CONCURRENT EFFECTS OF CONTEXTUAL CONSTRAINT AND
WORD FREQUENCY ON IMMEDIATE RECALL AND
LEARNING OF VERBAL MATERIAL

ENDEL TULVING AND JEANNETTE E. PATKAU
University of Toronto

One of the obvious operational distinctions among relevant dimensions of stimulus materials in verbal learning is the distinction between characteristics of individual items and of sequences of items. A list of words, for instance, may be more easily memorized because it contains more “meaningful” words, or because the sequence of words is more “meaningful,” or because of both conditions. Meaningfulness of items, in a very broad sense, can be varied independently of characteristics of their sequence, and properties of sequences can be varied while holding constant relevant characteristics of items.

In most experiments investigating the effects of the nature of material on verbal learning, task variables have been specified in terms of properties of individual items. Summaries of representative experiments are provided by McGeoch and Irion (1952), Noble (1961), and Underwood and Schulz (1960). The first known attempt to investigate the effect on memorization of quantitatively specified characteristics of sequences of items was that of Ebbinghaus (1885) who used the method of derived lists. More recently, Miller’s method of constructing sequences of words at various orders of approximation to English (Miller & Selfridge, 1950) has formed the basis of some systematic research (for example, Deese & Kaufman, 1957; Marks & Jack, 1952; Postman & Adams, 1960; Richardson & Voss, 1960; Sharp, 1958). In a slightly different vein, Epstein (1960) has demonstrated the effects of syntactical structure on the learning of nonsense material; and it is even conceivable that levels of “grammaticalness” (Chomsky, 1957) could be quantified to yield a continuously variable dimension relevant to learning and retention. By and large, however, experiments studying the effects of different sequences of items have been an exception rather than a rule.

The fact that it is possible to quantify characteristics of individual items independently of their membership in any experimentally controlled sequence, and those of sequences independently of constituent items, does not necessarily mean that the effects of these two broad classes of variables are mediated by disparate mechanisms. On the other hand, the assumption that the effects of variables such as associative value and taxonomic frequency of occurrence of words are additive to the
effects of characteristics of sequences should not be accepted without directly relevant experimental evidence. Such an assumption often has been implicit in experiments that have used Miller-Selfridge (1950) lists as learning materials. Miller and Selfridge's method assigns word frequency and contextual constraint to the same continuum defined in terms of redundancy of information. In fact, however, first-order lists differ from zero-order lists in word frequency, but not in the degree of sequential dependency, and from higher-order lists in sequential dependency, but not in word frequency. Consequently, zero-order lists are less redundant than second-order and higher-order lists because of differences in both the average probability of individual words and the average probability of sequences of words.

The main purpose of the present paper was to investigate the concurrent effects of word frequency and contextual constraint on recall of words on several successive practice trials. It is quite possible that the influence of word frequency on recall depends on the statistical structure of the list, or that contextual constraint has a different effect on recall of words at different levels of frequency of occurrence. The finding of the presence or the absence of such an interaction would contribute to our empirical knowledge of the effects of task variables on remembering, and provide still another testing ground for different theories of verbal learning.

**Method**

**General Design**

The present experiment can be thought of as a partial replication and an extension of Miller and Selfridge's (1950) experiment. Two levels of frequency of words were orthogonally combined with five degrees of contextual constraint to yield ten lists of 24 words. Each of two experimental groups of 25 Ss learned five lists at one level of word frequency. Six practice trials, each consisting of auditory presentation of the list in a fixed order and a written recall in free order, were given to Ss on all five lists. The main data of the experiment were provided by the number of words correctly recalled on each trial. An additional analysis compared the orders of words in recall with that of input lists.

**Materials**

As the first step towards construction of experimental lists, two populations of words were defined in terms of frequency of occurrence in printed English. One population consisted of all words occurring at least 100 times per million in the Thorndike-Lorge (1944) general word count. These words will be referred to as high-frequency (H) words. The second population consisted of words occurring four times per million in the general word count. These words will be called low-frequency (L) words. The sizes of these two populations are approximately equal at 1069 and 1064 words respectively. At one level of frequency, lists were made up exclusively of H words (H list). At the second level the list items alternated between
H and L words (HL list), since it is not possible to construct all the sequences used in the experiment of L words alone.

At each of the two levels of list frequency first, third, fifth, seventh, and text orders of approximation were constructed with the help of 12 volunteer Ss, graduate students in psychology, as follows. The general approach was the same as that used by Miller and Selfridge (1950), but certain modifications were dictated by the present design. In composing a list at the third, fifth, and seventh order of approximation S was given a sequence of two, four, or six words, either H words or HL words, constituting a part of a meaningful sentence. The S then selected a word, on the basis of a table of random numbers, from the population of either H or L words, as the case might require, and tried to make up another sentence in which the stimulus sequence was followed by the selected word. When this was found possible, the selected word was retained, the first word of the sequence dropped, and the remainder of the sequence plus the added word given to the next S who repeated the procedure. When the word first selected could not be made to fit the sequence, another word was selected, and the procedure repeated. Twenty-four words were successively generated by Ss in this manner for each of three lengths of the stimulus sequence, and at both levels of frequency. These words constituted the experimental lists at the third, fifth, and seventh orders of approximation. One of the Es constructed the first-order and text sequences, the former by a random method of selection, the latter by drawing words at random and by trying to construct a legitimate sentence. No proper names were used in the lists, homonyms were avoided as far as possible, and no word was permitted to appear more than once in each list.

Subjects

Fifty summer school students at the University of Toronto, 18 men and 32 women, served as Ss. They were randomly divided into two experimental groups, with the restriction that nine men and 16 women were assigned to each group. The mean age of Ss in the first group was 26 years, in the second group 25 years, and the over-all range in age was from 19 to 49 years.

Procedure

The first group of Ss learned all five H lists, the second all five HL lists. Each experimental group was randomly subdivided into five groups of five Ss each. All members of one subgroup were tested together. The order in which Ss learned lists was counterbalanced among the five subgroups of an experimental group, but the members of a subgroup received an identical order.

The lists were recorded on magnetic tape in a monotone female voice, at the rate of approximately one word per sec. The material was presented to Ss by playing the tape. The operative parts of the instructions were as follows: "This is an experiment to see how well people can remember lists of English words. You will listen, without writing anything, to a list of 24 words. Then . . . you will write as many words as you can remember, in any order you remember them. . . . You will hear the same list a second, . . . third, fourth, fifth, and sixth time and you will be asked to recall after each trial. . . . There will be five lists of words and you will hear

---

1There are no zero-order lists with the present procedure. The dimension between zero and first order corresponds to the dimension of word frequency.
2Lists of words used in this experiment are available from the authors.
3We are grateful to Miss Cecille Gold for the recording of lists.
each list six times as before. Your final score on any list is the total number of words correctly recalled on all six trials, so write down all of the words each time.”

The Ss recorded their recall in five separate six-page booklets of plain paper. They were allowed two minutes after each trial for recall. The complete experimental session occupied approximately one hour and 15 minutes.

RESULTS

Number of Recalled Words

The primary data from this experiment are provided by the number of words correctly recalled on each trial. Repetitions of list words, as well as extra-list intrusions were ignored in the analysis. The mean number of such “errors” was 0.25 per trial for H lists, and 0.16 for HL lists.

![Graph showing the percentage of words correctly recalled as a function of order of approximation to the statistical structure of English](image)

Figure 1. Mean immediate recall as a function of order of approximation for H words and HL words. Miller and Selfridge’s data are included for comparison.

The data will be first summarized for Trial 1, on which comparison with Miller and Selfridge’s (1950) results is possible. Then the analysis will be extended to cover all six trials.4

The percentage of correct recall on Trial 1 is shown in Figure 1. The lowest curve shows the change in the immediate recall of HL lists as a

---

4The raw data are given in full in Jeannette Patkau’s M.A. thesis, on file in the Department of Psychology, University of Toronto.
function of contextual constraint. Each point on the curve is the mean from 25 subjects. The same data for the H lists, based on means from the other group of 25 subjects, are shown by the middle curve. The upper curve in Figure 1 is an interpolation between 20-word lists and 30-word lists in Miller and Selfridge’s experiment.

It is reasonable to assume that in general the characteristics of individual words in Miller and Selfridge’s stimulus lists correspond to those in our H lists. The obvious discrepancy between the two sets of immediate recall data is therefore somewhat surprising. It is probably related to differences in methods of constructing lists. Since the selection of words for our lists was initially determined by the table of random numbers, inter-item associative bonds were probably not only more variable, but also generally weaker in our lists than the sequential connections among words in Miller and Selfridge’s lists.

The most prominent feature of our own immediate recall data, as shown in Figure 1, is the apparent interaction between frequency and contextual constraint. The effect of order of approximation on the recall of H lists is monotonic and noticeably stronger than for HL lists. For the present sample, the recall of HL lists is at a maximum at the fifth order of approximation. While we cannot make very strong statements about these relations on the basis of only single samples of material from each combination of level of frequency and contextual constraint, there seems little doubt that (1) the average frequency of list words has only a relatively small effect on the recall of unstructured material, such as at first-order approximation, and that (2) the advantage of H lists over HL lists increases with degree of contextual constraint.

An analysis of variance on the immediate recall data, with variance contributed by the five orders of learning lists ignored, yielded an $F$ of 20.2 (4 and 192 df) for the interaction effect between order of approximation and familiarity. This $F$ is highly significant and provides statistical support for the statement of interaction made on the basis of the graphical description of the data. A separate one-way analysis of the HL data alone yielded an $F$ of 3.50 (4 and 120 df) which is significant at .05 level. Thus we can conclude that although the improvement in immediate recall of HL material with higher orders of approximation is relatively small, it cannot be readily accounted for in terms of chance factors only. The effect of contextual constraint is simply considerably greater for H lists than for HL lists.

When the analysis of data was extended to include data from all six trials, a significant three-way interaction among frequency, contextual constraint, and practice was found ($F = 3.48, 20$ and $960$ df). Although this suggests that the rate of learning depends on both frequency and
contextual constraint, an inspection of the actual learning curves for the
ten lists shows that the effect is neither very large nor systematic.
The learning curves are shown in Figure 2 and Figure 3. Figure 2
shows the number of words correctly recalled from H lists on six trials,
with order of approximation as the parameter. Figure 3 depicts the same
data for HL lists. The triple interaction effect suggested by the variance
analysis seems to be mostly attributable to the small differences among
the slopes of the curves of the five H lists, since the differences among the
HL curves seem negligible, and the slope of the first-order HL curve is
almost identical with the slope of the first-order H curve.

These findings are clear-cut. The effects of word frequency are not
additive to those of contextual constraint. The question is, why not? We
might say that in sequences of high contextual redundancy the pre-
experimental associative bonds are stronger than in less redundant
sequences, and that they are more so in case of high frequency words
than in case of low frequency words, with the consequence that more
words are recalled from H lists and higher orders of approximation. This
seems to be essentially the interpretation put forth by Miller and
Selfridge at the time of their experiment: “Meaningful material is easy
to learn, not because it is meaningful per se, but because it preserves
the short range associations that are familiar to the Ss” (1950, p. 183).
The main difficulty with this explanation is that it is too vague to be very
useful. It cannot be readily extended to cover interactions such as we
found in this experiment.

Rather than looking for an ad hoc explanation of the findings, we
decided to look at the data again. After all, the findings presented so far
pertained to the number of words in recall, and words as response
units are as arbitrary as any other possible units. Words may be funda-
mental units in the sense that they are the end result of a lot of prior
encoding of their constituent letters or phonemes (Deese, 1961), but
there is no particular reason to assume that sequential encoding ends
with words. It may well extend over the boundaries of individual words,
particularly in well-organized contexts (Deese & Kaufman, 1957).

Sequential Analysis of Recall

In this experiment there were no external restrictions on the order in
which subjects had to recall words. The variability in the order of words
in recall, therefore, can be used as an additional source of information
about the effects of the task variables manipulated in the experiment.
Response measures based on order of recall have been used before in
Miller-Selfridge type experiments. Sharp (1958) counted couplets of
words in analysing his data, Postman and Adams (1960) have reported
Figure 2. Learning curves for H lists. Mean number of words correctly recalled is shown as a function of trials, with order of approximation as parameter.

Figure 3. Learning curves for HL lists. Mean number of words correctly recalled is shown as a function of trials, with order of approximation as parameter.
on the effect of order of approximation on clustering of words in recall, and Marks and Jack (1952) gave credit to subjects only for recalling words in the order of the stimulus list.

The analysis undertaken here was suggested by Miller's (1956a, 1956b) conceptual model of remembering. Miller holds the view that a subject's memory is limited to some seven, plus or minus two, "chunks" of information. Whenever it seems that the subject's memory span is greater than this relatively invariant number of basic units, such as we observe following repetition or in case of highly meaningful materials, it is primarily because the subject has recoded or organized into a chunk several individual items, and not because anything has happened to his basic capacity to handle a maximum number of seven or so independent units of information.

One of the difficulties with this model lies in the identification of chunks. If we observe a subject recall ten individual words, how can we decide how many chunks there are? The method we used in this experiment is based on the assumption that the subject adopts some sequentially organized groups of items from the input list. Thus we define an "adopted chunk" as any group of one or more items which occur in a subject's recall in the same sequence as in the input list. For example, if the stimulus list consists of a sequence A, B, C, D, ..., V, W, X, and the subject recalls T, U, V, A, B, C, D, E, F, R, K, in this order, we assume that the ten items the subject recalled are organized into five adopted chunks: TUV, AB, DEF, R, and K. The average size of these chunks, in terms of individual items, is 2.00. Since it is possible that some chunks are formed independently of the contiguity of input items (Tulving, 1962), the number of adopted chunks, as defined here, probably underestimates the number of all chunks in the subject's recall. In view of evidence from experiments such as Deese and Kaufman's (1957), it is also quite probable that the underestimation is more pronounced with less organized materials.

Because of certain complications that arise for this analysis as a result of the extended practice the subjects have with the material, the analysis of adopted chunks was restricted to data from Trial 1 only. These immediate recall data are shown in Table I. It shows the mean number of adopted chunks for all ten lists, together with their standard deviations.

An analysis of variance yielded a significant F only for order of approximation \( F = 2.95, 4 \) and 192 \( df, p < .05 \). Word frequency was not a significant variable \( (F < 1.00, 1 \) and 48 \( df) \); neither was the interaction between frequency and order of approximation \( (F = 1.68, 4 \) and 192 \( df) \). Thus, if we use adopted chunks as our response units, the findings are
TABLE I
MEAN NUMBERS OF ADOPTED CHUNKS IN SUBJECTS' RECALL ON TRIAL 1
AND THEIR STANDARD DEVIATIONS

<table>
<thead>
<tr>
<th>Order of approximation</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>Text</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>H lists</td>
<td>Mean</td>
<td>5.00</td>
<td>4.60</td>
<td>5.80</td>
<td>6.04</td>
<td>5.80</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.57</td>
<td>1.52</td>
<td>1.79</td>
<td>2.13</td>
<td>1.52</td>
</tr>
<tr>
<td>HL lists</td>
<td>Mean</td>
<td>4.96</td>
<td>5.44</td>
<td>6.04</td>
<td>5.40</td>
<td>5.12</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.84</td>
<td>1.57</td>
<td>2.20</td>
<td>2.00</td>
<td>2.13</td>
</tr>
</tbody>
</table>

Figure 4. Mean number of words and mean size of adopted chunks on Trial 1 as a function of order of approximation.

quite different from those based on the frequency of recall in terms of individual words. The over-all means for H lists and HL lists are almost identical at 5.46 and 5.39 chunks, and even the absolute differences among means for various levels of contextual constraint are rather small.

From the relative constancy of the number of adopted chunks in Table I and from the data shown in Figure 1, it follows that changes in the amount of recalled words, as a function of both word frequency and contextual constraint, should be accompanied by corresponding changes in the size of chunks. This correspondence is graphically depicted in
Figure 4, in which both the number of words recalled and the mean size of adopted chunks on Trial 1 are shown as functions of order of approximation (abscissa) and frequency of list words (parameter). The size of the adopted chunks is to be read from the right hand ordinate, and the number of words recalled, from the left. The values of the two ordinates have been selected to reflect the finding, based on the data in Table I, that a typical subject can remember no more than five or six unrelated words or chunks.

A legitimate question with respect to this kind of sequential analysis of recall concerns the mathematical relation between number of recalled words and size of adopted chunks. It is conceivable that chunking occurring by chance varies directly with the number of items in the recall sequence. If this chance effect is pronounced, the psychological significance of our findings is weakened.

Some chance chunking undoubtedly occurs. Postman and Adams (1960), who were faced with the same problem in their analysis of recall from Miller-Selfridge lists, generated recall data for 100 statistical subjects and indeed found some clustering. However, the amount of statistical clustering was relatively small, and there seemed to be no apparent increase in it for longer recall sequences. In the light of these data, the correspondence between frequency of recall and size of chunks in this experiment probably reflects a true psychological relation.

DISCUSSION

We began this study with the question about the combined effects of word frequency and contextual constraint on recall. We end by pointing out that answers to experimental questions depend upon units of analysis, and that for this reason our initial question can be answered in several ways.

When we used the number of recalled words as the dependent variable, we found that both frequency and contextual constraint influenced recall in a non-additive manner, and that the increase in recall of high frequency words with increasing orders of approximation was linear and more rapid than the increase in recall of lists containing less frequent words. However, when we used the number of adopted chunks as the response measure, our findings were different. Word frequency had no effect on recall following Trial 1, nor did it interact with contextual constraint. Only contextual constraint had a small, but statistically significant effect on the number of recalled chunks.

The finding that the number of recalled chunks is relatively invariant—it ranges from 4.6 to 6.0 in contrast to a range from 6.0 to 15.1 for the number of words—is quite interesting in the light of Miller's (1956a,
1956b) conceptual model of remembering. Subjects seem to remember some five or six chunks of information regardless of the nature of the material. This number is quite close to Ebbinghaus' (1885) memory span of seven syllables, and to Miller's (1956a) magical number seven, plus or minus two. The observed invariant number may reflect an artifact or a coincidence of some kind, but if these "coincidences" keep piling up, we ought to take them seriously.

When these data are placed into the framework of Miller's unitization hypothesis, the problem of why subjects remember more high frequency words than low frequency words and the problem of interaction between word frequency and contextual constraint in immediate recall disappear, or at best become uninteresting problems. The interesting finding is that subjects remember approximately the same number of chunks under all experimental conditions, but this finding itself constitutes no problem. It only confirms the basic assumption that human memory is limited by the number of "psychological units" (Miller, 1956b, p. 136) that it must handle.

It is difficult to get at these psychological units, since, as Miller points out (1956b), their essential characteristic is that they are imposed by the person who memorizes the material. Although this makes the subjective chunk a more elusive response unit than is, for instance, a word, the concept is not entirely metaphysical. Elsewhere (Tulving, 1962) a method of quantifying the organization that subjects impose on the sequentially unorganized material has been described. The "adopted chunks" defined in this paper, although quite arbitrary as response units, are probably much closer to the "real" subjective units than are words, at least in sequentially organized passages. As a matter of fact, the assumption that adopted chunks are only similar to, but not identical with subjective units, provides us with a convenient rationalization to account for the significant effect of order of approximation on the number of adopted chunks. According to Miller's unitization hypothesis there should be no such effect.

Although one can take the view that the two methods of analysis used in this paper are simply two different ways of looking at data, we believe that the analysis of recall in terms of adopted chunks, supported by the unitization hypothesis, in some sense "explains" the data for number of recalled words. Adopted chunks for H lists and higher orders of approximation contain more words than for HL lists and lower orders of approximation, and this relation is responsible for the apparent differences in recall when we counted individual words. The finding that chunks vary in number of words, however, is no more enigmatic than the fact that words vary in number of letters or phonemes. And if
we are willing to take isolated words of different lengths, say "dog" and "giraffe," as equivalent units (Deese, 1961), we should also be willing to admit chunks as useful units of analysis, even though they differ in "size."

It is perhaps paradoxical that in experiments which specify task variables in terms of properties of items, the subjects' task is usually to learn to recall sequences of items (methods of serial anticipation and paired associates), while in experiments studying the effects on recall of sequential organization of the stimulus list, performance is quite often measured in terms of number of recalled items regardless of order (method of free recall). It is perfectly obvious that in all verbal learning experiments counting words has the advantage of simplicity and objectivity. What may be less obvious is that in some experiments it may also have the disadvantage of complication and delusiveness. If in this experiment we had only counted individual words, regardless of their sequential correspondence to the input lists, we would be faced with the difficult problem of accounting for the observed interaction among our independent variables. The analysis of recall in terms of chunks suggests an entirely different theoretical problem—that of response integration, or of development of chunks. We think that this problem is not only more interesting but will be much easier to answer in terms of existing knowledge. A very large proportion of research in verbal learning is relevant to it, beginning with the early work of Ebbinghaus and ending with much contemporary research. What may be needed in addition to what we already do and know is the realization that subjective units need not correspond to experimenter's units, and that at least as much can be learned from studying subjective organization as from studying variables that influence subject's acquisition of some experimentally imposed organization.

Summary

This experiment was a partial replication and extension of the well-known experiment by Miller and Selfridge (1950). Recall on six successive learning trials was studied as a function of word frequency and contextual constraint. Two different kinds of analysis of recall data—counting recalled words, and counting word sequences adopted by subjects from input lists—yielded different conclusions. With number of words as the response variable, an interaction between word frequency and contextual constraint was found for immediate recall data. These two independent variables also interacted with practice over six trials. With "adopted chunks" as response units, no interaction between word frequency and contextual constraint was found in immediate recall, nor was there any difference in recall attributable to word frequency. Contextual constraint, however, did have a small, but significant, effect.
The results of the experiment were considered with reference to Miller's unitization hypothesis of remembering, which received some support from the data. The importance of reconsidering the usefulness of different kinds of response variables in studies of remembering was also discussed.

REFERENCES


(First received August 16, 1961)