

# ORGANIZATION AND MEMORY<sup>1</sup>

*George Mandler*

UNIVERSITY OF CALIFORNIA, SAN DIEGO  
LA JOLLA, CALIFORNIA

I. The Concept of Organization .....	328
II. The Limits of Memory and the Unitization Hypothesis .....	330
III. Clustering: The Organization of Recall .....	333
IV. Subjective Organization .....	336
V. Free and Constrained Conceptualization .....	338
VI. The Category-Recall Function .....	340
A. General Method and Procedure .....	341
B. Materials .....	342
C. Data Analysis and Rationale .....	342
D. Experimental Data .....	345
E. Summary of the Category-Recall Relation .....	349
F. Clustering and Categorization .....	351
G. Long-Term Memory of Organized Material .....	352
H. Organization and Recall as a Function of Instructions ....	354
I. Category Size and Recall .....	357
J. Related Studies .....	359
K. Category Size and Organization .....	363
VII. The Organization of Memory .....	366
References .....	371

Organization has had a somewhat tattered reputation in the history of modern psychology. Many theorists have talked about it and others have viewed it from a distance—with either affection or alarm—but

<sup>1</sup>The initial experiment on free categorization was presented at the meetings of the Psychonomic Society at Niagara Falls, Ontario, in October, 1964, where the hypothesis of the category-recall relation was outlined. The general model and some preliminary data were discussed at a Conference on the Quantification of Meaning in January, 1965, at La Jolla, California, and at a colloquium at the Center for Cognitive Studies, Harvard University, in February, 1965. The major experimental data were presented at the meeting of the Psychonomic Society in Chicago, Illinois, in October, 1965.

The preparation of this chapter and the research reported in it were supported by Grant GB 810 from the National Science Foundation, Grant APA 64 from the National Research Council, Canada, and a travel grant from the University of California, San Diego.

most of the efforts either to develop a generally acceptable class of variables that might be called organizational or to find a single acceptable measure of organization have tended to be short-lived. This has been particularly true in the area of memory and learning. Perception has fared somewhat better, although the efforts of the Gestalt psychologists to translate perceptual concepts into other areas of psychology have often produced a feeling of failure and ennui. Too often the concept of organization has become the rallying cry for theoretical battles; as a result, its connotation frequently has become emotional rather than scientific. There is no doubt that less strictly drawn battle lines in the 1920's, 1930's, and 1940's might have produced more fruitful attention to organizational concepts on either side of the fence.

Organizational variables have assumed a new importance in human psychology, particularly in the area of human memory. The present paper will be devoted to the illustration of three general principles: First, memory and organization are not only correlated, but organization is a necessary condition for memory. Second, the organization of, and hence memory for, verbal material is hierarchical, with words organized in successively higher-order categories. Third, the storage capacity within any one category or within any level of categories is limited.

Memory as used here comprises memory for specified verbal units—the memory for lists of words; long-term memory—memory extending over periods longer than a few seconds; and memory in free recall—memory for lists of words where no restrictions are placed on the order or time in which a person recalls these words. This chapter will not deal with unintegrated verbal material (such as nonsense syllables), memory for connected passages, short-term memory, nor serial or paired-associate learning. The paradigm to be explored is the free recall experiment in which the subject is presented with a list of words and then asked to recall as many of those words as possible in any order he wishes.

## I. The Concept of Organization

In 1940 George Katona wrote a book called *Organizing and Memorizing* in which he elaborated a Gestalt position as it applied to human memory; it is in a sense an unfortunate book, for Katona's preoccupation with grand theoretical points hides many of the book's empirical contributions, and some of his specific theoretical insights were inconsistent with the temper of a period that called for the development

of grand general schemes. Hidden among vague generalities—about undefined wholes, inner necessities, and “real” understanding—and general denigrations of the role of prior experience is the assertion that organization is a requirement for successful memorization. He avoids any specific definition of organization, but he does suggest that organization involves the formation and perception of groupings and of their relations. Organization is a process that establishes or discovers such relations. Throughout the book Katona emphasizes the grouping of verbal stimuli as an important variable in memorizing; for example, in one series of experiments he shows how an otherwise “unrelated” sequence of digits is better retained in memory when the digits are regrouped according to categories or principles supplied by the *E* or *S*. The notion that memory is limited when grouping is not used and that grouping overcomes such limitations was plainly stated for anyone who cared to pay attention. But for the next decade or two, few psychologists did.

Like his associationist brethren, Katona had little use for mnemonic devices. He neither cared to investigate them nor felt that they really aided memory; they excluded “real understanding.” Katona failed to see that all organizations are mnemonic devices; the “real” ones are those chosen by the experimenter or generally agreed to be more “relevant” to the problem. There was not then nor is there now any evidence that the use of socially learned organizations is better than or different from the use of unusual or idiosyncratic ones. Katona also maintained a distinction between rote or senseless memory on one hand and organized or meaningful memory on the other. The former he attacked as “pure” memorization, partly because associationists wanted to make “pure” memorization the basis of all learning or memory. Except possibly in the sense of immediate or primary memory (cf. Waugh & Norman, 1965), it is questionable whether the distinction between rote memory and other kinds can be maintained today (cf. Underwood, 1964).

Following Wertheimer (1921), Katona considered a set of stimuli to be meaningful when the existence and quality of the parts are determined by the structure of the whole. The notion that meaning, and hence organization, can be defined by the relations among the units of the set has persisted and has found its way into modern and more serviceable formulations.

Garner (1962), for example, prefers the term “structure” to “organization,” but is obviously talking about the same problem. His very attractive definition of structure is worthy of quotation.

By structure I mean the totality of the relations between events. When we say that a picture composed of randomly located dots is meaningless, we imply that we see no relations between the dots and that, therefore, the picture has no structure. If the same total number of dots is rearranged, however, we can perceive structure and the picture becomes meaningful. . . . Meaning . . . refers to the entire set of relations, not just to the significations of each individual word. A particular word may be meaningful in the sense of signification, but the entire language becomes meaningful only if some structure is perceived in the total set of symbols. I am definitely not implying that meaning as structure is simply the sum of the significations of the individual words, but rather that the structure is itself meaningful (Garner, 1962, p. 141).

Organization and structure are clearly related to the general problem of grouping. G. E. Müller had been concerned with grouping at the turn of the century, as had Selz early in this century; and in 1932 Thorndike had wrestled with the concept of "belongingness"; two events "belong" when it is apparent to the *S* that "this goes with that." Belongingness defined the boundaries of groupings.

We can propose a general use for the term organization at this point: A set of objects or events are said to be organized when a consistent relation among the members of the set can be specified and, specifically, when membership of the objects or events in subsets (groups, concepts, categories, chunks) is stable and identifiable.

## II. The Limits of Memory and the Unitization Hypothesis

The importance of organization and grouping was made obvious to psychologists interested in information processing in two papers by Miller (1956a; 1956b). Miller started with a puzzle. Evidence from a large number of sources had suggested that there were limitations on the capacity of the human organism for processing information; limitations that were observed over a range of tasks from the absolute judgment of unidimensional variables to immediate memory. In all these cases, Miller suggested, the limiting value—the "magical" number—was  $7 \pm 2$ . Subjects usually can not distinguish more than about seven alternatives of a unidimensional variable, nor remember more than about seven items from an input list in immediate memory. Given these limitations, some mechanism must be responsible for extending human judgment and memory, since we obviously do remember more than seven items and can judge across a wider range. Miller's solution to this puzzle was, in the case of human memory, the unitization hypothesis.

The unitization hypothesis (Miller, 1956b) states, first, that the memory limit cannot be extended by simply adding more sets of seven

items. The second set of seven apparently makes us forget the first and human memory can deal only with seven items at a time. The only way to extend the amount of information is to enrich each item, that is, to increase the amount of information each item conveys. Miller refers to informationally rich units in memory storage as chunks. The input items must be recoded or reorganized into new units or chunks. Miller talks about "grouping or organizing the input sequence into units or chunks" (1956a, p. 93), specifically suggesting that "by organizing the stimulus input simultaneously into several dimensions and successively into a sequence of chunks, we manage to break . . . [the] informational bottleneck" (1956a, p. 95). In summary, organization is absolutely necessary if memory is to exceed the limit of individual items that the system can deal with at any one time. This process of organization involves recoding the input material into new and larger chunks. Memory consists of the recall of a limited number of chunks (that is, about seven) and retrieval of the contents of these chunks.

The influence of this formulation on the area of human learning and memory has been both fruitful and decisive (cf. Mandler, 1967). However, relatively few extensions of the unitization hypothesis in the specific area of human verbal memory are available. Some of these extensions, by Tulving, Cohen, and others, will be discussed later, but first some further elaborations and extensions of Miller's model are necessary.

Miller's 1956 papers suggested that if the number of chunks is limited to about seven, the chunks themselves may contain apparently unlimited informational riches. The following excerpts illustrate Miller's suggestions:

The span of immediate memory seems to be almost independent of the number of bits per chunk, at least over the range examined to date (1956a, p. 93) . . . . The process of memorization may be simply the formation of chunks, or groups of items that go together, until there are few enough chunks so that we can recall all the items (1956a, p. 95). Since the memory span is a fixed number of chunks, we can increase the number of bits of information that it contains by building larger and larger chunks containing more information than before (1956a, p. 93).

The general import is that there is no limit to the amount of information a chunk may contain. Miller's major suggestion for enriching the information in a chunk was to increase the size of the set of alternatives from which an item is chosen. A second possibility involves a hierarchical arrangement in which the number of items in a chunk is limited to  $7 \pm 2$ , just as the number of chunks is initially limited to that

number. This does not imply that memory is limited to 49 items—seven items with seven items per chunk. Rather, the seven items in a chunk may in turn be informationally rich—containing again about  $7 \pm 2$  items. This extension of Miller's unitization hypothesis will form the major theoretical argument of this paper. A hierarchical system recodes the input into chunks with a limited set of items per chunk and then goes on to the next level of organization, where the first-order chunks are recoded into "superchunks," with the same limit applying to this level, and so forth. The only limit, then, appears to be the number of levels the system can handle, a problem to be discussed later. It might be noted that whereas Miller's early formulation advances the general notion of informationally rich chunks, later formulations discuss hierarchical systems similar to the one advocated here (Miller, 1962, p. 49; Miller, Galanter, & Pribram, 1960).

In the organization of words as items, chunking proceeds primarily by way of conceptualization or categorization of sets of words. A further assumption, therefore, is that a category is equivalent to Miller's chunk.

Given that a limit constrains the number of words that can be recalled from a category and that a similar limit constrains the number of subcategories, categories, and superordinate categories that can be recalled, what is the numerical value of that limit? Miller suggested "the magical number seven, plus or minus two." As he points out however, the span for monosyllabic English words is only about five (cf. also Tulving & Patkau, 1962) and, in fact, the immediate memory experiment may hide an artifact that spuriously inflates the limiting value. Since Miller's work there has been a recurrent interest in immediate or short-term memory, dealing with memory effects within 30–60 seconds following input (cf. Melton, 1963). When a *S* is required to memorize relatively large sets of words, the mechanism apparently involves two separate processes: Short-term, or primary, memory (Waugh & Norman, 1965), which produces recall of the words immediately preceding the output; and organized memory, which typically includes earlier words from the list (cf. Waugh, 1961). Thus the number of  $7 \pm 2$  may be made up of two components:  $4 \pm 1$  plus  $3 \pm 1$ . Since the present concern is not with short-term or immediate memory in the sense of recovering items from some temporary or buffer storage, it seems likely that the value of items to be recalled per chunk is below seven; for working purposes, and in light of some of our subsequent data, a value of  $5 \pm 2$  seems more appropriate.

To recapitulate: Given a set of words, a human organism catego-

rizes them and, if the length of the list requires an extended organization, arranges the categories in turn into superordinate categories. When a category contains more than about five words it may, if necessary and possible, again be subdivided into two or more subordinate categories. Thus it is assumed that if recall from a list contains more than five items, then the *S* has used more than one category and some of his categories contain more than one item. Conversely, if a list of words is categorized into several categories such that each category contains one or more words, then recall should be a direct function of the number of categories used during organization of the list. The experimental data presented in the following discussion pertain directly to the relation between the number of categories used in organization and the number of words recalled.

In the sense of the unitization hypothesis and its elaborations presented in this section, the process of memorization is a process of organization. Katona's formulation thus is correct; memorization or learning depends on organization and the organizational variables (rather than the number of trials, for example) determine memory.

If organization determines recall, then the categories available to a *S* should also determine the form of his output. Members of categories should be recalled together if the *S* remembers categories and then their content. Extensive work has been done on this effect of organization and we will briefly review some of the experimental studies on clustering in free recall.

### III. Clustering: The Organization of Recall

A large number of studies has been concerned with the tendency for categorized items to cluster during recall. This particular line of research was initiated with an experiment by Bousfield and Sedgewick (1944), who found that *Ss* instructed to produce all the items in a particular language category (such as birds) would cluster subcategories during recall. In 1953 Bousfield initiated a program of research to investigate further the tendency of members of a category to appear contiguously during recall. In the first experiment (Bousfield, 1953) *Ss* were given a randomized list of 60 items consisting of four categories (animals, names, professions, and vegetables) with 15 items per category. Following a single presentation *Ss* listed the items they could recall and Bousfield's data showed conclusively that such recall contained clusters of the categories built into the lists.

It should be noted that in this and almost all subsequent experiments

on clustering the categories investigated were only the categories pre-established by the experimenter. Such a procedure presents two serious problems of analysis: first, the problem of the distinction between the discovery and use of these categories, and second, the necessary tendency to ignore any clustering or organization used by the *S* but different from the organization expected by the *E*.

Concerning the distinction between discovery and use, consider the following problem. Assume that I make up a list of names of all the people in my acquaintance and then categorize this list into those people who are blood relations, those who are professional acquaintances, and those who are social acquaintances. In my recall of such a list, I would probably cluster these categories according to these three characteristics. If the same list were presented to a total stranger, it is unlikely that his recall would show any significant clustering in terms of my categories. If I were to present the list to a professional colleague, however, we would find some clustering in terms of one of the categories, and if I gave the list to my wife, there would probably be significant clustering for all three categories. Similarly, any pre-categorized list will show clustering to the extent that the *S* has available the categories that the *E* has put into the list. On the other hand, the *S* might know of these categories but may not discover that the list in fact contains them.

Closely related to this difficulty is the possibility that a *S* may in fact discover the categories but choose not to use them, or he may use some combination of these categories and some of his own construction. In either case the analysis of clustering in terms of the preestablished categories will usually underestimate the actual degree of clustering imposed by the subject.

When, as is usual, the data from a group of *Ss* are averaged in a typical clustering experiment, the idiosyncratic clusters are never examined. Subjects will be included in the analysis who use their own rather than the experimenter's organization, because of some idiosyncratic preference or because they never discovered the experimenter's categories. In general, then, both final performance and clustering data will contain an inordinate amount of noise and variance.

There is one aspect of these studies that is of general interest and is the point of major emphasis of workers in this field; it concerns those variables that affect discovery and clustering. For example, Bousfield, Cohen, and Whitmarsh (1958) have shown that if the categories contain items with high taxonomic frequency, the recall and clustering values will be significantly greater than for categories with items of



low taxonomic frequency. Taxonomic frequency thus is one variable that affects the likelihood that a category will be discovered and used. Similarly it is reasonable to suppose that the discovery and use of highly overlearned categories will produce a more stable organization and therefore better recall than the *ad hoc* categories a *S* may impose on the material. In a summary of a large program of research on factors that affect the organizational characteristics of free recall, Cofer (1965) arrived at a similar conclusion.

Cofer and his associates have also compared the occurrence of *E*- and *S*-defined clusters. Their data suggest that the more "obvious" the *E*-defined categories or pairs, the less likely it is that idiosyncratic *S*-defined clusters or pairs will occur in recall; that is, the more likely it is that the *E*'s and *S*'s categories coincide. Similarly, Marshall and Cofer (1961) have shown that if word relations have "some prominence," telling the *S*s to look for such relations increases the degree of clustering of these pairs in recall.

In summary, the various studies on clustering show that clustering of *E*-defined categories will occur and that such clustering is a function of the ease with which the *S* can discover these categories. In addition, important advances have been made to define the variables that will influence the ease of such discovery and, finally, the variety of different conceptual schemes that *S*s may use to categorize lists of words is illustrated by the large variety of *E*-defined schemes that affect clustering. These include categorical, associative, syntactic, and semantic factors, and probably extend to a variety of idiosyncratic schemas that are a function of the individual *S*'s past experience and past word usage.

More immediately relevant to our present interests are those studies that have investigated the relation between the number of categories in a list and free recall from that list. The earliest of these was a study by Mathews (1954); two more recent ones were those of Dallett (1964) and Tulving and Pearlstone (1966). The results obtained in these studies will be discussed after some of our data have been presented and the category-recall relationship has been discussed in greater detail. What should be noted now is that these studies also have used *E*-defined categories. In addition, they have held list length constant while varying number of categories, and have thereby confounded list length with items per category. Although such confounding is inevitable when list length, number of categories, and number of items per category are varied, it presents some difficulty in interpretation. This point will be discussed later, but as far as *E*-defined cate-

gories are concerned, the same criticism applies to these studies as to the clustering studies. Although in the clustering studies the relation between the discovery and use of categories is a minor problem if the main point of interest is the specification of the variables that will produce organization and clustering, any attempt to specify the relation between recall and number of categories becomes dubious when it is not known whether the *S* did in fact discover and use the categories built into the list. Only Tulving and Pearlstone paid detailed attention to this problem and demonstrated the occurrence of subjective clustering of objective categories.

It will be the major import of the studies presented here to show just such a relation. The difficulty of some of the prior studies on this topic will be better illuminated in that context, but it seems that one of two conditions must be met in order to be able to demonstrate the relation between categories and recall. The first possibility is to provide the *S* with the names or labels of the categories and to show him the actual categorical structure of the list. If that is done during input only, one might be fairly certain that an input organization has taken place that is at least similar to that desired by the *E*. The *S* may, of course, still impose some of his own category system on the list. If the *S* is also provided with category labels during the output, we can further be sure that he will remember all the categories specified by the *E*. Tulving and Pearlstone (1966), for example, have fulfilled these conditions.

Another possibility, the one used here, is to permit the *S* to impose his own organization on the input. Such a method not only avoids the problems mentioned earlier but also permits us to see how organization proceeds and what the preferred or optimal organizational schemas might be.

#### IV. Subjective Organization

There have been several recent attempts to investigate the organization that *Ss* impose on input materials. Tulving (1962) and Seibel (1964) have been most directly concerned with subjective organization in free recall. Tulving's paper raises the question—derived from some of Miller's (1956a) notions—whether or not the improvement in performance in multitrial free recall is a direct function of the increase in organization. In his demonstration study *Ss* were given a list of “unrelated” words, that is, words not organized by the *E*. The order within input lists was changed in a random fashion from trial to trial.

Tulving developed a measure of the sequential dependencies in the output of successive trials. This measure, called SO for "subjective organization," evaluated the Ss tendency "to recall items in the same order on different trials in the absence of any experimentally manipulated sequential organization among items in the stimulus list." Tulving concluded "that the Ss do impose a sequential structure on their recall, that this subjective organization increases with repeated exposures and recall of the material, and that there is a positive correlation between organization and performance" (1962, p. 352). A similar conclusion was reached by Bousfield, Puff, and Cowan (1964). Both their measure and the SO measure only evaluate pairwise dependencies and can only tell us that organizational activity is in fact revealed in the output. It is not designed to evaluate the categorical organization by the S, nor can it evaluate the occurrence of organized units larger than pairs.

The last deficiency is a major argument against using either a clustering or SO measure in evaluating the organization imposed by the S. A cluster of two items that occurs in repeated trials does not imply that the category to which the cluster belongs is not in fact larger than two, and this argument applies to an output cluster of any size. A S may in fact produce one or two members of a category, use his written output as a reminder while recalling other items, and then return to the category that he had previously started. For example, a list may contain a "furniture" cluster; the S recalls "table" and "chair," then recalls some other items, checks the list, and on seeing "table" and "chair" may then give additional items from the furniture category.

Finally, the SO and clustering approaches do not tell us what the organization at the time of input was, nor how output and input organization are related to each other or to performance. Tulving (1962) has suggested the need for studying both the input and output phases of free recall. Our experiments are more directly addressed to the relation between organization during input and performance. For the reasons just listed, less emphasis will be placed on the dependent measures of organization developed by Bousfield, Tulving, and others.

Seibel (1964) has reported some initial work in which the subjective organization of the input lists was related to performance and clustering. His Ss were presented with 40 words at a constant rate and were required to write these words on a study sheet containing an array of blank cells. On this sheet Ss could—though they were not instructed to do so—organize the input according to categories of their own choosing. After the presentation of the list, Ss were required to recall

at a constant rate. Seibel found that *Ss'* recall contained clusters that corresponded to the clusters on the study sheet and that performance of the experimental group was superior to the performance of a group that had been instructed to write down the words from the input list in the order in which they were given. The performance of the latter group was indistinguishable from that of another control group whose members did not write down the input list at all. In other words, subjective organization significantly improved recall and affected subsequent clustering in recall.

With this background on prior work on organization and memory, some of our experimental studies can be considered, starting with a recapitulation of a study on free and constrained conceptualization and the relevance of free conceptualization or organization to the problem at hand. This will be followed by a series of six studies on the category-recall relationship. Finally, a brief experiment on organizing and memorizing instructions will be discussed.

### V. Free and Constrained Conceptualization

In our discussion of experiments on clustering we have suggested that *E*-imposed categories may frequently hide the *Ss'* system of organization. In addition, a procedure that focuses on such categories gives us little information about how the average human organism might go about organizing an input list. Similar arguments can be addressed to the typical experiment in concept learning where *Ss* are required to attain some concepts specified by the *E*. Since categories and concepts are, in the present sense, interchangeable notions, an examination of an experiment by Mandler and Pearlstone (1966) on this topic provides the first step in the program of research reported here.

Mandler and Pearlstone argued that the typical concept-learning experiments not only hide important aspects of conceptual behavior but also present the *S* with an interference paradigm. It is assumed that any set of stimuli will invoke some categorization or conceptualization on the part of the *S*. The initial categories imposed by the *S* and those imposed by the *E* are not likely to be identical. To the extent, then, that the *S* must suppress, extinguish, or ignore his own system of conceptualization, such activity will interfere with the acquisition of the *E*-defined conceptual categories.

Subjects were given either free or constrained concept-learning tasks with four different kinds of materials, of which the high frequency words are of particular interest. Subjects in the "free" groups were given a

deck of 52 cards, each of which had a word printed on it. They were asked to sort these cards into anywhere from two to seven categories according to any system they wished to use. They were also told that following their first sorting trial they would be given another deck with the same words in a different order and would be asked to sort the words again and to continue in this manner until they had achieved identical categorizations in two successive trials. The Ss in the "constrained" group learned the same category systems as the free Ss by yoking one constrained S with one free S. The constrained S thus was in a typical concept-learning or attainment situation, with the target concepts being those of the yoked free Ss. Following attainment of the concept-sorting criterion, Ss in both groups were asked to recall as many words as possible from the set they had just sorted.

The major relevant findings were that constrained groups needed twice as many trials as free groups to reach criterion in sorting, but that both groups recalled the same number of words, about 20 out of 52. In addition, the various free groups used a fairly stable mean number of categories in sorting, that is, from 4.0 to 4.6, regardless of the materials they were asked to sort.

Thus the assumption that constrained concept learning represented an interference paradigm was supported by the data. Furthermore, numbers of sorting trials did not affect recall; if a stable categorization had been achieved by the two groups, recall was identical.

These findings supported our notion that the free or subjective organization of verbal materials could be fruitfully investigated and that the method used in this experiment provided one approach to the investigation of the organizing behavior of human Ss. We have already mentioned Tulving's (1962) approach to a similar problem; Imai and Garner (1965) have also made the distinction between free and constrained classificatory behavior.

The major finding of the Mandler and Pearlstone study that is relevant here, and the starting point for subsequent experiments, concerned the relation between recall (R) and number of categories (NC) used during sorting. Since Ss in the free group could use any number of categories from two to seven, it was possible to relate these values to free recall performance. Figure 1 shows the observed relation between NC and R for the 10 free Ss. The correlation between these two variables is .95 and the equation for a straight line fitted by the method of least squares has a slope of 5.59 and a  $y$ -intercept of -6.0. In other words, Ss remembered on the average 5.6 words for each category and their recall was a direct function of the number of categories used. It was this reassuring suggestion that the category-

recall relation could be directly investigated that launched the following set of experiments.

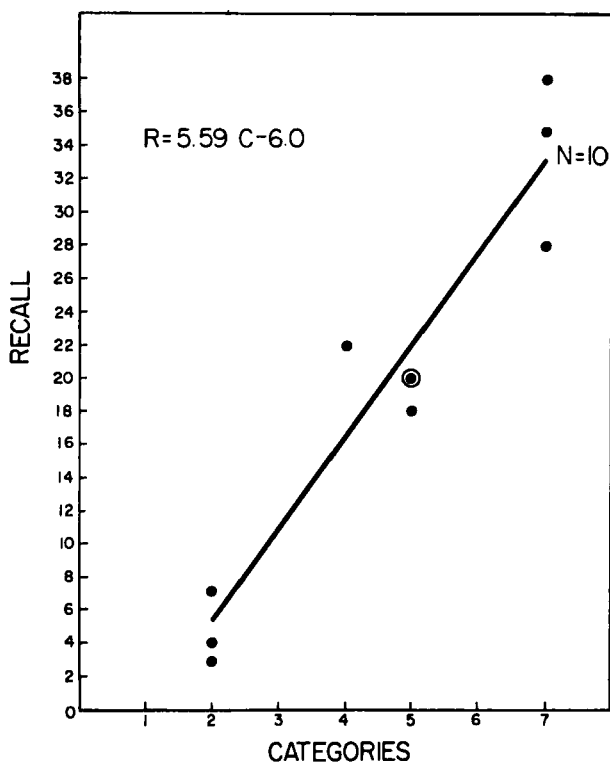


FIG. 1. Recall as a function of number of categories used in Experiment A. Data points are for individual Ss. The equation shown is for the line of best fit.

## VI. The Category-Recall Function

The following six experiments are all variations on the theme developed in the Mandler and Pearlstone study, which will be referred to as Experiment A. Free categorization was used to investigate the categorization behavior of the Ss and to establish stable and reliable category systems. Following such free categorization, free recall of the words used in sorting was tested in order to investigate further the category-recall relationship. The general method used in all six experiments will be described first, followed by the specification of the variations incorporated in each of Experiments B to G and the description of the relevant data from these experiments.

# A. GENERAL METHOD AND PROCEDURE

All *Ss* received identical basic instructions. They were given decks of cards that they had to sort in successive trials until they had achieved two identical sorts. Subjects were not allowed to put one card into one category and all others in another; apart from this restriction, any method of sorting or categorization was permitted. Following their last criterion sort they were asked to write on a sheet of paper all the words they could recall. They were given enough time to write down all the words immediately available to them. Recall terminated after a pause of about 1 minute with no recall had occurred. During sorting *Ss* were allowed to proceed at their own speed. However, if a *S* was not able to reach criterion within about 1¼ hours the session was discontinued. Column 4 of Table I shows the number of *Ss* in each group who

TABLE I  
SUMMARY OF DESIGN OF EXPERIMENTS A-G

Exp.	Initial <i>N</i>	Noncontent <i>Ss</i> <sup>a</sup>	Noncri- terion <i>Ss</i> <sup>b</sup>	<i>Ss</i> with more than 7 categories	Final <i>N</i>	Criterion (%)
A	10	2	0	—	10	100
B	70	17	10	—	43	100
C	32	7	12	6	7	100
D	40	8	17	—	15	100
E	39	11	3	5	20	100
F	49	16	8	—	25	95
G	64	19	14	3	28	95

<sup>a</sup> Over-all percentage of *Ss* who did not use content categories was 26%.

<sup>b</sup> Over-all percentage of *Ss* who failed to reach criterion was 21%.

failed to reach criterion. This number also contains some *Ss* who continuously changed numbers of categories and for whom no stable NC (number of categories) value could be determined.

Subjects were seated at a table with a sorting hamper with seven slots slightly larger than the 3- by 5-inch cards on which the words were printed. They were always able to inspect the top card of any category, but they were not allowed to inspect any other card during the sorting task. In some experiments *Ss* were allowed to use more than seven categories. In these cases the table top was subdivided with chalk lines into an array of 20 compartments. Otherwise the procedure was identical.

The *Ss* were first-year psychology students of both sexes at the University of Toronto who had not previously participated in any free recall experiments.

## B. MATERIALS

Three sets of words were used. The first set, identical with that used in the Mandler and Pearlstone study, consisted of 52 words from a fairly wide range of Thorndike-Lorge values—the *52-range* set. Frequencies ranged from 14 per million to AA. There were 50 nouns, 1 adjective, and 1 adverb, though 20 of the nouns also had verb functions and 3 also had adjective functions. Six different decks were prepared in random order, and Ss who used more than six trials were given the same decks over again.

The second set of words consisted of 52 AA nouns, 33 of which also had verb functions and 1 of which also an adjective function. The third set of words consisted of 100 AA words, the 52 from the previous set plus another 48. Of these 100 nouns, 59 also had verb functions and 1 had an adjective function. They were arranged in six decks just as the first one.

## C. DATA ANALYSIS AND RATIONALE

It has already been noted that Ss who did not reach criterion were excluded from our analysis. For Experiments B, C, D, and E, the criterion was the same as that for Experiment A, namely, two identical sorts in succession. The high degree of attrition suggested an attempt to use a less stringent criterion, which was applied in Experiments F and G, where the 100-word vocabulary was used. Criterion was reached when sorting on two successive trials differed for no more than 5 words out of the 100. In other words, a 95% sorting consistency criterion was used instead of 100%.

One other source of *S* loss must be discussed. Since Ss were told that they could use any sorting or categorization criterion they wished, some Ss used organizations based on systems other than word content. However, such organizations of words in terms of alphabetizing the initial letters or using word length or counting number of vowels is potentially or actually useless for purposes of recall. When words are sorted according to the alphabet, the *S* need only look at the first letter of the word. Since Ss were not told that they would be asked to recall the words following the categorization task, such noncontent sorting was fairly frequent. Column 3 of Table I shows the number of Ss in each experiment (except for Experiment A) who were discarded because they did not use content categories. The over-all percentage of such Ss was 26%. Since a subsequent experiment showed that additional *recall* instructions prior to categorization did not affect recall, future studies can avoid this source of attrition. If Ss are instructed to recall,



the use of alphabetical and other noncontent categories declines markedly.

Finally, for purposes of initial analysis only those *Ss* were used who restricted themselves to seven categories or less in those Experiments (C, E, and G) where *Ss* were permitted to use more than seven categories. The restriction to seven categories in most of our experiments and in all of our analyses arose out of the consideration of the major purpose of the studies. We are concerned with (a) the relation of number of categories to number of words recalled and (b) the number of words that are recalled per category. For present purposes it would have obscured some of our major findings if the task had been complicated by also including category recall. It was felt that with a maximum of seven categories and with several trials during which the *S* could become thoroughly familiar with his categories, the likelihood that a *S* would forget a category would be relatively small and that the results would not be a function of both category recall and recall within categories. This reasoning was fully borne out by our data. Using Cohen's (1963) criterion for category recall (that is, counting a category as recalled if at least one member of the category set is recalled), the incidence of failure to recall categories was extremely low. Out of a total of 680 categories appearing in our final protocols, only 13 categories, or less than 2%, did not appear in the recall data. Thus restricting both the *Ss* and the analysis to seven categories or less assures us that the data presented will present a picture of items recalled per category and will, except in rare cases, not be confounded by problems of category recall. Data on *Ss* who used more than seven categories will be presented separately.

The "Final *N*" column in Table I shows, for each of the experiments, the number of *Ss* included in the analysis in Table II. These *Ns* include only *Ss* who reached the criterion, used content criteria in their categories, and used seven categories or less—where that restriction is applicable.

Table II shows, in successive columns: (1) the final number of *Ss* (*N*); (2) the mean number of categories used (*NC*); (3) the mean number of trials needed to reach criterion (*T*); (4) the mean total recall (*R*); (5) the correlation between *NC* and *R*; (6) the correlation between *T* and *R*; (7) the partial correlation between *NC* and *R*, holding *T* constant; (8) the slope of the line of best fit for the *NC-R* function; (9) the intercept of that line; (10) the mean ratio of repetition (*RR*) developed by Bousfield to measure clustering (Cohen, Sakoda, and Bousfield, 1954). It is defined as  $R/(N-1)$  where *R* is the number of times a word from a category follows another word

TABLE II  
SUMMARY DATA FOR EXPERIMENTS A-G

Exp.	N	Mean cat. (NC)	Mean trials (T)	Mean recall (R)	Correlations			Slope	Intercept	RR <sup>a</sup>	Vocabulary
					NC-R	T-R	NC-R (T constant)				
A	10	4.6	4.6	19.5	.95 <sup>b</sup>	.45	.94 <sup>b</sup>	5.6	-6.0	.63	52 Range
B	43	4.4	6.9	23.4	.74 <sup>b</sup>	.16	.75 <sup>b</sup>	2.9	10.6	.56	52 Range
C	7	4.6	5.4	23.6	.39	.30	.30	1.7	15.7	.56	52 AA
D	15	5.5	6.1	35.3	.70 <sup>b</sup>	.24	.73 <sup>b</sup>	2.5	21.2	.72	52 AA
E	20	5.0	6.2	28.2	.60 <sup>b</sup>	-.22	.61 <sup>b</sup>	3.9	8.7	.70	52 Range
F	25	4.2	6.2	38.8	.84 <sup>b</sup>	-.02	.86 <sup>b</sup>	7.5	7.5	.68	100 AA
G	28	3.7	6.3	40.7	.64 <sup>b</sup>	.10	.64 <sup>b</sup>	7.2	14.4	.68	100 AA
Median		4.6	6.2		.70+	.16	.73+	3.9	10.6	.68	

<sup>a</sup> Mean ratio of repetition.

<sup>b</sup>  $p < .01$ .

from that category and  $N$  is the total number of words recalled; (11) the vocabulary used in the experiment.

It should be noted that the line of best fit for the NC-R function was obtained from the raw data of all final  $S$ s, even though Figs. 2-4 only show the mean recall for each NC value.

#### D. EXPERIMENTAL DATA

##### 1. Experiment B

In order to assure adequate sampling for all values of NC, it was decided to impose one additional constraint on the  $S$ s, namely, to instruct them on the number of categories they were to use. The additional instructions for each of the six NC groups who were told to use from 2 to 7 categories informed  $S$ s that they must use a certain number of categories but that they were free to use any category system within that limitation. Data were collected for 10  $S$ s in each NC group who reached the sorting criterion. The NC-R relation is shown in Fig. 2

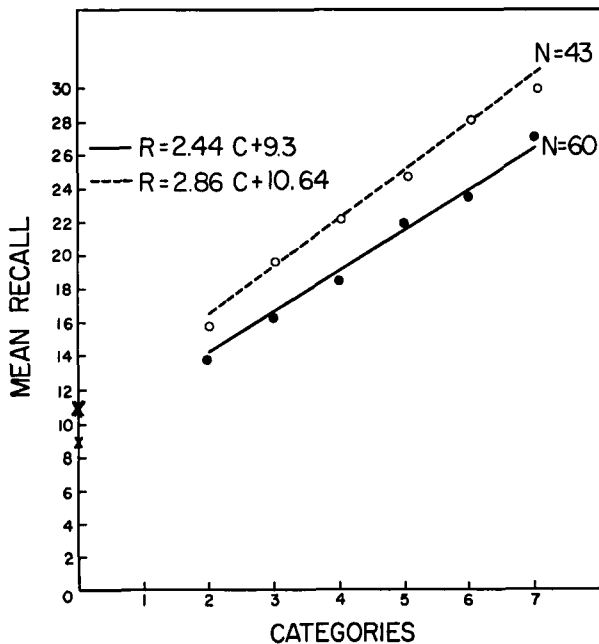


FIG. 2. Mean recall as a function of number of categories used in Experiment B. Filled circles and solid line are for original 60  $S$ s; open circles and dashed line, for the 43  $S$ s using content categories only. The equations shown are for the line of best fit.

for these 60 Ss as well as for 43 Ss who only used content categories. The data in Table II are for the latter group. The 17 Ss who used noncontent categories were rectangularly distributed across the NC groups; two came from the 2-NC group and three from each of the others.

The slope of the function is reduced from that obtained in Experiment A. But the straight line fit is obviously excellent. Subjects recall about 2.9 words per category and the  $y$ -intercept is positive. It might be noted that comparisons with Experiment A are probably spurious. The  $N$  in that experiment is small, the data include points from Ss who alphabetized, and the negative  $y$ -intercept is difficult to interpret. The general significance of the  $y$ -intercept will be discussed at the conclusion of the experimental discussion.

Table II shows that the Ss used a mean of 4.5 trials to reach criterion. The mean number of categories used is, of course, fixed in this experiment. There is highly significant correlation of .74 between NC and R and the T and R correlation is nonsignificant. The RR is .56.

## 2. *Experiments C and D*

It was thought that the constraint imposed on the Ss in Experiment B might have depressed the slope value of the NC-R function and Experiments C and D were conducted with a new, high frequency vocabulary. In Experiment C, Ss were allowed up to 20 categories; in Experiment D the conditions were similar to Experiment A. Unfortunately the attrition due to noncriterion Ss was heaviest in these two experiments, reaching a value of 40%.

In Experiment C, Ss chose more than seven categories more frequently than in the similar unlimited-category Experiments E and G. As a result the final  $N$  for Experiment C was only 7. In Experiment D a large number of Ss continued to switch numbers of categories and the final  $N$  was only 15.

For Experiment C the slope was 1.7 and the intercept 15.7. The correlation between NC and R was .39 and between 'T' and R, .30.

Experiment D produced another low slope value of 2.5 with a very high intercept of 21.2. The NC-R correlation was .70; the T-R correlation was .24.

In general these two experiments produced the most disappointing results of the series. In trying to remove the constraint imposed in Experiment B, we not only reduced the NC requirement but also provided Ss with a combination of the shortest list of the series (52 words) and the most familiar vocabulary (AA). The possibility exists

that these lists were so easily organized that Ss might have used the sorting task to impose organization over and above the obvious associations and clusters apparent in the list. In such a case, the organization produced by the categorization task is not in fact the organization that determines recall, and recall might have been a function of both the overt and other covert organizing schemes.

It should be noted that Experiment D was essentially a replication of Experiment A but produced a less stable, though still highly significant, relationship between NC and R.

### 3. Experiment E

On the assumption that the 52 AA list was too easy for the task in Experiment C, the original 52-range list was used in Experiment E in which Ss were allowed unlimited number of categories. Table I shows that only 5 Ss out of 36 criterion Ss used more than seven categories. The slope of 3.9 for the criterion content Ss is within the expected limits and the intercept is 8.7. The NC and R correlation is .60, and the recall and trials correlation is negative and low at  $-.22$ .

Figure 3 shows the NC-R relation for Experiment E. The fitted line

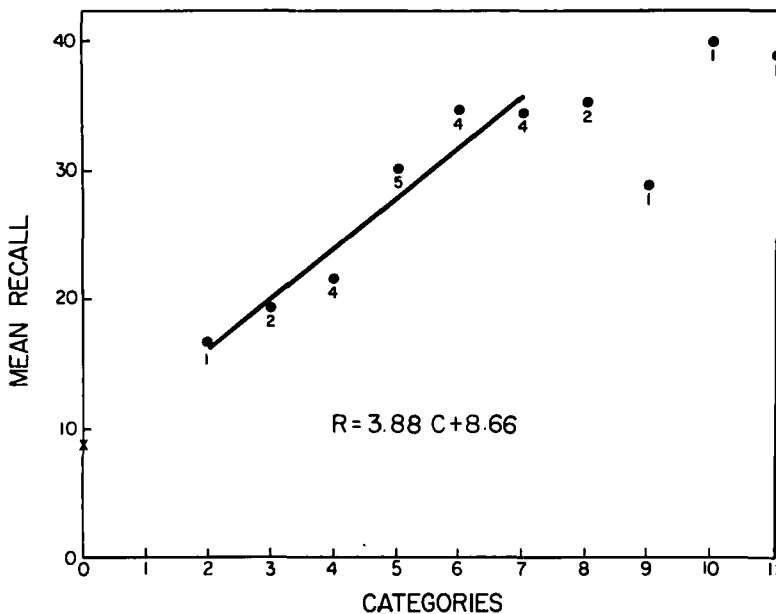


FIG. 3. Mean recall as a function of number of categories used in Experiment E. Numbers next to data points indicate number of Ss contributing to the mean value shown. The equation is for line of best fit for all Ss using seven categories or less.

is for the *Ss* with seven categories or less only. The data for the other five *Ss* are also shown. It is quite clear from inspection of this figure that the category-recall function does not hold beyond seven categories. In fact, the line of best fit for the five *Ss* who used more than seven NC has a slope of 1.65 with a high intercept of 20.6. In the previous unlimited category experiment (C), the slope for the comparable six *Ss* was 0.88, the intercept was 29.7.

#### 4. Experiments F and G

If the 52 AA vocabulary was too easy, in the sense that the list was relatively short considering the high familiarity of the items, another way to avoid parallel organizations, that is, those that the sorting task might not detect, is to increase the length of the list. Experiments F and G therefore used the 100 AA list, with F restricting *Ss* to seven categories, while G again permitted *Ss* to use unlimited NC. At the same time, the new criterion of two successive trials with only 95% overlap was introduced. The results are shown in Fig. 4 for the NC-R relation.

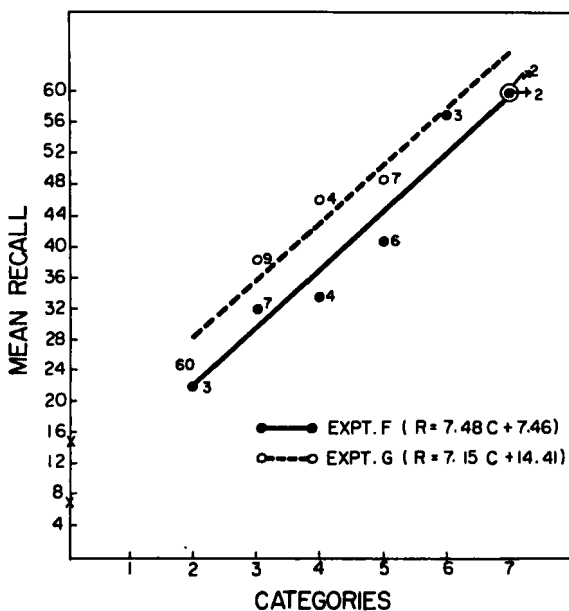


FIG. 4. Mean recall as a function of number of categories used in Experiments F and G. Number next to data points indicate number of *Ss* contributing to the mean value shown. Equations shown are for the lines of best fit.

The data for Experiment F in Table II are in the expected direction. The slope of the NC-R relation is 7.5 with an intercept of 7.5 and all other values are within the ranges that by now had become fairly stable.

Experiment G produced only 3 Ss who used more than seven categories; the remaining 28 content criterion Ss produced results very similar to Experiment F, with a slope of 7.2, an intercept of 14.4, and other comparable values.

In order to check on the acceptability of the new criterion, 10 of the 25 Ss in Experiment F who had reached the old 100% criterion provided data for a separate analysis. These Ss used a mean of 4.1 categories, 5.8 trials, and recalled 38.7 words. Their slope value was 8.0, the intercept was 6.3. Thus, the data can be said to be quite comparable whether the 100% or 95% criterion is used.

#### E. SUMMARY OF THE CATEGORY-RECALL RELATION

With Experiments F and G the present program of research terminated, since the relation between NC and R seems to have been fairly stably established under a variety of different conditions.

The consistency of some of the dependent variables is quite remarkable. The last row of Table II shows the median of the major variables of interest. The value of 4.6 categories for median NC is consistent with the Mandler and Pearlstone (1966) data that Ss will, on the average, select about 4-5 categories out of 7. While the range for all Ss is determined at 2-7, the preference for about  $5 \pm 2$  categories during conceptualization seems to be fairly well established. In all the unlimited-category experiments only 20% of the Ss used more than seven categories. The same general level applies if Ss are examined regardless of content or criterion restrictions.

A surprising finding was the general stability of the mean number of trials (T) needed to reach criterion. The median was 6.2, with a range of 4.6-6.3. While the materials used were, of course, quite homogeneous, the stability of this value is still somewhat unusual.

More important for understanding the trials variable is the relation among NC, T, and R. Table II shows that the median value for the correlation between NC and R is .70. This coefficient expresses the major thesis of this paper, that is, that there is in fact a highly significant positive relation between number of categories and recall. The significance level of the individual correlation coefficients also permits an evaluation of the stability of the straight line fit.

It might be argued that the NC-R relation hides a more basic

relation between trials and recall. Thus it would have been possible that Ss take more trials the more categories they use and that the more trials they had during which they were exposed to the input list, the better would be their recall. In that case we would expect a high NC-R correlation, mediated by the trials variable.

The column of correlation coefficients for the T-R relation suggests that such an argument is inadmissible. Not only is there a generally nonsignificant, if not negative, relation between trials and recall (median = .16), but the NC-R relation remains essentially stable when it is corrected for the mediating effects of trials. The median partial correlation between NC and R with T held constant is .73. Thus trials are an unimportant variable in the particular relations investigated here. (When total time expended during sorting is used instead of number of trials, the relation with both NC and recall is very similar to that obtained with trials.) Subjects need a certain number of trials to reach a criterion of organization, but it is the nature of that organization, not the number of trials needed to produce its stability, that determines recall. In other situations, the apparent effects of trials would, of course, be greater. Thus in a multitrial free recall experiment, organization would develop more slowly with trials, and there would be a much larger correlation between trials and recall, though we would argue that such a relation hides the basic category-recall function. Multitrial free recall will be discussed later in more detail.

The median slope value for the various experiments is 3.9, which is within the range of our hypothesized value of  $5 \pm 2$ . On the average, Ss in these experiments add about 4 words to recall for every additional category used, with a range of 1.7-7.5.

The median  $y$ -intercept value is 10.6. The theoretical meaning of the  $y$ -intercept is somewhat complicated. The straight line function suggests that when Ss use no category they would recall about ten words from the list. This is theoretically meaningless for our purposes, since we assume that categorization is essential for memory and, by definition, words cannot be classified in zero categories. If anything, the value of the intercept might be applicable to the situation where all items in the list are categorized into a single set or category, but the function does not, of course, make that prediction.

It seems reasonable to suggest that the function is in effect discontinuous between 0 and 1, primarily because of the theoretical and empirical lack of interpretation of a zero-category sort. The  $y$ -intercept can be used as an estimate of the amount of material recalled on the basis of organizations other than those assessed by the NC



variable. It is of course possible, as we have already discussed in relation to the 52 AA vocabulary, for Ss to use two or more concurrent organizational schemas. In that case we would expect the  $y$ -intercept to be fairly large and the effect of the NC variable should be interfered with. In fact, such a relation between the slope and the intercept does exist in our data. For the seven experiments there is a rank order correlation of  $-.68$  ( $p = .05$ ) between slope and intercept values. This is not artifactual since the intercept could, of course, change with the slope staying constant, and vice versa. It is therefore defensible to use the  $y$ -intercept value as an index of concurrent but unevaluated organizational schemas.

One other possible explanation that could be advanced for the category-recall relation needs to be discussed briefly. In all experiments except Experiment B the number of categories used was selected by individual Ss. It could be argued that the category-recall correlation is mediated by some individual capacity such as general intelligence, with the more intelligent Ss selecting more categories and also recalling more words. Such an explanation cannot be advanced for the data on Experiment B, however, where the number of categories to be used was randomly assigned to Ss. The correlation between NC and R in that experiment was  $.74$ , above the median for all the studies, and the slope was at  $2.9$ , below the over-all median. We have argued that the latter low figure was due to the additional constraint on the Ss. These data make an explanation based on self-selection less tenable.

#### F. CLUSTERING AND CATEGORIZATION

The clustering scores (RR) in Table II show a median of  $.68$  and a range of values, as one would expect, generally greater than that found in the clustering of  $E$ -defined categories.

Another way to evaluate the clustering behavior of Ss in this situation is to consider clustering scores as a function of NC. In order to be able to obtain relatively stable values,  $N$  per NC must be fairly large, and only Experiments B and G provided enough data for such an analysis. Figure 5 shows, for Experiment B ( $N = 43$ ), three values at each NC level: first, the obtained mean ratio of repetition (RR); second, the random RR value determined by randomizing each S's recall protocol; and finally, the maximum value that would be obtained if each S had recalled the words from each category in a single cluster.

Both the limiting values (random and maximum) are affected by NC; both decline, with the random value dropping much more steeply

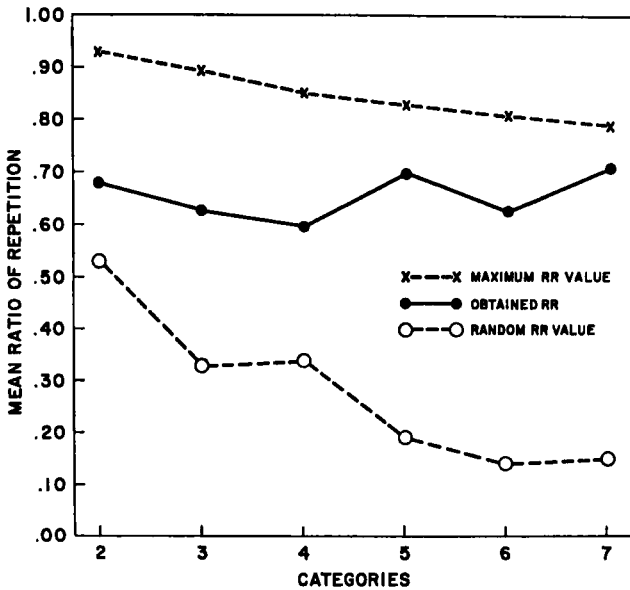


FIG. 5. Mean ratio of repetition (RR) as a function of number of categories used in Experiment B ( $N = 43$ ). Solid line shows observed data; dashed line with crosses shows maximum possible if perfect clustering had occurred; dashed line with open circles shows RR values for a random rearrangement of output.

as a function of NC. The empirical values, however, remain remarkably stable and are essentially unaffected by NC. In other words, as NC increases, Ss diverge more from the random model and approach the maximum or perfect clustering score. The same relationship was found when the clustering data for Experiment G were analyzed.

The clustering data show that the free organization of material produces a very strong tendency for members of the same category to be recalled in a cluster. This tendency apparently increases as the number of categories increases and, of course, as the total number of words recalled increases. With small NC values Ss apparently have more of a tendency (a) to switch from category to category during recall than with large NC values, where categories are recalled in more consistent clusters, or (b) to use categories not evaluated by the categorizing task.

#### G. LONG-TERM MEMORY OF ORGANIZED MATERIAL

Although no plans had been made to retest the Ss in the various experiments, it was discovered during the course of conducting Experi-

ment G that most of the *Ss* in Experiments C-G had been recruited from two lecture classes. One half week after conclusion of Experiment G, the available *Ss* in these two classes were retested for long-term recall. It should be noted that these *Ss* had no information at the time of the first recall that they would be retested, nor, as a matter of fact, had the *E*. The same *E* who had conducted the original session addressed the students in the two classes, reminded them of the experiment by describing the categorizing task and the recall, and then asked them to write on a sheet of paper all the words they could recall from that experiment.

Figure 6 shows the recall data for those *Ss* whose data were used in the final analyses of the original experiments and who were avail-

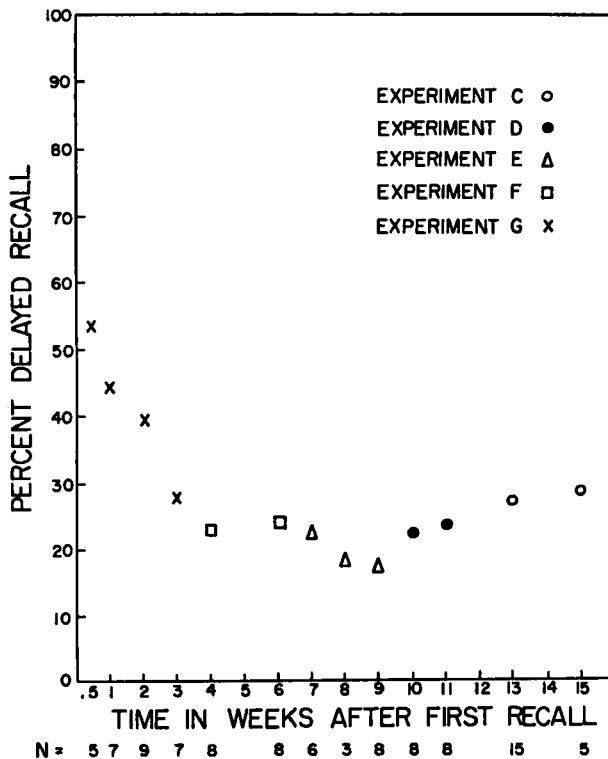


FIG. 6. Percentage of delayed recall (based on original recall) as a function of time since original recall for Experiments C-G. Bottom line on abscissa shows number of *Ss* contributing to each data point.

able at the retest. The figure shows the delayed recall as a percentage of the number of words the Ss were able to recall during the original experimental session. This percentage of delayed recall is plotted as a function of time since the original recall session. Delayed recall drops very sharply within the first 3-4 days to a little over 50% and reaches a relatively stable level of 20%-30% after about three weeks. While few data are available in the literature (cf. McGeoch and Irion, 1952) on memory decay over long periods of time, these data do seem to be more similar to long-term memory for connected, meaningful passages, rather than to the memory for lists. It appears that memory for organized material shows a sharp initial decay, but no further loss, even after three or four months. The nature of this long-term storage is further illustrated by category recall. Using a single word recalled from a category as an index of category recall, percentage of category recall is much more variable than percentage of word recall. However, category recall generally drops from about 90% to about 75% during the first six weeks, then drops to about 60% in the seventh week, whereupon it stabilizes at approximately 50%-60%. These figures should be compared with a category recall of 98% during immediate recall. It can be argued that the persisting memory of the list over 15 weeks is to some extent due to the high percentage of recall for the coding categories.

The importance of the category system for long-term recall is also supported by the clustering analysis for long-term recall. Even after 14 weeks recall is still clustered relative to a random measure. The RR for 10 Ss from Experiment C was .24 as against .16 for the random measure. However, the clustering score declines generally over time from .56 at  $\frac{1}{2}$  weeks to .24 after 14 weeks.

An important additional set of data on the relation between recall and category size will be discussed after the next section, which will present a final experiment on the category-recall function.

#### H. ORGANIZATION AND RECALL AS A FUNCTION OF INSTRUCTIONS

The following experiment was conducted for two reasons. First, having argued that organization is a necessary condition for recall and having shown that organization was directly related to recall, it was decided to explore the further step that organization is a sufficient condition for recall and that asking Ss to remember something implies that they are instructed to organize. Do instructions to organize have the same effect on recall as instructions to remember? Second, the previous series of experiments had shown that recall was apparently

unrelated to number of trials of exposure, that the recall-category relation was independent of number of trials. That demonstration, however, has been only statistical. The next experiment was designed to produce an experimental demonstration of the recall-trials-category relation.

### 1. *Method*

In the main study four groups of Ss were run, comprising the four cells of a  $2 \times 2$  design involving the presence or absence of instructions (a) to categorize or (b) to recall material. An additional group was tested for incidental memory. The five groups and the  $N$  in each group follow, with "Category" and "Recall" indicating the instructions given: (1) Category-Recall,  $N = 21$ ; (2) Category-No Recall,  $N = 21$ ; (3) No Category-Recall,  $N = 19$ ; (4) No Category-No Recall,  $N = 19$ ; (5) Incidental; No Category-No Recall,  $N = 15$ . All data were collected in five group sessions, one for each of the experimental groups; Ss were assigned to groups at random. Word lists were presented aurally at a 4-sec rate and had been pretaped for five presentation trials, each of which consisted of a random rearrangement of a 52-word list. The lists were the 52-range lists with five words changed because of their auditory confusability. Subjects in all groups were given booklets consisting of five sheets with seven columns and a final blank sheet. The instructions varied as follows.

(1) Category: Ss were told that this was an experiment in categorizing words and that they were to divide the words into any number of categories from two to seven. After hearing a word, they were to write it in a column and then add to that column any other word that went with it. They were asked to try to use the same organization on successive trials, and were not allowed to look at the category sheet of a previous trial. With a fixed number of trials, aural presentation, and written sorting, this procedure was similar to that used in the previous experiments.

(2) No Category: Ss were instructed to write the first word they heard in the first column, the second word in the second, and so forth, followed by the eighth word in the first column, and so on; and to continue this until the end of the trial.

(3) Recall: Ss were told at the beginning of the experiment that they would have to recall all the words at the end of the experiment.

(4) No Recall: Ss were given no additional instructions.

(5) Incidental: Ss were given the same words, but on a different tape on which each word was followed by a randomly selected digit from one to nine. They were asked to write the words down just as the

No Category group and were told that their task was to remember how often each of the digits appeared. At the end of each trial, they were required to write down their estimate of the number of times that each digit was heard.

At the end of the five trials, all groups were asked to write down all the words from the lists that they could remember.

## 2. Results

Table III shows the mean number of words recalled by each of the five groups. The only significant result for the four main groups was

TABLE III  
MEAN NUMBER OF WORDS RECALLED IN FIVE EXPERIMENTAL GROUPS

Categorization instructions	Recall instructions	
	Present	Absent
Present	31.4	32.9
Absent	32.8	23.5
Incidental condition		10.9

the difference between the No Category-No Recall group and the other three main groups ( $p < .01$ ). The Incidental group performance was significantly below the lowest of the other four groups ( $p < .01$ ). Either organization or recall instruction produce the same level of recall, significantly better than that produced in the absence of instructions. It could be assumed that the No Category-No Recall group was self-instructed and therefore organized to some extent, either because they expected to be asked to recall or more generally because Ss tend to organize material to some extent, even in the absence of instructions. It might be noted that half of the Ss in the No Category-No Recall group reported after the experiment that they expected to have to recall the words. It should be noted, however, that recall instructions produce the same results as organizing instructions, thus supporting our initial hypothesis. When organization is inhibited by directing Ss attention to another task, as in the Incidental group, recall drops to a level about one-third that of the organizing groups. Even in that last group, however, 4 of the 15 Ss reported that they expected to be asked to recall the words.

As far as the category-recall relationship is concerned, it should be noted that all Ss received five trials and identical amounts of time during the input presentation. Any relation between NC and R, therefore, is and must be independent of number of trials.

The change in procedure retarded speed of categorization. Only 3 of the 42 Ss in the categorizing groups reached the criterion of two identical successive trials and two of those Ss used alphabetic organization. But even with these relatively unstable organizational schemas, the same NC-R relationship is discernible as in the previous studies. The correlation between NC and R was .64 ( $p < .01$ ) for the Category-No Recall group and .53 ( $p < .02$ ) for the Category-Recall group. The respective equations for the line of best fit were  $3.7 \text{ Cat.} + 11.2$  and  $3.0 \text{ Cat.} + 13.9$ , respectively. Thus, despite a low level of organization and constant number of trials, we obtained a stable NC-R relation with three or more words added to recall per category.

This last experiment has shown that recall is a function of the number of categories used and that this relation cannot be derived from some mediating effects of number of trials; moreover, the data support the notion that both recall or organizing instructions produce equivalent organization and equivalent recall. Thus, with some of our assumptions more firmly anchored, we can return to the more general problems of the Category-Recall function. In particular, we shall now consider how category size affects recall from that category. Following this discussion we will also be able to evaluate some prior studies on the category-recall relationship.

### I. CATEGORY SIZE AND RECALL

How many items can a S recall from a category of a given size? We have already suggested the importance of this problem when we assumed that recall from sets of a given size follows the same general function whether the set is a category made up of word items or whether it is a set of categories from which categories must be recalled prior to word recall. At present we can evaluate only word recall from categories of given sizes, since our experiments were deliberately designed to maximize recall of all the categories. Data of category recall from other investigators will be compared with our data to demonstrate the generality of the relationship.

The basic data for the category size-recall function was obtained from all useable protocols in Experiments B-G. Category size varied from 1, which was permissible if Ss used more than two categories, to 96, which is, of course, only possible for 100-word lists. For each category size, all Ss who used that category size in any experiment were combined into a single group. The data from all available categories are shown in Fig. 7. For purposes of presentation a log scale is used for category size and the number of categories contributing to each value is shown on the abscissa. For large category sizes (beyond

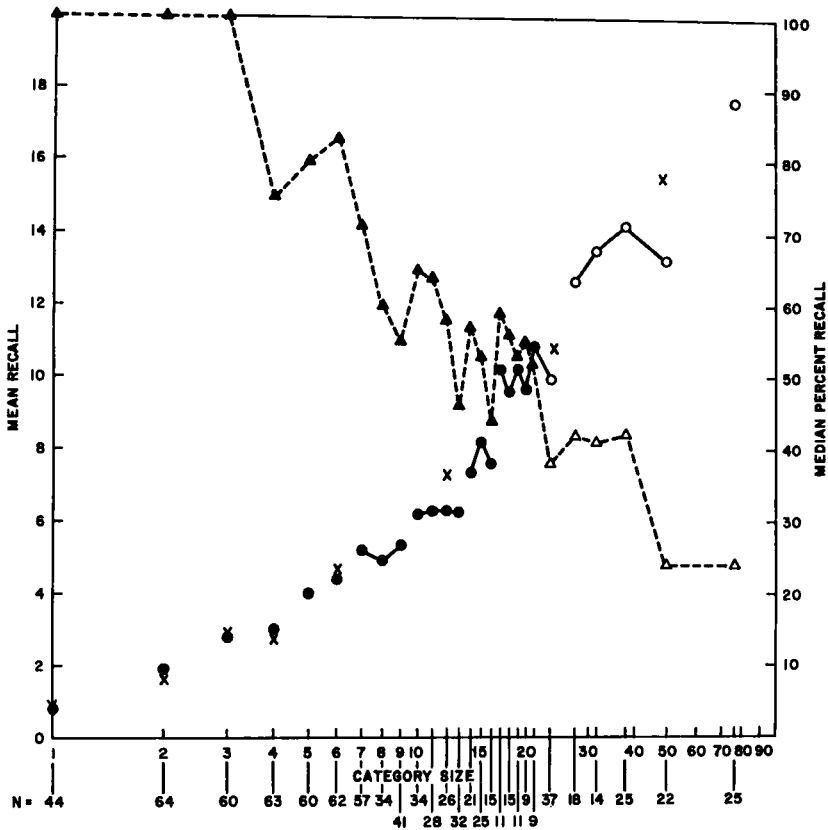


FIG. 7. Mean and median percent recall as a function of category size. Solid line, circles, and left-hand ordinate show mean recall. Triangles, dashed line, and right-hand ordinate show median percent of category size recalled. Open circles and triangles indicate that several category sizes have been combined. For the mean recall function, adjacent points that differ at the .01 level of statistical significance (*t* tests) have not been joined. The abscissa shows category size on a logarithmic scale. The bottom line on the abscissa shows the number of categories observed at each size. Single crosses show data from Tulving and Pearlstone (1966) see text for explanation.

21), data were combined for several sizes and are shown as empty circles. Figure 7 also shows recall as a percentage of category size.

The means shown in Fig. 7 form a fairly stable function of recall as a function of category size. It should be noted that the means of the low values of category size somewhat underestimate actual recall. The medians for category sizes 1, 2, 3, 4, 5, and 6 were 1, 2, 3, 3, 4,



and 5, respectively. For the larger values the means and medians agree more closely.

For small category sizes *Ss* recall close to 100% up to about size three, then drop to about 75% at sizes up to about seven items per category (IPC), after which the percentage function drops rapidly to the 50% level and reaches 25% for very large categories. In other words, the relative recall from categories decays rapidly after the  $5 \pm 2$  level has been passed. For the time being it can be assumed that up to an IPC size of seven, recall is fairly stable. But what of categories that contain more than seven items, and what of category recall as a function of the size of the set of categories? Before dealing with that problem directly, some data from several other studies must be considered.

#### J. RELATED STUDIES

Tulving and Pearlstone (1966) presented *Ss* with preorganized lists of varying lengths and varying numbers of items per category (IPC). Category size varied from 1 to 2 to 4 IPC, while list length varied from 12 to 24 to 48 words. Thus different lists contained from 3 to 48 categories—from list length 12 with 4 IPC to list length 48 with 1 IPC. During input, *Ss* were given category names and after one presentation, they were tested either with (cued) or without (noncued) presentation of the category name. Rather than investigating the relation between recall and number of categories, as Mathews (1954), for example, had done, Tulving and Pearlstone chose to investigate the relation between list length and IPC. They concluded that list length affected the number of categories recalled but not the number of words recalled within a recalled category. Their criterion of category recall was the same as Cohen's (1963), that is, recall of any one member of the category was taken as a criterion of category recall. It must be noted, however, that as list length increased, so did the number of categories within any value of IPC. Thus, category recall increased as a function of list length but also as a function of number of categories to be recalled. Tulving and Pearlstone suggest, in accord with the present orientation, that category recall and recall within a category are independent. So far so good; but they do not raise the question whether or not the mechanism of category recall is the same as the mechanism of word recall within categories. In other words, the number of items recalled from a particular category and the number of categories recalled from a set of categories should—and do—follow the same general function.

Their data permit a reanalysis that tests this proposal.<sup>2</sup> Word recall from categories was obtained from their cued condition in which Ss were given category names, thus assuring category recall, which is necessary to assess recall within a category. Recall per category was obtained from all categories of a given IPC value. This procedure provides values for recall from categories of sizes 1, 2, and 4. If, at the same time, the sets of categories are considered with 3, 6, 12, 24, and 48 categories per list, then the number of categories recalled, as defined earlier, provides data for recall of sets of sizes 3–48. These later values were taken from their Table 2, which shows number of categories recalled for different list lengths and different IPCs under noncued conditions. It is, of course, only in the noncued condition that category recall can be investigated. These values also were averaged for all sets with identical numbers of categories. The resulting data from both analyses for sets of sizes 1, 2, 3, 4, 6, 12, 24, and 48 are shown as crosses in Fig. 7. Despite the different Ss, materials, and procedures used, Tulving and Pearlstone's data fit nicely into the general relation between set size and mean recall shown for our various experiments. This concordance suggests that mechanisms for the recall of categories and of words within categories follow the same general function. Subjects recall a fixed number of words from any category of a particular size, and they recall the same proportion of categories from a set of categories of that size. Thus, although it is reasonable to argue that category recall and recall within categories are independent processes, they do follow the same size–recall function.

Since we are dealing with Tulving and Pearlstone's data, their major finding should be noted. When Ss were given category names during recall, their performance dramatically improved, suggesting that these additional words had been available but were accessible only when the category was also recalled. Interestingly enough, their data failed to show any superiority of cued recall when only three categories were used (list length 12 with 4 IPC). Considering the present argument this is not surprising since cuing for three categories cannot provide much if any advantage over the already high—if not perfect—recall of the three categories. Subjects should be able to recall three categories and cuing does not help. As soon as the number of categories becomes 6 or larger, cuing—the addition of category names—does and should produce significant advantages.

In a summary of an extended program of research on coding behavior dealing primarily with the recall of categorized word lists, Cohen (1966) has discussed the "some-or-none" characteristics of

<sup>2</sup>I want to thank the authors for making these data available to me.

coding behavior. He has shown that under a variety of different conditions *Ss* will recall a fairly constant proportion of the members of a particular category. His research has dealt primarily with category sizes 3 and 4, and therefore contributes relatively little to the generality of the present data, but some of his data on category sizes 2-5 suggest the same general function as the one shown in Fig. 7. His values are somewhat lower than those in Fig. 7, primarily because his procedure includes data from *Ss* who recall words from some categories that they have in fact not used in a categorical fashion. Thus, although Cohen points out that *Ss* recall a constant proportion from a category of a given size if they recall any word from that category, the means for category recall include data from *Ss* who may have recalled a word or two from a category without having discovered the complete categorical structure of the list. As was pointed out earlier, if *Ss* do not discover a category, then recall from that category will not follow the general mechanism apparent in category recall. The free conceptualization method insures that *Ss* not only discover these categories but, in fact, develop them in the first place.

Just as Tulving and Pearlstone did, Cohen distinguishes between category recall and recall within categories. Once a category has been recalled, "performance on these recalled categories appears to be invariant." He notes in passing that *Ss* recall about 10-14 categories from an 18- or 20-category list. These values are consistent with the function in Fig. 7. Recall within the category of categories is the same process as recall within a category.

In another study Cohen (1963) was concerned with the contrast between the recall of unrelated words and the recall of categories. He found that when presentation time was held constant, the recall of categories—scored as the recall of one or more words from a category—was similar to the recall of unrelated words. He concludes that these data support Miller's (1956a) notion that the recall of chunks should be constant, whether these chunks are single words or categorized groups of words. This general finding supports the assumption that category recall and IPC recall are limited by the same general mechanism, though they may not be influenced by the same variables.

Both Cohen, and Tulving and Pearlstone thus support the general notion that, given a particular category size, recall from that category will be constant if it occurs at all. In general, then, the recall of organized material depends (a) on the number of categories used, which determines the number of categories that will be recalled, and (b) on the size of the recalled categories, which determines the amount to be recalled from each category. When category size is constant,

the function is a simple multiplicative one of recalled categories times items recalled within categories, a relationship proposed by Tulving and Pearlstone in their paper.

This position suggests a simple trading relationship between size of the set of categories and the number of items within the categories. When the set of categories is small (6 or less), a large proportion of the categories will be recalled and, if the category is large, a relatively large amount per category will be recalled because of subcategorizing, and the total amount of recall will be large. But if the number of categories is large, the percentage of categories recalled will be relatively small and recall, particularly if the category size is small, will be relatively low. This relationship explains some data on the recall of categorized material.

Dallett, in one of his experiments (1964, Exp. IV), demonstrated a decreasing amount of recall with increasing number of categories. However, by keeping list length constant at 24 and keeping IPC constant within any one list, his lists varied from 2 categories with 12 IPC to 12 categories with 2 IPC. Given the multiplicative function suggested earlier, these two sets of values should produce identical recall. Dallett's Ss were not, however, given the category names nor told that the list contained, in the extreme case, 12 categories. In the absence of discovery of the category structure by the Ss, the set of categories from which category recall occurs is frequently smaller than 12. In that case we would expect a decreasing function as the number of categories increases. It should be noted that Tulving and Pearlstone's Ss were given the category names during input and some of their data suggest that their Ss had the category structure available at time of noncued output.

Mathew, in a pioneering experiment on categorization and recall, also used a constant-length list with varying numbers of categories and varying IPC. In her experiment, however, Ss had category names available to them at the time of recall. With 2, 3, and 6 categories per list and 12, 8, and 4 IPC, respectively, she found increasing recall as a function of number of categories. With no problem of category recall, the large drop in the proportion of items recalled between 4 and 12 IPC would predict such a function on the basis of the values shown in Fig. 7 and the simple multiplicative function discussed earlier.

The discussion of these various studies makes it apparent that experimenter-defined categories, varying numbers of categories, and varying IPC make possible any one of several functions between the number of categories used and the resultant recall. The values obtained from Fig. 7 permit the construction of ascending or descending functions, just as these various experiments have shown.

The data from our experiments permit the evaluation of the more general relationship between numbers of categories and recall. In the first instance, the categories are subject-selected and therefore not dependent on discovery during input. We can assume that all Ss have their categories available and accessible. Second, the extensive conceptualization training overtrains Ss on these categories so that there should be no loss due to forgetting of the categories up to about seven categories, and our data have shown this to be the case. Finally, by permitting Ss to vary category size freely there is no general interaction between category size and number of categories. It is true, of course, that Ss who use a large number of categories will also tend to use fewer items per category—on the average. But with any given choice of number of categories, the size of any one category may, and does, vary widely.

#### K. CATEGORY SIZE AND ORGANIZATION

The discussion of the generality of the relation between mean recall and category size should not obscure or confuse some of our earlier findings. It was concluded previously that in the recall of organized lists, Ss will, depending on conditions, recall about 3–7 words per category. In the first instance, that statement was limited to those cases where Ss use seven categories or less, and use categories that are immediately available (that is, not subject to forgetting). In the light of Fig. 7 and the subsequent discussion, what is the relevance of the number  $5 \pm 2$ ? Obviously Ss recall less when IPC is very small and as much as three times the implied limit when category size is large. For the lower values of IPC an obvious ceiling is operating and it has already been shown that up to an IPC value of about 4 the median value of recall is equal to the items in the category. For the larger values it seems reasonable to argue that whenever recall goes above 7–9 items, Ss are using subcategories within the categories. Some evidence on this point will be discussed later. For the time being all that needs to be said is that the number of items gained per category is, of course, an average value, made up of recall from different categories with varying values of IPC. The generality of the category–recall relation is not disturbed by the subsequent, more detailed, analysis that showed that there is a highly stable and general function between IPC and recall. Given a well-organized list of items, it may still be said that Ss will add about  $5 \pm 2$  items per category to their recall.

It has been noted that in situations where category recall occurs under conditions similar to IPC recall, the same function applies. Up till now we have dealt only with two levels of organization, categories

and items within categories. Is it reasonable to expect more extended organization, and what does such an assumption imply for the organization of human memory?

It is appropriate to consider evidence from some of the categories—necessarily large ones—where recall was well above the assumed limit of five to seven words per category. We will first examine such instances and then proceed to a more general discussion of higher level organization.

Even cursory inspection shows that whenever a *S* did use a very large category, the category name or label was vague and the concept used to determine category membership tended to be highly inclusive. Three *Ss* from Experiment F will be discussed here for illustrative purposes: *S* F1 used three categories to sort the 100 words and obtained a recall score of 31. His first category, labeled “senses,” contained 6 items and produced a recall of 5 items; his second category, containing 9 items, was labeled “spirit, mind, character,” and produced a recall of 5 items, while the third category, of interest here, contained 85 items, was labeled “everything else,” and produced a recall of 21 words. Looking at the recall protocol, it is quite obvious that “everything else” was subdivided by the *S* and the following list of the third-category items, in the order in which they were produced in recall, makes this quite obvious. Ellipses indicate intervening recall of items from the other categories.

GRASS	END	UNCLE
PAINT	MOUTH	GENTLEMAN
. . .	. . .	QUEEN
TABLE	MUSIC	. . .
KITCHEN	CLASS	WALK
DINNER	MEMBER	SUPPORT
. . .	. . .	. . .
SKIP	FEAR	SALT
. . .	FACT	. . .
FINGER	. . .	FLOWER

The list contains such obvious categories as “kitchen items” and “people” and other clusters that have been organized by being fitted into mnemonic devices such as images, syntactic clusters, and others.

Similar sequential clustering can be found for *S* F2, who used four categories: “verbs,” with 20 items and a recall of 5; “art,” with 17 items and a recall of 11; “abstract ideas,” with 10 items and a recall of 3; “ordinary words,” with 53 items and a recall of 19. The last category distributed its recall in the protocol as follows.

WIND	PAPER	. . .	TASTE
. . .	. . .	HOOR	. . .
CENTURY	UNCLE	YEAR	SENSE
. . .	. . .	. . .	SILVER
OFFICE	GENTLEMAN	COAT	FARMER
BUILDING	DINNER	. . .	. . .
NEWS	FOOD	SHOULDER	HAIR

And finally, the recall by *S* F3 of a category called "objects" with 64 items and a recall of 17 words was as follows.

LIFE	GENTLEMAN	BOAT
. . .	QUEEN	TABLE
SOUND	FINGER	. . .
. . .	FARMER	COMPANY
BUILDING	. . .	CROSS
MATERIAL	CENTURY	END
. . .	DAY	. . .
OIL	PERIOD	
. . .	. . .	

Some of these clusters are obvious, others are obscure, and some "intrusions" within clusters (for example, finger) are highly idiosyncratic.

These quite typical protocols suggest that whenever very large categories are used by the *Ss*, they tend to be superordinate categories that include several subordinate categories. Few categories could be more superordinate than "everything else." It should also be noted that the recall from these large categories occurs in clusters of from one to five words. While *Ss* frequently recall all the words from a small category in one sequential cluster, the recall from the large clusters is broken up into smaller sequential units.

In other words, when categories with large membership are formed, they tend to be superordinate categories with recall from these categories being determined by the number of subordinate categories they contain. We can now refer back to Fig. 7 and consider the general function represented there. It seems likely that up to about 10-15 IPC we may be dealing with a single function with recall being directly determined by category size. Above these values, both theoretical consideration and the foregoing suggestive evidence support the notion that these categories produce recall as a function of the categories they themselves contain. In Fig. 7, the adjacent points of the mean recall graph that are significantly different at the .01 level have not been joined. The data suggest that the function contains several discrete recall levels. Figure 7 shows particularly distinct plateaus at 8, 10, 14,

and 18 words of mean recall. Pending further investigation this suggests that categories of large sizes may contain, on the average, 2-5 subcategories with recall of about 3-5 words from each of these categories. The number of categories will depend on the size of the larger category, producing about 2 when category size is around 20 and rising to about 5 subcategories when category size is as large as 75.

The abscissa of Fig. 7 also shows the number of categories of each size. Although these values suggest that the preferred category sizes are 1-7, containing 44% of all categories, another 32% of the categories fall in the size range 8-15, with a respectable 24% in the 16-96 range.

## VII. The Organization of Memory

It is now possible to suggest the general outline of the organizational system. We assume first the basic limit of the organizing system at  $5 \pm 2$  per set of items. For any single chunk the organism can handle only that many units. Given that limitation, categories will be formed and 3-7 items assigned to them. We might note in passing that some categories will be smaller than that simply because the list may contain only one or two relevant items. Once these initial categories are filled up, new categories will be created to accommodate additional items. But in turn, there will be a limit of about  $5 \pm 2$  categories at this first level of categorization. When all slots are taken up with first-level categories, a second level of categories will be formed, each of which may contain up to about seven first-level categories, and so forth. In this manner, a hierarchical system of categories can be built up with an increasing level of complexity and an exponential growth in the size of the system. We will return shortly to some speculations about the size of these organized systems. First, two additional items of evidence for this general scheme are relevant.

Applying these notions to multitrial experiments of free recall suggests the following mechanism to explain the increase in free recall as a function of repeated exposure to a particular list of items. We agree with Tulving (1964) and others that the free recall experiment using words is not a learning but a retention situation; that is, the *S* must retain the items presented on any one trial. In the sense of response learning, the items have been "learned" prior to the experiment and in terms of retention, any single word could be retained if it were presented alone. Thus, an item is "learned" at the time of presentation.

The present analysis assumes that at the time of the first trial, when *S* is given information as to list length and the words contained



in the list, the processing system establishes the requisite number of categories, probably and preferably about five, to which the words are to be assigned. If list length is about 25 or less (that is, 5 categories with 5 items per category), this should present no particular problem. If it is longer, it is likely that superordinate categories are established in order to accommodate eventually all the words in the list. Recall after the first trial probably reflects category recall, that is, approximately one word from a large proportion of the categories. On subsequent trials these categories are then "filled up" with items up to the capacity of  $5 \pm 2$ . Given a constant number of categories, the optimal strategy might be to add one item to each category on each trial. There are serious limitations to such a process, since it is highly likely that the initial categorization might undergo changes in order to accommodate the items in the list, and it also might prove difficult to assign every word to a particular category as fewer items remain to be organized.

However, support for some such process can be found in the arguments and data presented by Tulving (1964). Tulving showed that, except for artifactual effects, intratrial retention (the number of items recalled on a trial that were not recalled on the previous trial) is constant across trials, whereas intertrial retention (items recalled on trial  $n$  that were also recalled on trial  $n - 1$ ) increases as a function of trials. The constant intratrial effect is consistent with the reasoning presented earlier. Tulving also presents some data that suggest that the value for the intratrial retention component increases with list length. With a list length of 22 words, both initial recall on Trial 1 and the intratrial retention value for early trials were about five. With a list length of 52, the intratrial retention value is close to nine. Furthermore, the SO (subjective organization) values increase as a function of trials and covary with intertrial retention. In terms of the present model, with increasing trials, more and more words from the category are recalled in clusters.

These suggestions also offer a possible explanation why single-trial recall varies as a function of list length. The longer the list, the more initial categories will be established after a single exposure. Finally, the model predicts that multitrial free recall experiments should produce relatively inefficient learning in the sense that performance cannot reach the asymptote of 100% recall for relatively long lists. The inefficiency and subsequent rigidity of the initial category system prevents the organization of all the items and eventually prevents some items from being recalled. Contrast such a system to one that permits the *S* to organize the list prior to recall, as in our studies.

For example, the three *Ss* in Experiment G who used 8 or more categories recalled an average of 76.7 items out of 100 with an average of only 4.6 categorization trials.

The rigidity of established categorical organizations has recently been illustrated by Ozier (1965). Her *Ss* were given varying numbers of trials of free recall followed by instructions to recall according to alphabetic categories. The data unequivocally show a drop in recall on the trial immediately following the instructions, with the size of the decrease being a direct function of the number of previous recall trials. In other words, the organization imposed by the *S* becomes increasingly fixed and more difficult to exchange for a new organizational schema. Similarly, it is unlikely that the organizational schema can be changed in the late trial of a free recall experiment in order to accommodate items that do not "fit" the previously established categories.

In the experiments presented here, no attempt was made specifically to investigate the power of superordinate categories. Cohen and Bousfield (1956), however, have studied the effect of single and double level of categorization on recall. Using forty word lists, they present data for three kinds of lists: (1) four-category single-level lists with 10 IPC; (2) eight-category single-level lists with 5 IPC; and (3) eight-category dual-level lists, that is, lists with four categories, each of which had two subcategories, and with 5 IPC. In terms of the present analysis, recall should improve in ascending order for these groups. In the first group, *Ss* must recall 4 categories and 10 IPC; in the second, 8 categories and 5 IPC, and in the last, 4 superordinate categories times 2 subordinate categories with 5 IPC. Taking approximate values for these groups from Fig. 7, the three recall means should be 19.2, 20.1, and 24.1 for lists 1, 2, and 3, respectively. Cohen and Bousfield's data show values of 15.6, 17.6, and 18.1. Since their lists were *E*-constructed categories, the lower levels of recall can be expected.

Another question that our analysis raises concerns the way in which words in general may be recalled when, for example, somebody is asked to say all the words he can think of, or all the animals or countries. The prediction must be that such recall from general storage should proceed in the same way as the special categorization imposed in the laboratory. Superordinate categories must be followed by subordinate, and so forth until the search system comes to a first-level category, recalls about five words from it, proceeds to another category, and so forth. Bousfield and Sedgewick (1944) have shown that when *Ss* are asked, for example, to list all the birds they can think of, they will in fact produce these in clusters of subcategories with some evidence

that these clusters occur in temporally discriminable sets, that is, with short pauses separating the clusters. In other words, recall from permanent vocabulary storage follows the same general organizational schema as the assignment of specific words in a memory experiment. We can also assume that the categories *Ss* use in the conceptualization and memory experiments are very similar to those that are represented in free emission. In that sense, the experimental situation simply utilizes the existing organization of the *Ss*' vocabulary. Some new organization may at times be imposed to accommodate unusual words or clusters, but generally the memory experiment is an experiment on the utilization of existing organizational schemas.

What are the limits of this kind of organization? Taking a value of five per set, we have suggested that the total content of an organizational schema rises exponentially with about five new units per level. It seems possible that the system also needs some limits on the number of levels that can be contained in a single organizational schema. One reason for this is the need to identify the level at which a particular search starts, since the level may influence a decision whether to go down the hierarchy, or up, or across. Similarly, the ease with which one can identify superordinate and subordinate concepts suggests that level identification is both useful and necessary. If levels are identified, we can assume that the limit on this task is also five, which then limits the content, in terms of final units, of an organizational schema to the value of approximately  $5^5$  or about 11–12 bits. Is this the limit of human memory? By no means; these speculations have only touched on a single schema. Obviously a particular unit may be contained in more than one schema, and some units or words may be in one schema and not in another. It is difficult to determine whether the number of such parallel schemas is in turn limited in light of the vast overlap of different organizations and the very specialized organizations that we construct. However, if such limits do in fact exist, they provide some interesting basis for further investigation into limits on the size of natural language vocabularies. On a highly speculative note, two suggestions might be entertained. First, the organization of any single coherent natural vocabulary may be limited to the value of  $5^5$  items. It is enticing to note that such divergent vocabularies as the basic sign language of the deaf, the ideographic vocabulary taught to the Japanese school child, and the basic vocabulary taught in foreign language schools all tend to fall at about 1500–2000 items, a value nicely between  $5^4$  and  $5^5$ .

Second, it is possible that separate, though overlapping, organizational schemas may be organized at a still higher level of schemas of schemas.

Again assuming that the identification of schema membership is necessary for storage and retrieval, such a superorganization would contain another five levels and would produce an estimate of  $5^{55}$  units that could be stored. Such a figure, in contrast to  $5^5$ , is reassuringly large. It involves about  $10^{17}$  units, certainly adequate for storing any reasonable set of human memory units.

A problem that this paper and most psychologists have avoided concerns the functional unit of memory. At the verbal level, a psychologist is tempted to say that the unit of behavior is the word, even though groups of words may, of course, make up larger units. The recent rapprochement between linguistics and psychology, on the other hand, has tempted some to speculate that verbal units may be phonemes or morphemes.

At the theoretical level it is necessary to speak of units as constructs. Such units have the main characteristics of being activated in an all-or-none manner and of being emitted in the same fashion; that is, it is not possible either to activate or emit part of a unit. If such partial activation or emission is possible, this would be *prima facie* evidence that the unit has constituents. None of these suggestions solve the problem of the psychological unit; they postpone the important issues. Eventually we must come to terms with the theoretical unit, which may be an image, an idea, a word, or a category (cf. Morton and Broadbent, 1964).

For the present we have confined ourselves to nouns, though the organization of other verbal units, such as adjectives or verbs, should follow similar laws. The restriction to nouns has also avoided the problem of the role of syntax in verbal memory, though grammatical considerations obviously play an important role in the organization of memory (cf. Cofer, 1965).

Within these rather restricting limitations, this chapter has talked about the organization of memory. But as we have seen previously, memory—the sheer recovery of a set of units—is just one outcome of organization. Given an organized set of units, we can recall some or all of these units according to rather simple rules. Given knowledge of the organization, we can predict with a fair degree of accuracy the amount of recall that is possible when the system is instructed to emit the constituents of the organized set. In our speculation about the organization of the available vocabulary, we have suggested that any memory experiment with words (that is, with units that are in the vocabulary), is just one way of tapping already existing organizations. In that sense, then, this chapter was not really about organization *and* memory, it was about the organization of parts of the human verbal repertory

and it used memory as a way of evaluating what that organization might be. Granted that the conditions of presentation or input present certain limiting conditions for what can or will be recalled, it seems quite certain that the major limit on the memory for words is the organization of verbal units. Such organization is fully developed in adult Ss and probably changes little over time. If we are to investigate how organization develops, we must go to the developmental study of language, semantics, and verbal behavior. That is probably the only source that will tell us about the development of organizational schemas.

#### ACKNOWLEDGMENTS

I am extremely grateful to Dr. Donald E. Broadbent of the Applied Psychology Research Unit, Cambridge, England, for making available the congenial facilities of his laboratory, where this chapter was written in July, 1965.

Mrs. Shirley Osler contributed faithful and hard work, imaginative suggestions, and endless patience to the execution and analysis of Experiments B-G. Miss Leslie Waghorn collected the data for the experiment on organizing and recall instructions. I am greatly indebted to William Kessen, George A. Miller, Endel Tulving, and Jean M. Mandler for critical comments on a previous draft of this paper. None of them can be blamed for any errors or misinterpretations that it still contains.

#### REFERENCES

- Bousfield, W. A. The occurrence of clustering in the recall of randomly arranged associates. *J. gen. Psychol.*, 1953, **49**, 229-240.
- Bousfield, W. A., Cohen, B. H., & Whitmarsh, G. A. Associative clustering in the recall of words of different taxonomic frequencies of occurrence. *Psychol. Rep.*, 1958, **4**, 39-44.
- Bousfield, W. A., Puff, C. R., & Cowan, T. M. The development of constancies in sequential organization during repeated free recall. *J. verb. Learn. verb. Behav.*, 1964, **3**, 489-495.
- Bousfield, W. A., & Sedgewick, C. H. W. An analysis of sequences of restricted associative responses. *J. gen. Psychol.*, 1944, **30**, 149-165.
- Cofer, C. N. On some factors in the organizational characteristics of free recall. *Amer. Psychologist*, 1965, **20**, 261-272.
- Cohen, B. H. Recall of categorized word lists. *J. exp. Psychol.*, 1963, **66**, 227-234.
- Cohen, B. H. Some-or-none characteristics of coding behavior. *J. verb. Learn. verb. Behav.*, 1966, 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. *J. gen. Psychol.*, 1956, **55**, 51-58.
- Cohen, B. H., Sakoda, J. M., & Bousfield, W. A., The statistical analysis of the incidence of clustering in the recall of randomly arranged associates. Tech. Rep. No. 10, ONR Contract Nonr-631(00), Univer. of Connecticut, 1954.
- Dallett, K. M. Number of categories and category information in free recall. *J. exp. Psychol.*, 1964, **68**, 1-12.
- Garner, W. R. *Uncertainty and structure as psychological concepts*. New York: Wiley, 1962.

- Imai, S., & Garner, W. R. Discriminability and preference for attributes in free and constrained classification. *J. exp. Psychol.*, 1965, **69**, 596-608.
- Katona, G. *Organizing and memorizing*. New York: Columbia Univer. Press, 1940.
- Mandler, G. Verbal learning. In *New Directions in Psychology III*. New York: Holt, 1967.
- Mandler, G., & Pearlstone, Zena. Free and constrained concept learning and subsequent recall. *J. verb. Learn. verb. Behav.*, 1966, **5**, 126-131.
- Marshall, G. R., & Cofer, C. N. Associative, category and set factors in clustering among word pairs and triads. Tech. Rep. No. 4, Contract Nonr 285 (47), Office of Naval Research, New York Univer., 1961.
- Mathews, Ravenna. Recall as a function of number of classificatory categories. *J. exp. Psychol.*, 1954, **47**, 241-247.
- McGeoch, J. A., & Irion, A. L. *The psychology of human learning*. New York: Longmans, Green, 1952.
- Melton, A. W. Implications of short-term memory for a general theory of memory. *J. verb. Learn. verb. Behav.*, 1963, **2**, 1-21.
- Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 1956, **63**, 81-97. (a)
- Miller, G. A. Human memory and the storage of information. *IRE, Trans. Inf. Theor.*, 1956, **2**, 129-137. (b)
- Miller, G. A. *Psychology: The science of mental life*. New York: Harper, 1962.
- Miller, G. A., Galanter, E., & Pribram, K. H. *Plans and the structure of behavior*. New York: Holt, 1960.
- Morton, J., & Broadbent, D. E. Passive vs. active recognition models or Is your homunculus really necessary. Paper read at the AFCRL Sympos. on models for the perception of speech and visual forms, Boston, November, 1964.
- Ozier, Marcia. Alphabetic organization in memory. Unpublished doctoral dissertation, Univer. of Toronto, 1965.
- Seibel, R. An experimental paradigm for studying the organization and strategies utilized by individual Ss in human learning and an experimental evaluation of it. Paper presented at meetings of The Psychonomic Soc., Niagara Falls, Ontario, October, 1964.
- Tulving, E. Subjective organization in free recall of "unrelated" words. *Psychol. Rev.*, 1962, **69**, 344-354.
- Tulving, E. Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychol. Rev.*, 1964, **71**, 219-237.
- Tulving, E., & Patkau, Jeannette E. Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canad. J. Psychol.*, 1962, **16**, 83-95.
- Tulving, E., & Pearlstone, Zena. Availability versus accessibility of information in memory for words. *J. verb. Learn. verb. Behav.*, 1966, **5**, 381-391.
- Underwood, B. J. The representativeness of rote verbal learning. In A. W. Melton (Ed.), *Categories of human learning*. New York: Academic Press, 1964. Pp. 47-78.
- Waugh, Nancy C. Free versus serial recall. *J. exp. Psychol.*, 1961, **62**, 496-502.
- Waugh, Nancy C., & Norman, D. A. Primary memory. *Psychol. Rev.*, 1965, **72**, 89-104.
- Wertheimer, M. Untersuchungen zur Lehre von der Gestalt. I. *Psychol. Forsch.*, 1921, **1**, 47-58.