ss' responses, the lower will be the rank of the median response. The rank of the median will depend on $k$, unfortunately, so this score will not be comparable across levels unless it is trans-

formed to a common scale. This is achieved in the Agreement Index (*AI*),

$$AI = \frac{\dfrac{Nk}{2} - x}{\dfrac{Nk}{2} - \dfrac{k}{2}} = \frac{N}{N-1} - \frac{2x}{k(N-1)},$$

where $N$ is the number of $S$s, $k$ is the total responses per $S$ at a given level, and $x$ is the obtained rank of the median word at that level. The *AI* index has the following rationale: (*a*) If all $Nk$ responses differ (i.e., no agreement), then the median rank, $x$, will be $Nk/2$ and *AI* will equal 0; and (*b*) if all $N$ $S$s agree in giving the same $k$ responses, then the median rank, $x$, will be $k/2$, and *AI* will equal 1. Therefore, *AI* scales from 1 for complete agreement down to 0 for complete disagreement.

### Results

The values of these indices for the four associative hierarchies of Exp. IV and for the six conceptual hierarchies of Exp. I are shown in Table 7. The total proportion correct for the overall hierarchy is obtained by a weighted average of the proportions correct at each level, with weighting coefficients equalling the proportions of words in the hierarchy contributed by the words at a given level. The column labeled NR reports the total number of different responses (types) given for the entire hierarchy by the 10 $S$s, counting omissions as unique responses (i.e., considering them to have been filled in by random selection from a dictionary). Excluding the level-1 name of the hierarchy which $S$ was given at the start, the target associative hierarchies had 21 different words, and the target conceptual hierarchies had from 25 to 30 different words, averaging 27. The NR measure thus gives a rough index of associative variety relative to the target hierarchies. The final column, labeled Recall, gives the average recallability of these hierarchies when they were presented in blocked form, pooling all four trials from Exp. IV for the associative hierarchies and the

TABLE 7

MEAN PROPORTION CORRECT ($p$), AGREEMENT INDEX (AI), TOTAL NUMBER OF
DIFFERENT RESPONSES (NR), AND RECALL FOR THE 12 HIERARCHIES

| Associative hierarchy | Level 2 | | Level 3 | | Level 4 | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | $p$ | AI | $p$ | AI | $p$ | AI | $p$ | NR | Recall |
| Salt | .33 | .74 | .03 | .61 | .27 | .83 | .21 | 115 | .74 |
| Cheese | .10 | .44 | .20 | .78 | .14 | .58 | .18 | 134 | .77 |
| Hammer | .43 | .89 | .20 | .78 | .09 | .64 | .18 | 126 | .79 |
| Wish | .37 | .81 | .20 | .80 | .14 | .72 | .21 | 123 | .81 |
| Average | .31 | .72 | .16 | .74 | .16 | .69 | .19 | 124 | .78 |

| Conceptual hierarchy | $p$ | AI | $p$ | AI | $p$ | AI | $p$ | NR | Recall |
|---|---|---|---|---|---|---|---|---|---|
| Instruments | .30 | .67 | .23 | .76 | .42 | .93 | .37 | 145 | .71 |
| Minerals | 0 | .56 | .02 | .64 | .51 | .95 | .39 | 120 | .74 |
| Body | .23 | .48 | .55 | .96 | .48 | .94 | .47 | 103 | .81 |
| Plant | .06 | .50 | .07 | .63 | .29 | .76 | .24 | 160 | .77 |
| Animal | 0 | .48 | .22 | .72 | .27 | .90 | .24 | 147 | .72 |
| Transportation | .70 | .96 | .13 | .81 | .21 | .86 | .27 | 117 | .72 |
| Occupation | .20 | .63 | .16 | .78 | .34 | .90 | .30 | 140 | .64 |
| Clothing | .20 | .39 | .23 | .85 | .36 | .96 | .35 | 95 | .71 |
| Average | .21 | .58 | .20 | .77 | .36 | .90 | .33 | 128 | .73 |

first two trials pooling Exps. I and II for the conceptual hierarchies.

The scores in Table 7 vary widely and fall into few discernible patterns. For the conceptual hierarchies, the proportion correct and the agreement indices increase at deeper levels of the tree. For the associative hierarchies, the average proportion correct decreases somewhat at deeper levels of the tree, but the average agreement index remains constant across levels. These patterns of changes in *AI* for the conceptual vs. associative hierarchies are understandable in terms of the increasing constraints upon responses imposed by the greater contextual information at the deeper levels of the conceptual but not the associative hierarchies. The overall proportion correct is higher for the conceptual hierarchies, but this is due to the preponderance of level-4 words which are more constrained and better guessed in the conceptual hierarchies.

The NR or associative variety measure reveals the open-ended or nonexhaustive character of all these hierarchies. The number of different responses given by ten *S*s is consistently 4–7 times more than the number of words in the hierarchies, with the proportions in excess of the target words being somewhat higher for the associative than for the conceptual hierarchies. Again, this seems attributable in part to the greater constraints in the conceptual hierarchies imposed by the restricted inclusion or part–whole relation between successive nodes.

Finally, we examine the relation between recallability of the hierarchies and some overall index of naive *S*s' ability to generate the different hierarchies. The mean recall probabilities for the blocked hierarchies are shown in the final column of Table 7. For the associative hierarchies, there is only very small variation in recallability and in overall generation probability for the four hierarchies. Because the range exhibited in both variables is small relative to the probable error variance of these measures, no systematic correlation of the two variables could possibly be shown in this situation. A similar statement applies for relating generation probability of the associative hierarchies to their recallability from a random list; the range of variation in recall from the randomly presented lists (Exp. IV) was only 5%, which is too small to consider correlating that variable with generation probability.

Turning to the conceptual hierarchies, the situation there is somewhat better for investigating the relation between generation probability and recall, since there was larger variation in both factors. However, the covariation in generation probability and recallability is not at all strong in this case. The *body* hierarchy was highest on both variables, but there were several large discrepancies in rankings of other hierarchies on the two variables. If the eight hierarchies are rank ordered on both variables, the correlation of the rank orders is only +.14, indicating a very weak relationship between the two variables. The Pearson correlation coefficient between the two variables is +.40, but this is well below the value of .62 required to reject the null hypothesis of zero correlation at the 5% level. The generation probabilities of the conceptual hierarchies were also correlated with the recallability of these words when they were presented in random order (condition WR in Exp. I and II). Here, too, the correlations were low and insignificant: Rank-order $\rho = +.24$; Pearson $r = +.40$ but seven of the eight hierarchies differed by only 3% in average recallability. These results show that recallability of the individual hierarchies is not predictable from the probabilities with which naive *S*s can generate the target hierarchies. In a similar vein, recallability of the conceptual hierarchies correlates very poorly ($\rho = +.15$) with the NR index of associative variety. Possibly other indices of hierarchic integration or association could be composed which will correlate better with recallability of these 12 hierarchies; but the clear failings of these obvious variables prognosticate little success for such a search.

One might entertain the hypothesis that the observed differences in free recall between the Blocked vs. Random input conditions in the prior experiments simply reflects the differential guessing of the words when $S$ does or does not know the principle of hierarchic construction. For example, in Exp. IV with the associative hierarchies, the Blocked vs. Random difference in recall, averaged over the four trials, was .19, which is exactly the average generation probability for these hierarchies in Table 7. However, there are several deficiencies in this account of the Blocked vs. Random differences in recall. First, with the conceptual hierarchies, the average generation probability (.33) considerably underestimates the recall advantage of the Blocked list on Trials 1 and 2 (average difference ca. .48). Second, it is not true that Random $S$s would never guess words on these hierarchies, since the general categorical nature of the word sets was still fairly obvious to $S$s receiving the random presentations, as shown by the categorical nature of the intrusions by these $S$s; so this factor would tend to reduce the guessing differential below the .19 or .33 figures of Table 7. Third, if $S$s were really guessing words from the conceptual hierarchies, then one should find a very large number of categorical intrusions, since the NR measure in Table 7 shows that these hierarchies are very open-ended. In fact, however, recall intrusions never exceeded 1% in either the Blocked or Random condition on any trial.

The upshot of these considerations is that the differential recall produced by the Blocked vs. Random presentations is not consistent with a differential guessing probability exhibited by $S$s who do vs. $S$s who do not know the principle of hierarchic construction. Nor is the recallability of a given hierarchy, with either Blocked or Random presentation, related in any significant degree to the average probability with which a naive $S$ can generate the words in that hierarchy from the construction principle. We are therefore left with the theoretical reconstruction given earlier for explaining the effect of Blocked vs. Random presentation; namely, the Blocked spatial array enables $S$ to directly strengthen particular category-instance associations and it also provides him with a systematic (hierarchic) plan for cuing these candidates for recall, which candidates are recalled only if they pass a recognition test for list membership.

## DISCUSSION

The message of these studies is simple: If $S$ can discover or learn a simple rule or principle which characterizes the items on a list and which relates them to one another, then he uses that rule as a retrieval plan in reconstructing the items from memory, with a consequent improvement in his performance. The principle characterizing a hierarchy is that of recursive rewriting by associative transitions. For our conceptual hierarchies, the associative transitions are primarily of one restricted type, namely, the relationship of class inclusion or part-whole. But Exp. IV showed facilitation even when the base-type of associative transition varied widely over the set of words. The potent effect of such retrieval plans on free recall was illustrated in Exp. I, where $S$s were recalling 112 words perfectly after two or three input trials, which may be compared with any number of other reports in the literature showing much poorer recall of shorter, less organized lists.

A question to which we must return is whether this result is trivial, whether it is analogous in some obscure way to asking $S$ to "recall" the numbers from 1 to 112. To answer this objection adequately would require a fairly detailed analysis of the operations involved in "memory experiments" (free recall in particular), the involvement of response-generation rules in such experiments, and the difference between "remembering" items from a presented set versus generating those items given only knowledge of a rule characterizing the set. These are difficult questions, but we would nominate them as

tremely important questions to answer if one ants to know what is involved in the remembering" seen in free recall.

The observable behavioral difference between remembering a set of items and generating a list according to a principle is in the nature and extent of intrusion errors. Of course, whether a particular candidate generated by the rule is or is not an intrusion depends on how well the rule generates only those items on the list and no others. To the extent that the rule fails to discriminate well between list and nonlist candidates, to that extent specific item-memory has to be used along with the rule to make it an adequate recall device. And it has been conjectured here that this item-memory is essentially "occurrence" recognition, mediated by some kind of recency or contextual frequency tag stored along with a lexical item in S's long-term semantic memory. The evidence collected in Exp. V tends to discredit the view that Blocked Ss are "merely guessing" from their long-term semantic memory in a manner consistent with their knowledge of the list-construction principle.

It is of interest to compare the present experiments to those of Whitman and Garner (1962) and Miller (1958) which also involved free recall of rule-generated lists. In Miller's experiment, Ss saw many strings of 3–5 letters and then tried to recall them. For one list, the strings were a subset of those produced by a simple Markov generator, which is a set of left-to-right rewriting rules for a finite vocabulary. For example, the three rewrite rules $A \to B$, $B \to (C \text{ or } B)$, and $C \to (A \text{ or } C)$, will generate strings like AB, BBC, CCAB, CCABBC but not CB, BA, or BAC. Miller found that a list of such rule-generated strings was more easily learned and recalled than was a list of random strings of comparable length with the same letters. If the list subset had had a simple structure (e.g., all permissible strings of length 4 beginning with C), then knowledge of the rewrite rules would have permitted S to "recall" all the presented strings. But in

Miller's experiment, the list subset was not especially restricted (different initial letters and strings of differing lengths) and S did not know the principle of list construction (the rewrite rules) at the start of the task. However, there can be little doubt that the recall improvement observed with practice was in part due to the S learning some or all of the rewrite rules exemplified by the list strings (see also Smith, 1966).

The experiments by Whitman and Garner (1962) (also Garner and Whitman, 1965) are relevant because their Ss in principle knew how to generate all possible items of the population from which a subset was to be recalled. A typical task might expose S to eight of the possible 16 geometric figures that could be composed by combining four binary attributes (e.g., one or two, large or small, red or blue, circles or squares). The principle for constructing the population of patterns was simple enough that most Ss could probably generate all 16 patterns in the population after only a few exposures. The problem for the free-recall S in such experiments is not one involving an inability to generate items on the presented list, but rather one of suppressing intrusions, of inferring a restricting rule which will generate only those patterns presented and no others of the population. Whitman and Garner showed that the difficulty of free recall in this situation depended upon the complexity of the restrictive rule characterizing the subset of presented patterns. For instance, if the presented subset were to occupy an elementary partition of the population (e.g., the eight red figures), then free recall would be extremely easy. Whitman and Garner also point out the basic similarity between such recall experiments and concept learning experiments, if one identifies the presented subset for free recall with the list of "positive instances" of a concept experiment. A rule for generating free recall of the presented list would have as an essential ingredient the concept characterizing the members of that subset of the total population.

The relevance of these studies by Miller and Whitman and Garner to our own experiments is relatively direct. In each experiment, the presented items were characterized by some structural principle which could be used to generate these items, and learning was largely a matter of discovering the rules and then recognizing the presented from nonpresented candidates that could be generated by the rules applied to the base vocabulary. In each experiment, of course, the difficulty of recall depended upon the structural characteristics of the whole list and was not some simple combination of the difficulty of the individual items considered in isolation. One might say that the availability or recallability of a given item depended upon "wholist" properties of the list in which it was embedded.

According to the viewpoint espoused here, free recall is mediated by S using a retrieval plan for cuing or generating plausible candidates to recall, and an "executive editor" which checks these candidates for recency recognition before overtly recalling them. Dale (1967), Kintsch (1968) and others have proposed similar views. The last few input items in the echo box of short-term memory may be recalled directly without mediated cuing, but the hypothesis supposes that most of the other items in free recall would be generated by a cuing system. The cuing system could be a set of rules, some structural information about the composition of the list, an alphabetic scheme (Earhard, 1967), or a pegword system (Wood, 1967a). Last but not least the effective cue for recall of a given word may be prior recall of other words. But even this has to regress eventually back to an implicit cuing system. The improvement in free recall over multiple trials would result, on this view, from several factors: (*a*) more discriminating occurrence-information being stored for list words, (*b*) increasing integration of subjective clusters or groups of words which S treats as units, and (*c*) development of a more adequate retrieval plan for S to cue recall of his subjective units. The characteristic of the cuing plan is that it should be a familiar or easily learned, summary abbreviation suggesting or leading to the words on the list. The word "plan" need not suggest any elaborate mentalistic or cognitive reconstruction of behavior: Millensen (1967) and Suppes (1968) have shown how hierarchical TOTE units (cf. Miller, Galanter, and Pribram, 1960) can be analyzed in terms of conditional implicit stimulus–response connections, so this language is noncommittal with respect to S–R analyses of recall.

The evidence to date, collected when S is given a systematic retrieval plan or mnemonic, tells us that such plans are sufficient to produce very high levels of recall. The evidence does not yet prove that such a plan is necessary or required for S to produce high levels of recall, but this hypothesis is beginning to look increasingly attractive.

## REFERENCES

ANISFIELD, M., AND KNAPP, M. Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology*, 1968, **77**, 171–179.

BOWER, G. H. A multicomponent theory of the memory trace. In K. W. Spence and J. T. Spence (Eds.) *The psychology of learning and motivation.* Vol. 1. New York: Academic Press, 1967.

COFER, C. N., BRUCE, D. R., AND REICHER, G. M. Clustering in free recall as a function of certain methodological variables. *Journal of Experimental Psychology*, 1966, **71**, 858–866.

COHEN, B. H. Some-or-none characteristics of coding. *Journal of Verbal Learning & Verbal Behavior*, 1966, **5**, 182–187.

COHEN, B. H., AND BOUSFIELD, W. A. The effects of a dual-level stimulus-word list on the occurrence of clustering in recall. *Journal of General Psychology*, 1956, **55**, 51–58.

DALE, H. C. A. Familiarity and free recall. *Quarterly Journal of Experimental Psychology*, 1967, **19**, 103–108.

EARHARD, M. Cued recall and free recall as a function of the number of items per cue. *Journal of Verbal Learning & Verbal Behavior*, 1967, **6**, 257–263.

GARNER, W. T., AND WHITMAN, J. R. Form and amount of internal structure as factors in free-recall learning of nonsense words. *Journal of Verbal Learning & Verbal Behavior*, 1965, **4**, 257–266.

KINTSCH, W. Recognition and free recall of organized lists. *Journal of Experimental Psychology*, 1968, **78**, 481–487.

MANDLER, G. Organization and memory. In K. W. Spence and J. T. Spence (Eds.) *The psychology of learning and motivation*, Vol. 1. New York: Academic Press, 1967.

MANDLER, G. Association and organization: Facts, fancies, and theories. In T. R. Dixon and D. L. Horton (Eds.) *Verbal behavior and general behavior theory*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1968.

MILLENSON, J. R. An isomorphism between stimulus-response notation and information-processing flow diagrams. *Psychological Record*, 1967, **17**, 305–319.

MILLER, G. A. The free recall of redundant strings of letters. *Journal of Experimental Psychology*, 1958, **56**, 485–491.

MILLER, G. A., GALANTER, E., AND PRIBRAM, K. *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston, Inc., 1960.

MILLER, G. A., AND SELFRIDGE, J. Verbal context and the recall of meaningful material. *American Journal of Psychology*, 1950, **63**, 176–185.

PAIVIO, A., YUILLE, J. C., AND MADIGAN, S. A. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement*, 1968, **76**, No. 1, Part 2, 1–25.

PARKS, T. E. Signal detectability theory of recognition memory performance. *Psychological Review*, 1966, **73**, 44–58.

POSTMAN, L. The present status of interference theory. In C. N. Cofer (Ed.) *Verbal learning and verbal behavior*. New York: McGraw-Hill, 1961.

POSTMAN, L., AND KEPPEL, G. Retroactive inhibition in free recall. *Journal of Experimental Psychology*, 1967, **74**, 203–211.

SMITH, K. H. Grammatical intrusions in the free recall of structural letter pairs. *Journal of Verbal Learning & Verbal Behavior*, 1966, **5**, 447–454.

SUPPES, P. Stimulus-response theory of finite automata. Technical Report 133, Psychology Series, Institute for Mathematical Studies in the Social Sciences, Stanford University, Stanford, Cal., 1968.

TULVING, E., AND PEARLSTONE, Z. Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior*, 1966, **5**, 381–391.

TULVING, E., AND THORNTON, G. B. Interaction between proaction and retroaction in short-term retention. *Canadian Journal of Psychology*, 1959, **13**, 255–265.

UNDERWOOD, B. J. False recognition by implicit verbal responses. *Journal of Experimental Psychology*, 1965, **70**, 122–129.

WHITMAN, J. R., AND GARNER, W. R. Free recall learning of visual figures as a function of form of internal structure. *Journal of Experimental Psychology*, 1962, **64**, 558–564.

WOOD, G. Mnemonic systems in recall. *Journal of Educational Psychology Monograph Supplement*, 1967a, **58**, no. 6, Part 2, 1–27.

WOOD, G. Category names as cues for the recall of category instances. *Psychonomic Science*, 1967b, **9**, 323–324.

WOOD, G., AND UNDERWOOD, B. J. Implicit responses and conceptual similarity. *Journal of Verbal Learning & Verbal Behavior*, 1967, **6**, 1–10.