

have a further demonstration of the influence of concurrent load on learning. At this point, we therefore shifted to measures of comprehension based on verification latency rather than retention of information. This has the advantage that the material to be comprehended can be continuously present so that there is no necessary involvement of memory.

Semantic memory and WMG

Sentence verification

We went on to carry out a series of experiments using the technique devised by Collins and Quillan (1969) whereby the subject is presented with a brief sentence describing some commonly known aspect of the world. A typical sentence might be *Canaries have wings*, or some equivalent but obviously false statement such as *Canaries have gills*. The subject is required to decide on the truth of each sentence and press a 'true' or a 'false' key as rapidly as possible.

The task was initially devised to test a specific theory regarding the structural storage of knowledge in semantic memory. This assumed a hierarchy with general concepts such as *LIVING THINGS* being split into more specific concepts such as *ANIMAL* or *PLANT*, some of which themselves split into such subconcepts as *BIRDS*, *FISH*, *TREES*, and *FLOWERS*, which in turn led to particular instances such as *CANARY* or *ROSE*. It was suggested that features that apply to most examples of a category, for example that birds have wings, are stored with that category rather than with each individual instance. This was assumed to lead to economy in storage space, but to have a cost in retrieval time. Verifying that a canary has wings for example involves two steps, verifying that a canary is a bird and then verifying that birds have wings. In contrast, a statement about canaries that is peculiar to canaries, for example *Canaries are yellow*, was said to involve fewer steps and hence, to lead to faster responding.

Our experiments intended to explore the role of WMG in comprehension using a range of examples involving different numbers of hypothetical steps. Unfortunately, however, we, like others failed to replicate the original Collins and Quillan hierarchical effect (see Baddeley 1979; Baddeley and Lewis 1981; Conrad 1972). However, provided one merely treats the sentences as broadly equivalent, then speed and accuracy of verification can be used as a convenient general measure of the efficiency with which subjects can interrogate their knowledge of the world. We have found the test to be highly reliable, and sensitive to a range of stressors including alcohol (Baddeley 1981), high pressure (Logie and Baddeley 1983) and brain damage (Sunderland, Harris, and

Baddeley 1983). This test also appears to provide a plausible example of general semantic processing in so far as our as yet unpublished results indicate that it correlates highly with both the Mill Hill vocabulary test, and with verbal fluency as measured by performance on a task involving generating items from a given semantic category.

Although we have carried out a number of experiments using the Collins and Quillan technique (Baddeley 1979; Baddeley and Lewis 1981), for the present purpose a single experiment will suffice, a study in which the subjects attempted to verify visually presented sentences while holding a load ranging from zero to eight digits spoken at a rate of one per second. We used a mixed design whereby on any given trial, a subject did not know in advance how many digits she would be required to hold. Her task was to listen to the experimenter and repeat whatever sequence she heard, continuing to articulate the sequence until after she had completed the sentence verification response. The sentence was always presented after the subject had begun to repeat the spoken digits. Where no digits were to be repeated, the experimenter said the word 'nothing' whereupon the subject was instructed to remain silent.

Sentences were typed on index cards which were stacked behind a shutter. When the shutter opened a timer was started. The subject's task was to decide whether the sentence was true or false, and press a left or right key accordingly. As soon as she did so, the shutter dropped and the clock stopped, whereupon the experimenter recorded the time and replaced the card. A total of 14 female subjects were tested on a random mixture of 20 sentences at each level of concurrent load.

Performance on this task is shown in Fig. 4.1. Overall, latency increased with concurrent digit load. There is, however, a slight paradox in that the zero load condition is in fact slower than the condition involving the rehearsal of a single item. It is probable that this stems from the fact that the zero load condition was the only one in which the subject was not required to repeat what the experimenter said. Since the subject did not know in advance what condition to expect, and since this occurred on only one occasion in every nine, it seems likely that the need to obey this atypical instruction caused some slight slowing in sentence processing. The most appropriate baseline would, therefore, seem to be the one digit load condition which is equivalent to a condition of articulatory suppression. When this is used, we find that a digit load of even two items is marginally significantly slower than that of one, three items clearly slower and so forth, with each additional digit causing an increase in response latency (for further details see Baddeley, Eldridge, Lewis, and Thomson 1984).

The question arises as to whether increasing digit load produces a continuous or discontinuous function. This has obvious implications for the underlying theory of what produces the effect. It might, for example,

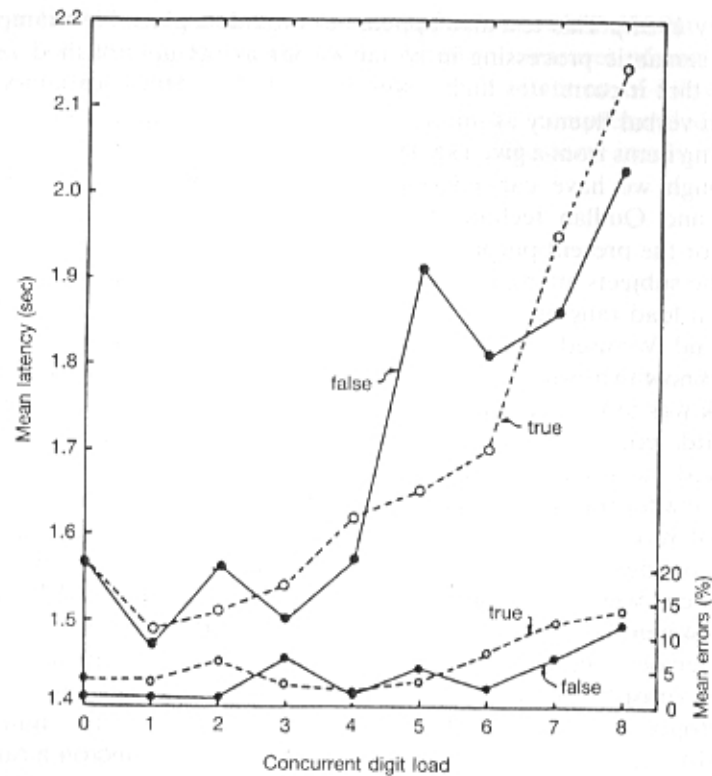


Fig. 4.1. The effects of a concurrent digit load on speed and accuracy of sentence verification (Baddeley, Eldridge, Lewis, and Thomson 1984).

be the case that no effect occurs until some relatively passive store is overloaded, at which point performance begins to deteriorate. The fact that we have obtained either a very small or no effect of two or three digits in earlier studies seemed to point in this direction. However, the more systematic data from this experiment do not support the view of a discontinuous function. It should, however, be borne in mind that such a discontinuity would only become apparent in group data if the assumed passive store had about the same capacity for all subjects. If different subjects had different capacities, then the discontinuity would appear at different points; averaging subjects might well produce what appeared to be a continuous curve. However, in the present study neither the individual nor the group curves show any obvious discontinuity. Individual curves themselves are, of course, based on averages, and as such should be interpreted with some caution. As far as they go, however, they give no apparent support to the discontinuity hypothesis.

This generalization is less obviously true in the case of error data, also shown on Fig. 4.1. Unlike latency, errors are not significantly impaired by the concurrent task until the load approaches six items, while loads of seven or eight items both have a highly reliable effect on error probability. It is, of course, possible that this merely reflects the lower sensitivity of the error measure, based as it is on relatively small numbers of erroneous responses. It does, however, resemble the result obtained from the previous comprehension tasks, both of which used a percentage correct measure, and from the previously described learning studies which again used accuracy rather than latency, and which typically show little or no effect of digit loads of less than six items.

This experiment showed that a concurrent task involving retaining sequences of digits consistently slows down the verification of simple sentences. Although subjects vary as to how steep the function is, the same characteristic effect is shown by all our subjects. These results certainly suggest that some aspect of comprehension is dependent on a limited capacity WMG, but give little evidence as to the more detailed nature of the interference. More specifically, the observed disruption might occur at the level of the processing and comprehension of the sentence, or of the retrieval of the appropriate response, or of the overt execution of that response, or indeed all of these processes might be impaired. We had initially hoped that the use of different types of sentence involving different degrees of the hierarchical processing suggested by Collins and Quillian (1969) would bear directly on this issue. However, since the expected inter-sentence differences did not materialize this was clearly not feasible. The next experiment therefore, selected two phenomena within the area of semantic memory that were known to be replicable, one of which seemed likely to reflect the process of accessing a given concept, while the other was based on the difficulty of the subsequent categorization decision.

Semantic categorization

One of the most commonly used tasks for studying semantic memory involves presenting the subject with a category, for example *animals*, and an instance, for example *horse*, and requiring him to decide whether the instance does or does not belong to that category. Using this procedure, Wilkins (1971) showed that the time to verify such a statement is a function of the instance frequency or saliency of the item within the category. If people are asked to produce as many items as possible from a given category such as *birds*, then items such as *robin* and *sparrow* which are in some sense more salient or typical birds are produced earlier and by more subjects than less typical items such as *ostrich*, *penguin*, or *chicken*. It seems plausible to assume that accessing such atypical items from the category is more demanding than accessing more salient

instances. If this is so, and if the process of moving from category to instance uses some limited capacity WMG system, then one might expect a concurrent load to be particularly disruptive of performance on the more difficult low saliency items. If on the other hand, previously observed concurrent load effect impairs some other component process, such as category access or response selection, then there is no reason to assume an interaction between the effects on latency of instance saliency and of concurrent load.

The second phenomenon to be studied in the experiment concerns the nature of the negative instances. Schaeffer and Wallace (1969) showed that given a category such as *trees*, a negative item from a similar category e.g. *daffodil*, would take longer to reject than an item from a very different category e.g. *kettle*. This presumably reflects the increased difficulty of the decision; trees and flowers have much more in common than trees and kitchen utensils. If this decision process is dependent on a limited capacity WMG, then once again one might expect the difficult negative responses to be more impaired by a concurrent memory load than the easier negatives based on items from categories that are clearly different.

It is perhaps worth pointing out, however, that both this and the previous prediction are made purely on the basis of general plausibility. There is no logical necessity to assume that two operations that place demands on the same system will necessarily lead to an interaction between their effects, nor that the absence of such an interaction necessarily means that the two processes affect different aspects of the system (see Broadbent 1984 and McClelland 1979 for a discussion of this). However, as we shall see in the case of a subsequent reasoning task, interactions between linguistic variables and concurrent load are sometimes observed, and their presence or absence may provide an important cue to the underlying processes, albeit not a completely unequivocal one. However, this issue will be discussed in more detail later.

The task we used therefore, was as follows. The subject was first presented with a category name (e.g. *trees*), which was then followed by an instance (e.g. *pine*). In the case of positive instances, half were taken from the items generated most frequently by subjects in a category generation task (Battig and Montague 1969) and half were from those generated least frequently. In the case of the negative items, half came from categories that were similar to that presented, for example a bird category might be tested with an animal as a negative instance. The remaining half comprised negatives from categories that were clearly different, for example the category animal paired with a kitchen utensil. The procedure involved presenting a sequence of six random digits which the subject was required to continue to repeat until he had completed the category judgement task. This was compared with an articulatory suppression control condition in which the subject was

required to count repeatedly from one to six. The primary comprehension task involved a series of index cards, on each of which was typed a category name and an instance. The card was placed behind a shutter which was raised as the subject began to articulate either the digit load or the counting task. This started a timer which was terminated when the subject pressed a 'yes' or 'no' key with her left or right hand respectively. The experimenter then noted the time, the subject's response and changed the card. A total of 18 female subjects were tested.

The results of the study are shown in Fig. 4.2, from which it is clear that there is an overall effect of concurrent load, as one would expect from the previous study. The two additional variables also had the expected effect, with low saliency items taking consistently longer to verify than high saliency, and similar negative instances taking longer to reject than dissimilar. However, although there is a general tendency for the concurrent load to slow down the negative responses somewhat more than the positive, neither the saliency of positive items nor the similarity of negative items interacts significantly with concurrent load. We thus have no evidence that the limited capacity WMG system is particularly involved in either the item access, or the decision components of the task.

The error scores for the various conditions were also analysed. There were significant effects on error rate of both saliency in the case of positive instances, and of similarity in the case of negatives, but no effect

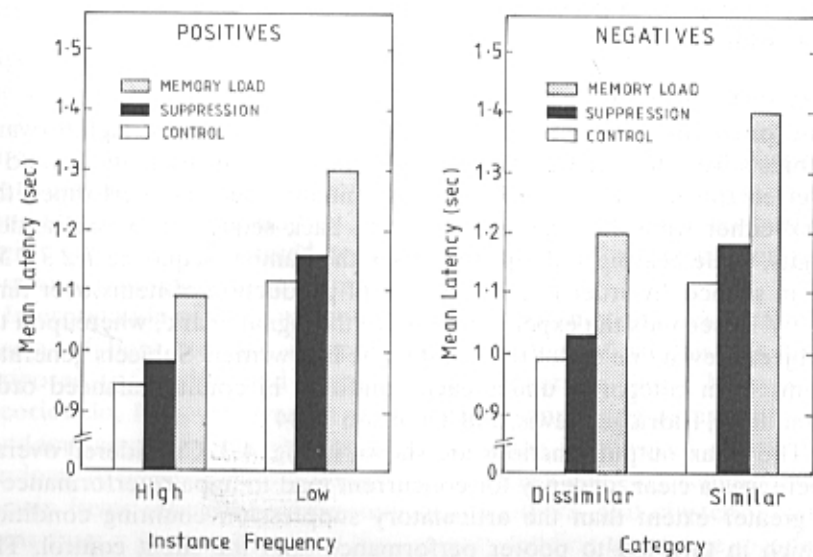


Fig. 4.2. The effects of memory load and articulatory suppression on categorization (Baddeley, Eldridge, Lewis, and Thomson 1984).

of concurrent load. Once again we find that when measured in terms of items correct, concurrent load has no effect on retrieval. When measured in terms of latency, an effect is observed, but this does not appear to interact with the difficulty of the retrieval task. Before going on to discuss this, however, one further study will be described.

Our data suggest unequivocally that a concurrent load will impair comprehension in some way. On the other hand, evidence on the retrieval component of comprehension, like that of the retrieval component in our learning studies suggests that a concurrent task may influence latency but not accuracy. It is as if the retrieval process itself proceeds independently of available WMG capacity, but that WMG is necessary for initiating the response.

It could however be argued that our comprehension data are irrelevant to this question since all the studies we have so far conducted have involved the subject in both processing the input and retrieving the response. It is conceivable that all the effects we have observed stem from the input process, with none of the comprehension effects based on retrieval *per se*. The next experiment explores this possibility by requiring subjects to generate items from specific semantic categories. Since the category name is given in advance, no input processing is involved in this task. Consequently, if the previously observed effect of concurrent load on semantic memory stems purely from input, we would expect no influence of a secondary task on generation performance. On the other hand, if item retrieval is dependent on some aspect of WMG, we would expect subjects to generate fewer items per unit time under concurrent load conditions.

Category generation

The procedure used in this experiment was relatively straightforward; subjects were required to write down as many instances as they could of a given category within a period of 2 minutes. Subjects performed this task either while hearing and repeating back sequences of six random digits, while hearing and repeating back the number sequence 1 2 3 4 5 6, or in silence. In order to plot the rate of production of items over time, every 15 seconds the experimenter gave the signal 'mark', whereupon the subject drew a line under the last item he had written. Subjects generated items from categories under each condition in counterbalanced order (Baddeley, Eldridge, Lewis, and Thomson 1984).

The mean output functions are shown in Fig. 4.3. Considered overall, there was a clear tendency for concurrent load to impair performance to a greater extent than the articulatory suppression counting condition, which in turn led to poorer performance than the silent control. This latter difference should, however, be treated with caution since although suppression appeared to have a clear effect at the beginning of the

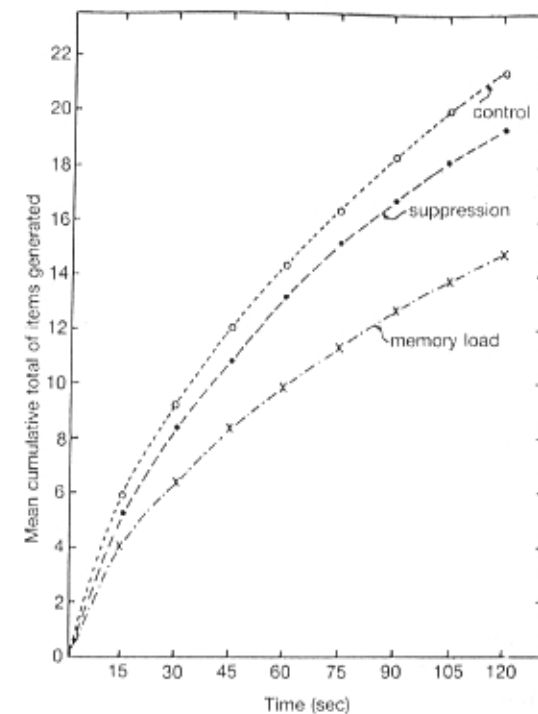


Fig. 4.3. The effects of memory load and articulatory suppression on category generation (Baddeley, Eldridge, Lewis, and Thomson 1984).

experimental session, those subjects tested in the counting condition at the end of the experiment showed no difference between suppression and control. While this result should be treated with circumspection since comparison across groups was involved, nonetheless it suggests that subjects are capable of coping with the problem of simultaneously generating items and counting. This was not the case with the concurrent digit load which continued to lead to very significantly poorer performance throughout the experimental period.

Interpretation of the functions in question depends on one's particular theory of how the task is performed. A range of models has been proposed (c.f. Indow and Togano 1970; Nickerson 1980). Many such theories do, however, tend to involve such simplifying assumptions as random sampling from the category. Sampling is however, clearly not random. A subject's output tends both to reflect category saliency, with certain items emerging consistently before others, and clustering, with items from a given sub-area of the category tending to emerge together, (e.g. *household pets*, *farm animals*, *African animals*, etc.). Both of these phenomena suggest that sampling is non-random.

We therefore, doubt whether any of these models is sufficiently developed to throw useful light on our data. What we can do, however, is to ask the more general question of whether the effect of concurrent load is constant throughout the 2-minute period of generation. It is conceivable for example that the task begins by being dominated by a relatively automatic retrieval process and then switches to the more attention-demanding recollection procedure (Nickerson 1980). If so, one might expect the effect of the concurrent load to be slight during the initial phase, but substantial during the later phase of generation. Since subjects generate about 1.4 items in the control condition for every item they generate under concurrent load, we can test the stability of the concurrent load effect by multiplying each point on the curve by the same amount, 1.4. If item generation switches from a relatively automatic retrieval process to attention-demanding recollection, one might expect an over correction of the early points on the curve and an under-correction of later points, producing a much flatter function. This does not occur, suggesting that the effect of concurrent load is relatively constant throughout. It would be of interest to explore this point further by manipulating the nature of the generation task. It seems likely that some tasks such as generating palindromes, words like *nun* or *minim* which are identical whether spelled forward or backwards, would place much heavier demands on the active recollective process and would therefore be more susceptible to the influence of a concurrent load. We have not, however, so far explored this possibility.

Reasoning and WMG

We have so far shown a clear effect of concurrent load on a range of learning tasks, and on comprehension, at least when measured in terms of latency. Our results for retrieval are less straightforward, and we have suggested that there may well be at least two components, an automatic search process, together with a more active recollective component that has some resemblance to problem-solving, and which is dependent on the operation of a limited capacity WMG. The last two experiments to be described in this section explore this possibility further by studying a syntactic verification task selected so as to place relatively heavy demands on the subject's reasoning capacity. The task in question is the grammatical reasoning task devised as a means of presenting subjects with a large number of simple and overlearned reasoning questions in a short period of time (Baddeley 1968*b*). It takes advantage of the considerable psycholinguistic literature suggesting that subjects are faster at processing affirmative than negative sentences and faster at processing actives than passives (Wason and Johnson-Laird 1972). The test correlates highly with intelligence, is highly reliable and very sensitive to

a range of stresses (Baddeley 1968*b*). It gives rise to consistent effects of both active vs passive voice and positive vs negative response, together with the well established interaction between these, and although subjects show consistent learning on this task, the learning function is relatively linear with the result that a test in which subjects perform the task for only one minute is sufficient to give a good and stable measure of their reasoning performance (Carter, Kennedy, and Bittner 1981).

We have carried out a number of experiments on the influence of concurrent load on this task (Baddeley and Hitch 1974; Hitch and Baddeley 1976), but will confine discussion here to two studies. The first of these (Baddeley and Hitch 1974 Experiment III) used the concurrent digit load task, comparing the effect of a concurrent load of six random digits with an articulatory suppression condition involving counting repeatedly from one to six, and a silent control. The digits whether random or sequential, were presented auditorily and the reasoning task visually using the shutter device described previously. Subjects were required to continue to rehearse the random digits while verifying a range of sentences claiming to describe the order of two subsequent letters, A and B. The sentences ranged from simple activity declaratives such as *A follows B—BA* to which the subject should respond by pressing the 'true' button through passives such as *A is preceded by B—AB* (false), negatives such as *B does not follow A—BA* (true) and combined negative passives such as *A is not preceded by B—AB* (true).

Figure 4.4 shows the mean time taken by our subjects to verify sentences of these various types as a function of concurrent task. The results indicate first of all that the nature of the sentence has a clear influence on verification latency, secondly, that latency is consistently slowed down by the concurrent digit task, with subjects being significantly slower in this condition than when suppressing articulation, which is in turn marginally significantly slower than the silent control. Finally, we observed an interaction between sentence type and concurrent load, with the more difficult sentences showing a greater decrement than the simpler. These results are clearly consistent with the concept of a WMG which is of limited capacity. Taking up some of that capacity by means of a secondary task impairs reasoning performance, with the effect being particularly great for the more demanding sentences.

The final experiment in this sequence is one in which we explored the reasoning effect in rather more depth by studying seven female subjects over five days, and systematically varying the load from zero to eight digits on each day. The design used was essentially that employed in the previous study of semantic processing under zero to eight concurrent digits. The major difference was that we were interested in collecting sufficient observations to draw reliable conclusions about each of the four types of sentence. In addition, the longer period of testing allowed

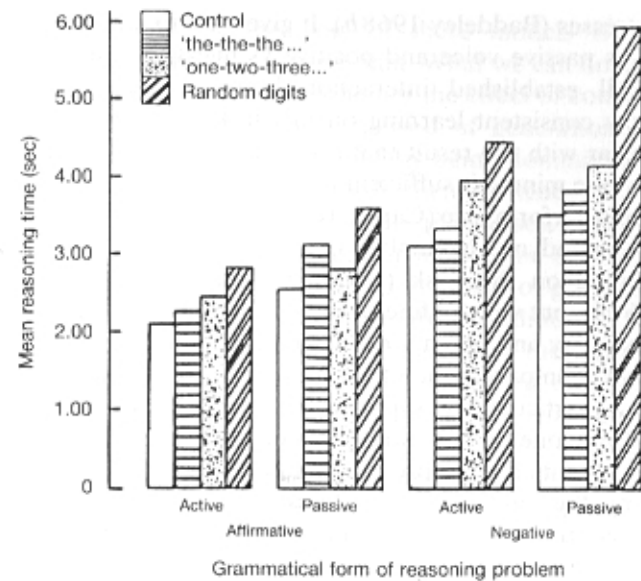


Fig. 4.4. The effects of memory load and articulatory suppression on speed of grammatical reasoning (Baddeley and Hitch 1974).

us to study the effects of practice and also to obtain a clearer indication of individual differences and performance on this task.

Figure 4.5 shows the overall effect of concurrent load on verification time. It is clear from this, that the effect of concurrent load on RT is replicated. We observed a very substantial practice effect over days, but this did not remove or interact with the influence of concurrent load. A simple index of this is the percentage increment to the verification time under control conditions produced by the maximum load of eight digits. If we calculate this over the five successive days, mean increment is 32, 30, 41, and 41 per cent respectively. In short, although our subjects are improving very substantially on this task, they do not appear to be escaping the deleterious effect of a concurrent digit load.

In one respect however, our results differ from those of the previous study; although there are clear and highly significant effects of the active/passive and positive/negative variables, these did not interact with concurrent load. It is far from clear why this should be, or indeed what are the implications of interaction effects, or their absence. Specifying such implications demands a more detailed model of the underlying process than has so far been proposed (c.f. Broadbent 1984; McClelland 1979).

Figure 4.5 also displays the error data for the sentence verification task. It is noteworthy that subjects appear to be able to maintain a

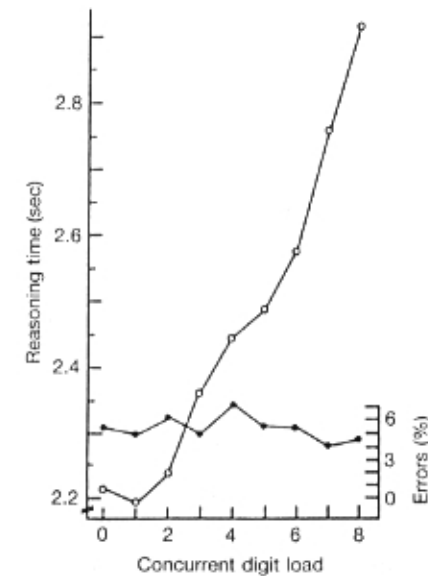


Fig. 4.5. Speed and accuracy of grammatical reasoning as a function of concurrent digit load (Baddeley and Lewis unpublished).

constant error rate of about 5 per cent regardless of concurrent digit load. This is fortunate since it suggests that subjects are continuing to treat the sentence verification task in a consistent way, and are not trading off reasoning accuracy in order to perform the digit task. Verification times for erroneous responses tended to be slower than for correct ones, with an overall mean latency of 3.99 seconds. However, those erroneous responses on which a digit error also occurred did not appear to be substantially different from those in which digit recall was perfect.

Analysis of the secondary digit task indicated that subjects were succeeding remarkably well in carrying out the concurrent task, with a substantial error rate occurring only in the eight digit load condition (See Table 4.1). Since five of our seven subjects had a span of eight or less, this represents virtually no decrement in memory as a result of the concurrent reasoning task. Even allowing for some improvement in span over the five days of practice, this is a remarkable result creating a major problem for any concept of working memory relying on the modal model, which would presumably predict that a difficult reasoning task performed with a 95 per cent accuracy should have a disastrous effect on a subject's concurrent ability to remember a digit sequence equivalent to her span.

Table 4.1 also gives the mean verification time for those sentences that were processed correctly, but associated with an error on the concurrent

no evidence of a phonological similarity effect when they are required to suppress articulation. This is an important finding, which has been replicated several times; Levy (1971), Estes (1973), and Peterson and Johnson (1971) all observed that the phonological similarity effect vanishes when items are presented visually and the subject prevented from subvocal rehearsal by means of an articulatory suppression task.

The word length effect

At about the time that we were developing the initial working memory model, we discovered another source of evidence for speech coding in STM, namely the word length effect (Baddeley, Thomson, and Buchanan 1975). Subjects who are asked to repeat back sequences of words are much more likely to do so correctly if the words are short such as *sum*, *wit*, *hate*, than if they are required to remember long words of equal frequency in the language, such as *opportunity*, *university*, *aluminium*. We were at first concerned that this might reflect linguistic rather than durational differences between the two sets, since our short words tended to be relatively simple and probably of largely Anglo-Saxon origin, in contrast to the more complex Latinate long words we had used. We therefore repeated the experiment using country names, a type of material which one would not expect to differ systematically between long and short items. Hence, we might compare memory for a sequence such as *Chad*, *Burma*, *Greece*, *Cuba*, *Malta* with a sequence such as *Czechoslovakia*, *Somaliland*, *Nicaragua*, *Afghanistan*, *Yugoslavia*. Subjects averaged 4.17 words out of 5 for the short names and 2.80 for the long names, a highly significant difference.

A subsequent study attempted to tease out whether the crucial factor was the duration of the spoken word, or its length in terms of syllables. Either might be plausible. A trace decay hypothesis might suggest that duration is important since longer words take longer to say and hence, allow the memory trace to be refreshed less frequently, leading to more forgetting. An interference theory or displacement model might argue that the crucial feature is number of syllables. A number of theorists have argued that one component of STM is a system containing a limited number of slots or memory locations; when the number of items to be remembered exceeds this number, forgetting occurs. If each slot were to hold a fixed number of syllables, then polysyllabic words would overload the system more rapidly than monosyllables.

In order to decide between these two hypotheses, we compared immediate memory for two sets of words. Both comprised two syllables, but in one set, spoken duration was relatively brief e.g. *bishop*, *wicket*, while for the other group spoken duration was relatively long, e.g. *harpoon*, *Friday*. We observed a consistent tendency for subjects to do better on the short-duration words, despite the fact that the two sets were matched not only

for number of syllables, but in a second study also for number of phonemes.

A very neat extension of this result was produced by Ellis and Hennessey (1980). They had noticed that the Welsh language version of the Wechsler intelligence test contained a puzzling anomaly, namely that the digit span norms for Welsh children were quite substantially below those for the original American norms. Ellis and Hennessey point out that Welsh digits, although containing the same number of syllables as English, nevertheless tend to contain longer vowel sounds and hence, take longer to articulate. They therefore explored the hypothesis that the observed difference in norms could be due to this difference in spoken duration. They used as their subjects bilingual students for whom Welsh was their first language. They showed first of all that these subjects had a longer digit span in English than in Welsh. They then measured the time taken by their subjects to articulate digits in English and then in Welsh, observing as anticipated that producing the Welsh digits took longer. When this information was used to measure the memory span of their subjects in the two languages in terms of spoken duration, then their English and Welsh spans became equal. Finally, when subvocalization was prevented by articulatory suppression, the difference between the digit span of their subjects in English and Welsh again disappeared. It appears to be the case that Welsh digit span norms are inferior to US norms simply because Welsh digits take longer to articulate. Preliminary evidence suggests that a comparable state of affairs may occur quite commonly in languages such as Italian, where most digits involve at least two syllables.

The fact that word duration is a crucial variable in memory span is consistent with a trace decay hypothesis rather than the interpretation in terms of a limited number of storage slots. Let us suppose that the presentation of an item leaves a memory trace which decays over time. Re-presentation of an item either by the experimenter, or by the subject himself rehearsing the item will refresh the trace and arrest the process of decay. Amount retained will therefore be a joint function of decay rate and rehearsal rate. With very few items, the subject will be able to rehearse the complete sequence in less time than it takes the memory trace to decay, hence allowing him to maintain the sequence indefinitely (Vallar and Baddeley 1982). As the length of the sequence increases, so will the time needed by the subject to rehearse the entire sequence, until a point is reached at which the decay time for an individual item is less than the time to rehearse the total sequence. When this happens, errors will begin to occur.

Looked at in this way, it is possible to express the memory span in terms of either the number of items or the total spoken duration. Whereas a span measured in terms of numbers of item is likely to depend

on the duration of those items, the total recall of time should be constant. We tested this hypothesis using the long- and short-duration disyllabic words described previously. We measured articulation rate in two ways, by presenting the words as typed lists and requiring the subject to read them as rapidly as possible, and by presenting subjects with groups of three words that they were required to articulate a total of ten times as rapidly as possible. Reading and articulation rate were then used to convert the observed memory spans into temporal spans. The results suggested that a subject's span for long words was equivalent to the amount he could read in 1.62 seconds for long words and 1.67 seconds for short words, a difference which did not approach significance. When measured in terms of articulation rate, long words gave a span of 1.33 seconds of articulation and short a span of 1.31 seconds, again indicating no difference as a function of word length.

In order to explore the temporal duration hypothesis over a wider range, we carried out an experiment in which subjects attempted to recall sequences of words of 1, 2, 3, 4, or 5 syllables. The words were matched for frequency and category membership, the set being shown in Table 5.1. Subjects were tested at each length on sequences of five words, presented visually at a rate of 2 seconds per word. Reading rate for each set of words was determined by requiring the subjects to read lists of 50 words, each of which comprised five occurrences of each item in the relevant set. Subjects read each list four times and were timed by stopwatch.

TABLE 5.1
Set of words used in Experiment II

Set No.					
1.	mumps	measles	leprosy	diphtheria	tuberculosis
2.	stoat	puma	gorilla	rhinoceros	hippopotamus
3.	Greece	Peru	Mexico	Australia	Yugoslavia
4.	Maine	Utah	Wyoming	Alabama	Louisiana
5.	zinc	carbon	calcium	uranium	aluminium

Figure 5.1 shows the mean recall and reading rate scores for the various sets of words as a function of length. As expected, word length has a very pronounced effect on both reading rate and memory, with each condition being significantly different from each other. A more crucial issue, however, is the question of the relationship between reading rate and span. It will be recalled that the decay hypothesis predicts that this should be constant. Figure 5.1 shows the result observed, a linear relationship between reading rate and recall that is described by the equation $S = c + kR$, where S is the memory span, R is reading rate and k

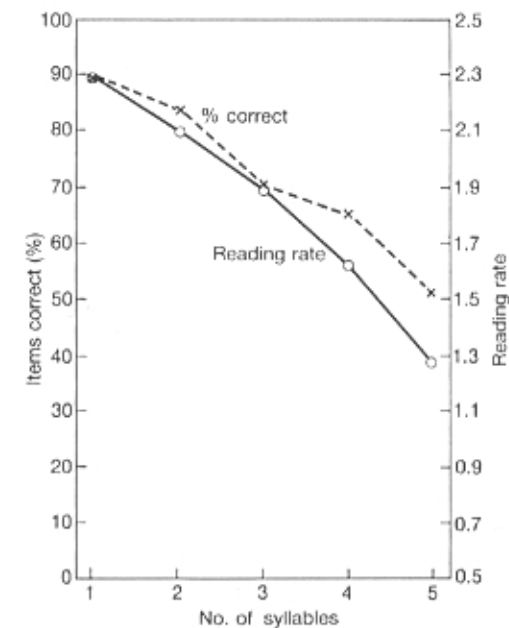


Fig. 5.1. Mean reading rate and percentage correct recall of sequences of five words as a function of length.

and c are constants. This suggests that the relationship predicted by a simple decay hypothesis does indeed hold for the group results. We next explored the question of whether it also held for differences between individuals; in short, are fast articulators good rememberers? Our data suggest that they are since we obtained a substantial correlation between memory span and reading rate ($R = 0.685$, $p < 0.005$).

The evidence so far then hangs together very neatly in suggesting some form of phonological store that relies on a fading trace which can be maintained by subvocal rehearsal. The question then arises as to whether the phonological store implied by the word length effect is the same as that suggested by the phonological similarity effect. If this is the case, then one might expect that the influence of articulatory suppression on the word length effect would be just as drastic as its effect on phonological similarity. We therefore carried out a further experiment in which subjects were presented visually with sequences of five words from either a short word set comprising monosyllables, or a set comprising words of five syllables. Under one set of conditions, subjects were left free to rehearse while in the other, rehearsal was prevented by requiring the subject to count repeatedly from one to six. The results were clear; the standard word length effect was present under control conditions but was abolished by articulatory suppression.

So far then, all the data seemed to fit into the simple concept of a time-based loop, somewhat analogous to a closed loop on a tape recorder. The loop was assumed to be based on articulation and to consist of a bank of articulatory programmes which were able to feed the process of articulation. This in turn refreshed the stored articulatory programmes and prevented them from fading. Such a simple device seemed to be capable of handling all the results we have so far described. The phonological similarity effect was assumed to be due to confusion among articulatory programmes, with items involving similar sounds also involving similar articulatory patterns. Murray's observation that the more intense the degree of articulation the better the recall also fitted, given the assumption that more intensive articulation laid down clearer or stronger articulatory traces. The effect of articulatory suppression fitted in neatly since a set of irrelevant speech sounds would tend to preempt the articulatory system, disrupting its use for short-term storage, while the evidence for phonological confusions in congenitally deaf children who are good articulators also fits in particularly neatly with such a view. The first articulatory loop model, therefore, was essentially a simple tape loop analogy.

Problems with the initial articulatory loop model

Although the initial model appeared to offer a simple explanation of a wide range of data, there remained a cluster of results that did not fit this model. Although the phonological similarity effect is disrupted by articulatory suppression, this is only the case when material is presented visually. With auditory presentation, the phonological similarity effect withstands suppression. In a final experiment in our initial word length paper (Baddeley, Thomson, and Buchanan 1975), we observed a similar effect for word length; although articulatory suppression removed the word length effect with visual presentation, when material was presented auditorily a clear word length effect was observed even when subjects were suppressing articulation. If articulatory suppression pre-empts the articulatory loop, then it should abolish the phonological similarity and word length effects regardless of whether presentation is visual or auditory. The next section describes how this apparent anomaly, together with data from other paradigms led to a modification and elaboration of the initial articulatory loop model.

In a previous study concerned with rate of presentation we had observed a difference in the effect of articulatory suppression on performance depending on whether suppression occurred only during input or during both input and recall (Baddeley and Lewis 1984). It appears to be the case that with moderately rapid presentation, subjects are able to use rehearsal during the recall period to maintain items of memory. It seemed possible at least that the word length effect observed under

auditory presentation and suppression might have been dependent on this. While this may not seem particularly likely on *a priori* grounds the issue was sufficiently important theoretically to merit exploring this possibility.

Giuseppe Vallar and I therefore carried out an experiment in which subjects heard and attempted to recall sequences of five words comprising either one syllable or five syllables each (Baddeley, Lewis, and Vallar 1984, Experiment 4). Subjects were either free to rehearse, or required to suppress articulation by counting rapidly from one to eight throughout both input and written recall. In order to equate writing time, we allowed subjects to abbreviate items to their first three letters. The results of the study are shown in Fig. 5.2. The highly significant word length effect that occurred under control conditions was reduced to a nonsignificant trend under suppression, a difference that was reflected by a significant interaction between word length and suppression.

In view of the importance of this result to the Articulatory Loop model, it clearly merited replication. An experiment by Muriel Woodhead and myself repeated the study using sequences of 4, 5, or 6 words, again drawn from sets comprising words of one or five syllables (Baddeley, Lewis, and Vallar 1984, Experiment 5). The results were virtually identical with the previous study, with a significant word length effect under control conditions, a significant interaction between word length and suppression, but no significant effect of word length when subjects suppressed articulation during both input and recall. There was, however, in both studies a slight trend in the direction of a word length effect under suppression, which though not significant, suggests that it would be unwise to conclude that suppression completely abolishes the

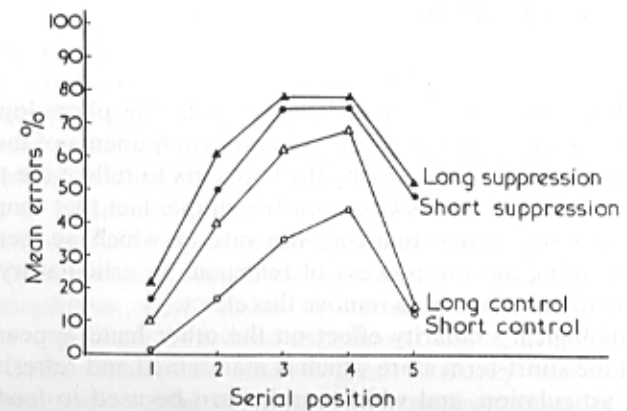


Fig. 5.2. The influence of word length and articulatory suppression on immediate memory for auditorily presented words. Suppression occurred during both presentation and recall (Baddeley, Lewis, and Vallar 1984).

effect of word length. However, even if the effect did reach significance, it might simply reflect a tendency for subjects to intersperse an occasional subvocal rehearsal between suppression responses.

In general however, these results suggest that the word length does indeed stem from the process of articulatory rehearsal *per se*; long words take longer to rehearse and allow more forgetting. If rehearsal is prevented, then word length ceases to be an important variable (Baddeley, Lewis, and Vallar 1984). The result does, however, raise two further questions. First, why is suppression during input enough to abolish the word length effect with visual presentation? Secondly, what happens to the phonological similarity effect when items are presented auditorily and suppression occurs throughout input and recall? If the phonological similarity effect manifests itself through the process of articulatory rehearsal, one might expect it to be abolished by suppression provided this occurs throughout the input and recall. If on the other hand, the phonological similarity effect operates through some other component of the system and is not dependent on articulation for its occurrence, then one might expect it to survive the effects of continuous articulatory suppression. Two experiments exploring this point will be described before going on to reconsider the Articulatory Loop hypothesis.

We carried out a total of three experiments in which material was presented auditorily, and the influence of phonological similarity was studied with and without articulatory suppression (Baddeley, Lewis, and Vallar 1984, Experiments 1-3). The results of all three were consistent in showing that similarity continues to have a marked effect whether suppression is at input only or at both input and recall. This results contrasts with the case of visual presentation, where suppression eliminates the phonological similarity effect (Estes 1973; Levy 1971; Murray 1968; Peterson and Johnson 1971).

A revised articulatory loop model

The available evidence, therefore, suggests that the phonological similarity and word length effects reflect different components of the articulatory loop system. The word length effect appears to reflect the process of articulatory rehearsal, and is based on the simple fact that longer words take longer to say, hence reducing the rate at which an item can be rehearsed. Cutting out the process of rehearsal by articulatory suppression appears to be sufficient to remove this effect.

The phonological similarity effect on the other hand appears to be a function of the short-term store which is maintained and refreshed by the process of articulation, and which can in turn be used to feed the articulatory process. This store appears to be accessible either through auditory presentation or by the articulatory coding of visually presented material. However, although auditory presentation seems to provide

direct access to this phonological store, it does not appear to guarantee rehearsal. If it did, then overall level of performance should not be impaired by articulatory suppression, provided materials were presented auditorily. As we have seen, decrement does occur under these conditions (e.g. Baddeley and Lewis 1984).

We are left then with a very simple system comprising a phonological store and an articulatory control process. The sections that follow describe the application of this simple model to a number of further issues. These cause some additional development, and suggest other components of the working memory system, but on the whole the revised articulatory loop model appears to offer an adequate and useful account of the available evidence.

The inner voice and the inner ear

While the articulatory loop component of working memory is probably the one that has been most extensively explored, we are left with at least one major loose end. It may be recalled that articulatory suppression impairs the retention of visually presented words, eliminates the phonological similarity and word-length effects and obliterates the influence of unattended speech. All these results suggest that articulatory suppression prevents visually presented items being phonologically encoded.

On the other hand, subjects appear to be quite capable of making judgements of rhyme or of homophony on visually presented words while suppressing articulation. Indeed not only are subjects capable of continuing to perform this task, they do so with little impairment in speed or accuracy (Baddeley and Lewis 1981; Besner *et al.* 1981). Further evidence that articulatory suppression does not prevent phonological coding comes from a study by Besner and Davelaar (1982). This involved two experiments, one concerning phonological similarity and the other word length. In the first experiment subjects were tested for immediate serial recall of sequences of items that were either pseudohomophones (e.g. *phood*, *newd*, *chood*) or were non-homophones (e.g. *thude*, *snude*, *beued*). Furthermore, the sequences were either phonologically similar or dissimilar, and were remembered under either control conditions or under articulatory suppression. As would be expected, a phonological similarity effect occurred under control conditions but was abolished by articulatory suppression. On the other hand, items that were pseudohomophones were easier to recall under both control and suppression conditions. A second experiment was equivalent except that length of item, rather than phonological similarity was manipulated. Again, presentation was visual and subjects performed under control or suppression conditions. Pseudohomophones were easier than non-homophones under both conditions, whereas the effect of item length was present only in the control condition. Besner and Davelaar conclude that