BEHAVIORAL AND BRAIN SCIENCES (2000) **24,** 87–185
Printed in the United States of America

The magical number 4 in short-term memory: A reconsideration of mental storage capacity

Nelson Cowan
Department of Psychological Sciences,
University of Missouri, Columbia, MO 65211
cowanN@missouri.edu **www.missouri.edu/~psycowan**

Abstract: Miller (1956) summarized evidence that people can remember about seven chunks in short-term memory (STM) tasks. However, that number was meant more as a rough estimate and a rhetorical device than as a real capacity limit. Others have since suggested that there is a more precise capacity limit, but that it is only three to five chunks. The present target article brings together a wide variety of data on capacity limits suggesting that the smaller capacity limit is real. Capacity limits will be useful in analyses of information processing only if the *boundary conditions for observing them* can be carefully described. Four basic conditions in which chunks can be identified and capacity limits can accordingly be observed are: (1) when information overload limits chunks to individual stimulus items, (2) when other steps are taken specifically to block the recoding of stimulus items into larger chunks, (3) in performance discontinuities caused by the capacity limit, and (4) in various indirect effects of the capacity limit. Under these conditions, rehearsal and long-term memory cannot be used to combine stimulus items into chunks of an unknown size; nor can storage mechanisms that are not capacity-limited, such as sensory memory, allow the capacity-limited storage mechanism to be refilled during recall. A single, central capacity limit averaging about four chunks is implicated along with other, noncapacity-limited sources. The *pure STM capacity limit* expressed in chunks is distinguished from *compound STM limits* obtained when the number of separately held chunks is unclear. Reasons why pure capacity estimates fall within a narrow range are discussed and a capacity limit for the focus of attention is proposed.

Keywords: attention; enumeration; information chunks; memory capacity; processing capacity; processing channels; serial recall; short-term memory; storage capacity; verbal recall; working memory capacity

1. Introduction to the problem of mental storage capacity

One of the central contributions of cognitive psychology has been to explore limitations in the human capacity to store and process information. Although the distinction between a limited-capacity primary memory and an unlimited-capacity secondary memory was described by James (1890), Miller's (1956) theoretical review of a "magical number seven, plus or minus two" is probably the most seminal paper in the literature for investigations of limits in short-term memory (STM) storage capacity. It was, in fact, heralded as one of the most influential *Psychological Review* papers ever, in a 1994 centennial issue of the journal. Miller's reference to a magical number, however, was probably a rhetorical device. A more central focus of his article was the ability to increase the effective storage capacity through the use of intelligent grouping or "chunking" of items. He ultimately suggested that the specific limit of seven probably emerged as a coincidence.

Over 40 years later, we are still uncertain as to the nature of storage capacity limits. According to some current theories there is no limit in storage capacity per se, but a limit in the duration for which an item can remain active in STM without rehearsal (e.g., Baddeley 1986; Richman et al. 1995). This has led to a debate about whether the limitation is a "magic number or magic spell" (Schweickert & Boruff 1986) or whether rehearsal really plays a role (Brown & Hulme 1995). One possible resolution is that the focus of attention is capacity-limited, whereas various supplementary storage mechanisms, which can persist temporarily without attention, are time-limited rather than capacity-limited (Cowan 1988; 1995). Other investigators, however, have long questioned whether temporary storage concepts are necessary at all, and have suggested that the rules of learning and memory could be identical in both the short and long term (Crowder 1993; McGeoch 1932; Melton 1963; Nairne 1992; Neath 1998).

At present, the basis for believing that there is a time

NELSON COWAN (Ph.D. 1980, University of Wisconsin–Madison) is Professor in the Department of Psychological Sciences at the University of Missouri–Columbia. He has written one book (Cowan, N. 1995. *Attention and memory: An integrated framework.* Oxford University Press) and edited another (1997. *The development of memory in childhood.* Psychology Press), and has 100 other publications on working memory, and its relation to attention. He is former Associate Editor of the *Journal of Experimental Psychology: Learning, Memory, and Cognition* (1995–1999) and current Associate Editor of the *Quarterly Journal of Experimental Psychology (section A).* He won the 1998 University of Missouri Chancellor's Award for Research and Creative Activities.

 0140-525X/00 $12.50

limit to STM is controversial and unsettled (Cowan et al. 1997a; 1997b; Crowder 1993; Neath & Nairne 1995; Service 1998). The question is nearly intractable because any putative effect of the passage of time on memory for a particular stimulus could instead be explained by a combination of various types of proactive and retroactive interference from other stimuli. In any particular situation, what looks like decay could instead be displacement of items from a limited-capacity store over time. If the general question of whether there is a specialized STM mechanism is to be answered in the near future then, given the apparent unresolvability of the decay issue, general STM questions seem more likely to hinge on evidence for or against a chunk-based capacity limit.

The evidence regarding this capacity limit also has been controversial. According to one view (Wickens 1984) there is not a single capacity limit, but several specialized capacity limits. Meyer and Kieras (1997) questioned the need for capacity limits to explain cognitive task performance; they instead proposed that performance scheduling concerns (and the need to carry out tasks in the required order) account for apparent capacity limits. The goal of this target article is to provide a coherent account of the evidence to date on storage capacity limits .

One reason why a resolution may be needed is that, as mentioned above, the theoretical manifesto announcing the existence of a capacity limit (Miller 1956) did so with considerable ambivalence toward the hypothesis. Although Miller's ambivalence was, at the time, a sophisticated and cautious response to available evidence, a wealth of subsequent information suggests that there is a relatively constant limit in the number of items that can be stored in a wide variety of tasks; but that limit is only three to five items as the population average. Henderson (1972, p. 486) cited various studies on the recall of spatial locations or of items in those locations, conducted by Sperling (1960), Sanders (1968), Posner (1969), and Scarborough (1971), to make the point that there is a "new magic number 4 ± 1." Broadbent (1975) proposed a similar limit of three items on the basis of more varied sources of information including, for example, studies showing that people form clusters of no more than three or four items in recall. A similar limit in capacity was discussed, with various theoretical interpretations, by others such as Halford et al. (1988; 1998), Luck and Vogel (1997), and Schneider and Detweiler (1987).

The capacity limit is open to considerable differences of opinion and interpretation. The basis of the controversy concerns the way in which empirical results should be mapped onto theoretical constructs. Those who believe in something like a 4-chunk limit acknowledge that it can be observed only in carefully constrained circumstances. In many other circumstances, processing strategies can increase the amount that can be recalled. The limit can presumably be predicted only after it is clear how to identify independent chunks of information. Thus, Broadbent (1975, p. 4) suggested that "The traditional seven arises . . . from a particular opportunity provided in the memory span task for the retrieval of information from different forms of processing."

The evidence provides broad support for what can be interpreted as a capacity limit of substantially fewer than Miller's 7 ± 2 chunks; about four chunks on the average. Against this 4-chunk thesis, one can delineate at least seven commonly held opposing views: (View 1) There do exist capacity limits but they are in line with Miller's 7 ± 2 (e.g., still taken at face value by Lisman & Idiart 1995). (View 2) Short-term memory is limited by the amount of time that has elapsed rather than by the number of items that can be held simultaneously (e.g., Baddeley 1986). (View 3) There is no special short-term memory faculty at all; all memory results obey the same rules of mutual interference, distinctiveness, and so on (e.g., Crowder 1993). (View 4) There may be no capacity limits per se but only constraints such as scheduling conflicts in performance and strategies for dealing with them (e.g., Meyer & Kieras 1997). (View 5) There are multiple, separate capacity limits for different types of material (e.g., Wickens 1984). (View 6) There are separate capacity limits for storage versus processing (Daneman & Carpenter 1980; Halford et al. 1998). (View 7) Capacity limits exist, but they are completely task-specific, with no way to extract a general estimate. (This may be the "default" view today.) Even among those who agree with the 4-chunk thesis, moreover, a remaining possible ground of contention concerns whether all of the various phenomena that I will discuss are legitimate examples of this capacity limit.

These seven competing views will be re-evaluated in section 4 (sects. 4.3.1–4.3.7). The importance of identifying the chunk limit in capacity is not only to know what that limit is, but more fundamentally to know whether there is such a limit at all. Without evidence that a consistent limit exists, the concepts of chunking and capacity limits are themselves open to question.

1.1. Pure capacity-based and compound STM estimates

I will call the maximum number of chunks that can be recalled in a particular situation the *memory storage capacity*, and valid, empirically obtained estimates of this number of chunks will be called *estimates of capacity-based* STM. Although that chunk limit presumably always exists, it is sometimes not feasible to identify the chunks inasmuch as long-term memory information can be used to create larger chunks out of smaller ones (Miller 1956), and inasmuch as time- and interference-limited sources of information that are not strictly capacity-limited may be used along with capacity-limited storage to recall information. In various situations, the amounts that can be recalled when the chunks cannot be specified, or when the contribution of non-capacity-limited mechanisms cannot be assessed, will be termed *compound STM estimates.* These presumably play an important role in real-world tasks such as problem-solving and comprehension (Daneman & Merikle 1996; Logie et al. 1994; Toms et al. 1993). However, the theoretical understanding of STM can come only from knowledge of the basic mechanisms contributing to the compound estimates, including the underlying capacity limit. The challenge is to find sound grounds upon which to identify the pure capacity-based limit as opposed to compound STM limits.

1.2. Specific conditions in which a pure storage capacity limit can be observed

It is proposed here that there are at least four ways in which pure capacity limits might be observed: (1) when there is an information overload that limits chunks to individual stimulus items, (2) when other steps are taken specifically to

block the recoding of stimulus items into larger chunks, (3) when performance discontinuities caused by the capacity limit are examined, and (4) when various indirect effects of the capacity limit are examined. Multiple procedures fit under each of these headings. For each of them, the central assumption is that the procedure does not enable subjects to group items into higher-order chunks. Moreover, the items must be familiar units with no pre-existing associations that could lead to the encoding of multi-object groups, ensuring that each item is one chunk in memory. Such assumptions are strengthened by an observed consistency among results.

The first way to observe clearly limited-capacity storage is to overload the processing system at the time that the stimuli are presented, so that there is more information in auxiliary or time-limited stores than the subject can rehearse or encode before the time limit is up. This can be accomplished by presenting a large spatial array of stimuli (e.g., Sperling 1960) or by directing attention away from the stimuli at the time of their presentation (Cowan et al. 1999). Such manipulations make it impossible during the presentation of stimuli to engage in rehearsal or form new chunks (by combining items and by using long-term memory information), so that the chunks to be transferred to the limited-capacity store at the time of the test cue are the original items presented.

The second way is with experimental conditions designed to limit the long-term memory and rehearsal processes. For example, using the same items over and over on each trial and requiring the recall of serial order limits subjects' ability to think of ways to memorize the stimuli (Cowan 1995); and rehearsal can be blocked through the requirement that the subject repeat a single word over and over during the stimulus presentation (Baddeley 1986).

The third way is to focus on abrupt changes or discontinuities in basic indices of performance (proportion correct and reaction time) as a function of the number of chunks in the stimulus. Performance on various tasks takes longer and is more error prone when it involves a transfer of information from time-limited buffers, or from long-term memory, to the capacity-limited store than when it relies on the contents of capacity-limited storage directly. This results in markedly less accurate and/or slower performance when more than four items must be held than when fewer items must be held (e.g., in enumeration tasks such as that discussed by Mandler & Shebo 1982).

Fourth, unlike the previous methods, which have involved an examination of the level of performance in the memory task, there also are indirect effects of the limit in capacity. For example, lists of items tend to be grouped by subjects into chunks of about four items for recall (Broadbent 1975; Graesser & Mandler 1978), and the semantic priming of one word by another word or learning of contingencies between the words appears to be much more potent if the prime and target are separated by about three or fewer words (e.g., McKone 1995).

1.2.1. Other restrictions on the evidence. Although these four methods can prevent subjects from amalgamating stimuli into higher-order chunks, the resulting capacity estimates can be valid only if the items themselves reflect individual chunks, with strong intra-chunk associations and weak or (ideally) absent inter-chunk associations. For example, studies with nonsense words as stimuli must be excluded because, in the absence of pre-existing knowledge of the novel stimulus words, each word may be encoded as multiple phonemic or syllabic subunits with only weak associations between these subunits (resulting in an underestimate of capacity). As another example, sets of dots forming familiar or symmetrical patterns would be excluded for the opposite reason, that multiple dots could be perceived together as a larger object with non-negligible inter-dot associations, so that each dot would not be a separate chunk (resulting in an overestimate of capacity). It also is necessary to exclude procedures in which the central capacity's contents can be recalled and the capacity then re-used (e.g., if a visual array remains visible during recall) or, conversely, in which the information is not available long enough or clearly enough for the capacity to be filled even once (e.g., brief presentation with a mask). In section 3, converging types of evidence will be offered as to the absence of inter-item chunking in particular experimental procedures (e.g., a fixed number of items correctly recalled regardless of the list or array size).

Finally, it is necessary to exclude procedures in which the capacity limit must be shared between chunk storage and the storage of intermediate results of processing. One example of this is the "*n*-back task" in which each item in a continuous series must be compared with the item that occurred *n* items ago (e.g., Cohen et al. 1997; Poulton 1954) or a related task in which the subject must listen to a series of digits and detect three odd digits in a row (Jacoby et al. 1989). In these tasks, in order to identify a fixed set of the most recent n items in memory, the subject must continually update the target set in memory. This task requirement may impose a heavy additional storage demand. These demands can explain why such tasks remain difficult even with $n = 3$.

It may be instructive to consider a hypothetical version of the *n*-back task that would be taken to indicate the existence of a special capacity limit. Suppose that the subject's task were to indicate, as rapidly as possible, if a particular item had been included in the stimulus set previously. Some items would be repeated in the set but other, novel items also would be introduced. On positive trials, the mean reaction time should be much faster when the item had been presented within the most recent three or four items than when it was presented only earlier in the sequence. To my knowledge, such a study has not been conducted. However, in line with the expectation, probed recall experiments have resulted in shorter reaction times for the most recent few items (Corballis 1967).

The present view is that a strong similarity in pure capacity limits (to about 4 chunks on average) can be identified across many test procedures meeting the above four criteria. The subcategories of methods and some key references are summarized in Table 1 (see sect. 3), and each area will be described in more detail in section 3 of the target article.

1.3. Definition of chunks

A chunk must be defined with respect to associations between concepts in long-term memory. I will define the term *chunk* as a collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use. (This definition is related to concepts discussed by Simon 1974.) It would be assumed

Table 1. *Types of evidence of a capacity limit of about four items, with selected key references (numbered according to the relevant section of the article)*

3.1. Imposing an information overload
3.1.1. Visual whole report of spatial arrays (Sperling 1960)
3.1.2. Auditory whole report of spatiotemporal arrays (Darwin et al. 1972)
3.1.3. Whole report of unattended spoken lists (Cowan et al. 1999)
3.2. Preventing long-term memory recoding, passive storage, and rehearsal
3.2.1. Short-term, serial verbal retention with articulatory suppression (see Table 3 references; also Pollack et al. 1959; Waugh & Norman 1965)
3.2.2. Short-term retention of unrehearsable material (Glanzer & Razel 1974; Jones et al. 1995; Simon 1974; Zhang & Simon 1985)
3.3. Examining performance discontinuities
3.3.1. Errorless performance in immediate recall (Broadbent 1975)
3.3.2. Enumeration reaction time (Mandler & Shebo 1982, Trick & Pylyshyn 1993)
3.3.3. Multi-object tracking (Pylyshyn et al. 1994)
3.3.4. Proactive interference in immediate memory (Halford et al. 1988; Wicklegren 1966)
3.4. Examining indirect effects of the limits
3.4.1. Chunk size in immediate recall (Chase & Simon 1973; Ericsson 1985; Ericsson et al. 1980; Ryan 1969; Wickelgren 1964)
3.4.2. Cluster size in long-term recall (Broadbent 1975; Graesser & Mandler 1978)
3.4.3. Positional uncertainty in recall (Nairne 1991)
3.4.4. Analysis of the recency effect in recall (Watkins 1974)
3.4.5. Sequential effects in implicit learning and memory (Cleeremans & McClelland 1991; McKone 1995)
3.4.6. Influence of capacity on properties of visual search (Fisher 1984)
3.4.7. Influence of capacity on mental addition reaction time (Logan 1988; Logan & Klapp 1991)
3.4.8. Mathematical modeling parameters (Halford et al. 1998; Kintsch & van Dijk 1978; Raaijmakers & Shiffrin 1981)

that the number of chunks can be estimated only when inter-chunk associations are of no use in retrieval in the assigned task. To use a well-worn example inspired by Miller (1956), suppose one tries to recall the series of letters, "fbicbsibmirs." Letter triads within this sequence (FBI, CBS, IBM, and IRS) are well-known acronyms, and someone who notices that can use the information to assist recall. For someone who does notice, there are pre-existing associations between letters in a triad that can be used to assist recall of the 12-letter sequence. If we further assume that there are no pre-existing associations between the acronyms, then the four of them have to occupy limited-capacity storage separately to assist in recall. If that is the case, and if no other optional mnemonic strategies are involved, then successful recall of the 12-item sequence indicates that the pure capacity limit for the trial was at least four chunks. (In practice, within the above example there are likely to be associations between the acronyms. For example, FBI and IRS represent two U.S. government agencies, and CBS and IBM represent two large U.S. corporations. Such associations could assist recall. For the most accurate pure capacity-based limit, materials would have to be selected so as to eliminate such special associations between chunks.) Notice that the argument is not that long-term memory fails to be involved in capacity-based estimates. Long-term memory is inevitably involved in memory tasks. The argument is that *the purest capacity estimates occur when long-term memory associations are as strong as possible within identified chunks and absent between those identified chunks.*

If someone is given new material for immediate recall and can look at the material long enough before responding, new associations between the original chunks can be formed, resulting in larger chunks or, at least, conglomerates with nonzero associations between chunks. McLean and Gregg (1967, p. 455) provided a helpful description of chunks in verbal recall, as "groups of items recited together quickly," helpful because recall timing provides one good indication of chunking (see also Anderson & Matessa 1997). McLean and Gregg (p. 456) described three ways in which chunks can be formed: "(a) Some stimuli may already form a unit with which S is familiar. (b) External punctuation of the stimuli may serve to create groupings of the individual elements. (c) The S may monitor his own performance and impose structure by selective attention, rehearsal, or other means."

The practical means to identify chunks directly is an important issue, but one that is more relevant to future empirical work than it is to the present theoretical review of already-conducted work, inasmuch as few researchers have attempted to measure chunks directly. Direct measures of chunks can include empirical findings of item-to-item associations that vary widely between adjacent items in a list, being high within a chunk and low between chunks; item-to-item response times that vary widely, being relatively short within a chunk and long between chunks; and subjective reports of grouping. For studies in which the main dependent measure is not overt recall, measures of chunking for a trial must follow the trial immediately if it cannot be derived from the main dependent measure itself. Schneider and Detweiler (1987, pp. 105–106) provide an excellent further discussion of how chunks can be identified through convergent measures.

For most of the research that will be summarized in section 3 below, however, the researchers provided no direct evidence of chunking or its absence. The present assumption for these studies is that chunk size can be reasonably inferred from the presence of the task demands described above in section 1.2, which should prevent inter-item chunking. The present thesis is that the great similarity of empirically-based chunk limits derived using these guidelines, reviewed in section 3, supports their validity because the guidelines yield a parsimonious, relatively uniform de-

scription of capacity limits of three to five chunks as the population average (with a maximum range of two to six chunks in individuals).

2. Theoretical framework

The most important theoretical point here is the identification of conditions under which a capacity limit can be observed (see sect. 1); reasons for this limit also are proposed. The theoretical model in this section provides a logical way to understand the empirical results presented in section 3. A fuller analytic treatment, consideration of unresolved issues, and comparison with other approaches is provided in section 4.

The basic assumptions of the present theoretical framework are (1) that the focus of attention is capacity-limited, (2) that the limit in this focus *averages about four chunks in normal adult humans,* (3) that no other mental faculties are capacity-limited, although some are limited by time and susceptibility to interference, and (4) that any information that is deliberately recalled, whether from a recent stimulus or from long-term memory, is restricted to this limit in the focus of attention. This last assumption depends on the related premise, from Baars (1988) and Cowan (1988; 1995), that only the information in the focus of attention is available to conscious awareness and report. The identification of the focus of attention as the locus of the capacity limit stems largely from a wide variety of research indicating that people cannot optimally perceive or recall multiple stimulus channels at the same time (e.g., Broadbent 1958; Cowan 1995), although most of that research does not provide estimates of the number of chunks from each channel that occupy the focus of attention at any moment. There is an additional notion that the focus of attention serves as a global workspace for cognition, as described, for example, by Cowan (1995, p. 203) as follows:

> Attention clearly can be divided among channels, but under the assumption of the unity of conscious awareness, the perceived contents of the attended channels should be somehow integrated or combined. As a simple supporting example, if one is instructed to divide attention between visual and auditory channels, and one perceives the printed word "dog" and the spoken word "cat," there should be no difficulty in determining that the two words are semantically related; stimuli that can be consciously perceived simultaneously can be compared to one another, as awareness serves as a "global workspace." (Baars 1988)

Cowan (1995) also suggested two other processing limits. Information in a temporarily heightened state of activation, yet not in the current focus of attention, was said to be time-limited. Also, the transfer of this activated information into the focus of attention was said to be rate-limited. Important to note, however, only the focus of attention was assumed to be capacity-limited. This assumption differs from approaches in which there are assumed to be multiple capacity limits (e.g., Wickens 1984) or perhaps no capacity limit (Meyer & Kieras 1997).

The assignment of the capacity limit to the focus of attention has parallels in previous work. Schneider and Detweiler (1987) proposed a model with multiple storage buffers (visual, auditory, speech, lexical, semantic, motor, mood, and context) and a central control module. They then suggested (p. 80) that the control module limited the memory that could be used:

> 50 semantic modules might exist, each specializing in a given class of words, e.g., for categories such as animals or vehicles. Nevertheless, if the controller can remember only the four most active buffers, the number of active semantic buffers would be effectively only four buffers, regardless of the total number of modules. . . . Based on our interpretations of empirical literature, the number of active semantic buffers seems to be in the range of three to four elements.

The present analysis, based on Cowan (1988; 1995), basically agrees with Schneider and Detweiler, though with some differences in detail. First, it should be specified that the elements limited to four are chunks. (Schneider & Detweiler probably agreed with this, though it was unclear from what was written.) Second, the justification for the particular modules selected by Schneider and Detweiler (or by others, such as Baddeley,1986) is dubious. One can always provide examples of stimuli that do not fit neatly into the modules (e.g., spatial information conveyed through acoustic stimulation). Cowan (1988; 1995) preferred to leave open the taxonomy, partly because it is unknown and partly because there may in fact not be discrete, separate memory buffers. Instead, there could be the activation of multiple types of memory code for any particular stimulus, with myriad possible codes. The same general principles of activation and de-activation might apply across all types of code (e.g., the principle that interference with memory for an item comes from the activation of representations for other items with similar memory codes), making the identification of particular discrete buffers situation-specific and therefore arbitrary. Third, Cowan (1995) suggested that the focus of attention and its neural substrate differ subtly from the controller and its neural substrate, though they usually work closely together. In particular, for reasons beyond the scope of this target article, it would be expected that certain types of frontal lobe damage can impair the controller without much changing the capacity of the focus of attention, whereas certain types of parietal lobe damage would change characteristics of the focus of attention without much changing the controller (see Cowan 1995). In the present analysis, it is assumed that the capacity limit occurs within the focus of attention, though the control mechanism is limited to the information provided by that focus.

In the next section, so as to keep the theoretical framework separate from the discussion of empirical evidence, I will continue to refer to evidence for a "capacity-limited STM" without reiterating that it is the focus of attention that presumably serves as the basis of this capacity limit. (Other, non-capacity-limited STM mechanisms that may be time-limited contribute to compound STM measures but not to capacity-limited STM.) Given the usual strong distinction between attention and memory (e.g., the absence of memory in the central executive mechanism as discussed by Baddeley 1986), the suggested equivalence of the focus of attention and the capacity-limited portion of STM may require some getting used to by many readers. With use of the term "capacity-limited STM," the conclusions about capacity limits could still hold even if it were found that the focus of attention is not, after all, the basis of the capacity limit.

A further understanding of the premise that the focus of attention is limited to about four chunks requires a discussion of working assumptions including memory retrieval, the role of long-term memory, memory activation, maintenance rehearsal, other mnemonic strategies, scene co-

herence, and hierarchical shifting of attention. These are discussed in the remainder of section 2. In section 3, categories of evidence will be explained in detail. Finally, in section 4, on the basis of the evidence, the theoretical view will be developed and evaluated more extensively with particular attention to possible reasons for the capacity limits.

2.1. Memory retrieval

It is assumed here that explicit, deliberate memory retrieval within a psychological task (e.g., recall or recognition) requires that the retrieved chunk reside in the focus of attention at the time immediately preceding the response. The basis of this assumption is considerable evidence, beyond the scope of this article, that explicit memory in direct memory tasks such as recognition and recall requires attention to the stimuli at encoding and retrieval, a requirement that does not apply to implicit memory as expressed in indirect memory tasks such as priming and word fragment completion (for a review, see Cowan 1995). Therefore, any information that is deliberately recalled, whether it is information from a recent stimulus or from long-term memory, is subject to the capacity limit of the focus of attention. In most cases within a memory test, information must be recalled from both the stimulus and long-term memory in order for the appropriate units to be entered into the focus of attention. For example, if we attempt to repeat a sentence, we do not repeat the acoustic waveform; we determine the known units that correspond to what was said and then attempt to produce those units, subject to the capacity limit.

A key question about retrieval in a particular circumstance is whether anything about the retrieval process makes it impossible to obtain a pure capacity-based STM estimate. A compound STM estimate can result instead if there is a source of information that is temporarily in a highly accessible state, yet outside of the focus of attention. This is particularly true when a subject's task involves the reporting of chunks one at a time, as in most recall tasks. In such a situation, if another mental source is available, the subject does not need to hold all of the to-be-reported information in the focus of attention at one time. In a trivial example, a compound, supplemented digit capacity limit can be observed if the subject is trained to use his or her fingers to hold some of the information during the task (Reisberg et al. 1984). The same is true if there is some internal resource that can be used to supplement the focus of attention.

2.2. The role of long-term memory

Whereas some early notions of chunks may have conceived of them as existing purely in STM, the assumption here is that chunks are formed with the help of associations in long-term memory, although new long-term memory associations can be formed as new chunks are constructed. It appears that people build up data structures in long-term memory that allow a simple concept to evoke many associated facts or concepts in an organized manner (Ericsson & Kintsch 1995). Therefore, chunks can be more than just a conglomeration of a few items from the stimulus. Gobet and Simon (1996; 1998) found that expert chess players differ from other chess players not in the number of chunks but in the size of these chunks. They consequently invoked the term "template" to refer to large patterns of information that an expert can retain as a single complex chunk (concerning expert information in long-term memory, see also Richman et al. 1995).

The role of long-term memory is important to keep in mind in understanding the size of chunks. When chunks are formed in the stimulus field on the basis of long-term memory information, there should be no limit to the number of stimulus elements that can make up a chunk. However, if chunks are formed rapidly through new associations that did not exist before the stimuli were presented (another mechanism suggested by McLean & Gregg 1967), then it is expected that the chunk size will be limited to about four items because all of the items (or old chunks) that will be grouped to form a new, larger chunk must be held in the focus of attention at the same time in order for the new intrachunk associations to be formed (cf. Baars 1988; Cowan 1995). This assumption is meant to account for data on limitations in the number of items per group in recall (e.g., see section 2.7). It should be possible theoretically to increase existing chunk sizes endlessly, little by little, because each old chunk occupies only one slot in the capacity-limited store regardless of its size.

2.3. Memory activation

It is assumed that there is some part of the long-term memory system that is not presently in the focus of attention but is temporarily more accessible to the focus than it ordinarily would be, and can easily be retrieved into that focus if it is needed for successful recall (Cowan 1988; 1995). This accessible information supplements the pure capacity limit and therefore must be understood if we are to determine that pure capacity limit.

According to Baddeley (1986) and Cowan (1995), when information is activated (by presentation of that information or an associate of it) it stays activated automatically for a short period of time (e.g., 2 to 30 sec), decaying from activation unless it is reactivated during that period through additional, related stimulus presentations or thought processes. In Baddeley's account, this temporary activation is in the form of the phonological buffer or the visuospatial sketch pad. As mentioned above, there is some question about the evidence for the existence of that activation-and-decay mechanism. Even if it does not exist, however, there is another route to temporary memory accessibility, described by Cowan et al. (1995) as "virtual short-term memory" and by Ericsson and Kintsch (1995), in more theoretical detail, as "long-term working memory." For the sake of simplicity, this process also will be referred to as *activation*. Essentially, an item can be tagged in long-term memory as relevant to the current context. For example, the names of fruits might be easier to retrieve from memory when one is standing in a grocery store than when one is standing in a clothing store because different schemas are relevant and different sets of concepts are tagged as relevant in memory. Analogously, if one is recalling a particular list of items, it might be that a certain item from the list is out of the focus of attention at a particular point but nevertheless is temporarily more accessible than it was before the list was presented. For example, if one is buying groceries based on a short list that was not written down, a fruit forgotten from the list might be retrieved with a process resembling the following stream of thought: "I recall that there were three

fruits on the list and I already have gotten apples and bananas . . . what other fruit would I be likely to need?" The data structure in long-term memory then allows retrieval. One difference between this mechanism and the short-term decay and reactivation mechanism is that it is limited by contextual factors rather than by the passage of time.

If there is no such thing as time-based memory decay, the alternative assumption is that long-term working memory underlies phenomena that have been attributed to the phonological buffer and visuospatial sketchpad by Baddeley (1986). In the present article, the issue of whether short-term decay and reactivation exists will not be addressed. Instead, it is enough to establish that information can be made temporarily accessible (i.e., in present terms, active), by one means or another and that this information is the main data base for the focus of attention to draw upon.

2.4. Maintenance rehearsal

In maintenance rehearsal, one thinks of an item over and over and thereby keeps it accessible to the focus of attention (Baddeley 1986; Cowan 1995). One way in which this could occur, initially, is that the rehearsal could result in a recirculation of information into the focus of attention, reactivating the information each time. According to Baddeley (1986), the rehearsal loop soon becomes automatic enough so that there is no longer a need for attention. A subject in a digit recall study might, according to this notion, rehearse a sequence such as "2, 4, 3, 8, 5" while using the focus of attention to accomplish other portions of the task, provided that the rehearsal loop contains no more than could be articulated in about 2 sec. In support of that notion of automatization, Guttentag (1984) used a secondary probe task to measure the allocation of attention and found that as children matured, less and less attention was devoted to rehearsal while it was going on.

It appears from many studies of serial recall with rehearsal-blocking or "articulatory suppression" tasks, in which a meaningless item or short phrase is repeated over and over, that rehearsal is helpful to recall (for a review, see Baddeley 1986). Maintenance rehearsal could increase the observed memory limit as follows. An individual might recall an 8-item list by rehearsing, say, five of the items while holding the other three items in the focus of attention. Therefore, maintenance rehearsal must be prevented before pure capacity can be estimated accurately.

2.5. Other mnemonic strategies

With the possible exception of maintenance rehearsal, other well-known mnemonic strategies presumably involve the use of long-term memory. In *recoding,* information is transformed in a way that can allow improved associations. For example, in remembering two lines of poetry that rhyme, an astute reader may articulate the words covertly so as to strengthen the temporary accessibility of a phonological or articulatory code in addition to whatever lexical code already was strong. This phonological code in turn allows the rhyme association to assist retrieval of activated information into the focus of attention. Another type of recoding is the gathering of items (i.e., chunks corresponding to stimuli as intended by the experimenter) into larger chunks than existed previously. This occurs when an individual becomes aware of the associations between items, such as the fact that the 12-letter string given above could be divided into four 3-letter acronyms. *Elaborative rehearsal* involves an active search for meaningful associations between items. For example, if the items "fish, brick" were presented consecutively, one might form an image of a dead fish on a brick, which could be retrieved as a single unit rather than two unconnected units. Recoding and elaborative rehearsal are not intended as mutually exclusive mechanisms, but slightly different emphases on how long-term memory information can be of assistance in a task in which memory is required. These, then, are some of the main mechanisms causing compound STM limits to be produced instead of pure capacity-based STM limits.

2.6. Scene coherence

The postulation of a capacity of about four chunks appears to be at odds with the earlier finding that one can comprehend only one stream of information at a time (Broadbent 1958; Cherry 1953) or the related, phenomenologically-based observation that one can concentrate on only one event at a time. A resolution of this paradox was suggested by Mandler (1985, p. 68) as follows:

> The organized (and limited) nature of consciousness is illustrated by the fact that one is never conscious of some half dozen totally unrelated things. In the local park I may be conscious of four children playing hopscotch, or of children and parents interacting, or of some people playing chess; but a conscious content of a child, a chess player, a father, and a carriage is unlikely (unless of course they form their own meaningful scenario).

According to this concept, a coherent scene is formed in the focus of attention and that scene can have about four separate parts in awareness at any one moment. Although the parts are associated with a common higher-level node, they would be considered separate provided that there are no special associations between them that could make their recall mutually dependent. For example, four spices might be recalled from the spice category in a single retrieval (to the exclusion of other spices), but salt and pepper are directly associated and so they could only count as a single chunk in the focus of attention.

This assumption of a coherent scene has some interesting implications for memory experiments that may not yet have been conducted. Suppose that a subject is presented with a red light, a spoken word, a picture, and a tone in rapid succession. A combination of long-term memory and sensory memory would allow fairly easy recognition of any of these events, yet it is proposed that the events cannot easily be in the focus of attention at the same time. One possible consequence is that it should be very difficult to recall the serial order of these events because they were not connected into a coherent scene. They can be recalled only by a shifting of attention from the sensory memory or the newly formed long-term memory representation of one item to the memory representation of the next item, which does not result in a coherent scene and is not optimal for serial recall.

2.7. Hierarchical shifting of attention

Attentional focus on one coherent scene does not in itself explain how a complex sequence can be recalled. To understand that, one must take into account that the focus of attention can shift from one level of analysis to another.

McLean and Gregg (1967, p. 459) described a hierarchical organization of memory in a serial recall task with long lists of consonants: "At the top level of the hierarchy are those cueing features that allow S to get from one chunk to another. At a lower level, within chunks, additional cues enable S to produce the integrated strings that become his overt verbal responses." An example of hierarchical organization was observed by Graesser and Mandler (1978) in a long-term recall task. The assumption underlying this research was that, like perceptual encoding, long-term recall requires a limited-capacity store to operate. It was expected according to this view that items would be recalled in bursts as the limited-capacity store (the focus of attention) was filled with information from long-term memory, recalled, and then filled and recalled again. Studies of the timing of recall have indeed found that retrieval from long-term memory (e.g., recall of all the fruits one can think of) occurs in bursts of about five or fewer items (see Broadbent 1975; Mandler 1975). Graesser and Mandler (1978, study 2) had subjects name as many instances of a semantic category as possible in 6 min. They used a mathematical function fit to cumulative number of items recalled to identify plateaus in the response times. These plateaus indicated about four items per cluster. They also indicated, however, that there were lengthenings of the inter-cluster interval that defined superclusters. Presumably, the focus of attention shifted back and forth between the supercluster level (at which several subcategories of items are considered) and the cluster level (at which items of a certain subcategory are recalled). An example would be the recall from the fruit category as follows: "apple–banana–orange–pear (some common fruits); grapes–blueberries–strawberries (smaller common fruits); pineapple–mango (exotic fruits); watermelon–cantaloupe–honeydew (melons). By shifting the focus to higher and lower levels of organization it is possible to recall many things from a scene. I assume that the capacity limit applies only to items within a single level of analysis, reflecting simultaneous contents of the focus of attention.

3. Empirical evidence for the capacity limit

3.1. Capacity limits estimated with information overload

One way in which long-term recoding or rehearsal can be limited is through the use of stimuli that contain a large number of elements for a brief period of time, overwhelming the subject's ability to rehearse or recode before the array fades from the time-limited buffer stores. This has been accomplished in several ways.

3.1.1. Visual whole report of spatial arrays. One study (Sperling 1960) will be explained in detail, as it was among the first to use the logic just described. It revealed evidence for both (1) a brief, pre-attentive, sensory memory of unlimited capacity and (2) a much more limited, post-attentive form of storage for categorical information. Sperling's research was conducted to explore the former but it also was informative about the latter, limited-capacity store. On every trial, an array of characters (e.g., 3 rows with 4 letters per row) was visually presented simultaneously, in a brief (usually 50-msec) flash. This was followed by a blank screen. It was assumed that subjects could not attend to so many items in such a brief time but that sensory memory outlasted the brief stimulus array, and that items could be recalled to the extent that the information could be extracted from that preattentive store. On partial report trials, a tone indicated which row of the array the subject should recall (in a written form), but on whole report trials the subject was to try to recall the entire array (also in written form). The ability to report items in the array depended on the delay of the partial report cue. When the cue occurred very shortly after the array, most or all of the four items in the cued row could be recalled, but that diminished as the cue delay increased, presumably because the sensory store decayed before the subject knew which sensory information to bring into the more limited, categorical store.

By the time the cue was 1 sec later than the array, it was of no value (i.e., performance reached an asymptotically low level). Subjects then could remember about 1.3 of the cued items from a row of 4. It can be calculated that at that point the number of items still remembered was 1.3×3 (the number of rows in the array) or about 4. That was also how many items subjects could recall on the average on trials in the "whole report" condition, in which no partial report cue was provided. The limit of four items was obtained in whole report across a large variety of arrays differing in the number, arrangement, and composition of elements. Thus, a reasonable hypothesis was that subjects could read about four items out of sensory memory according to a process in which the unlimited-capacity, fading sensory store is used quickly to transfer some items to a limited-capacity, categorical store (according to the present theoretical framework, the focus of attention).

One could illustrate the results of Sperling (1960) using Figure 1, which depicts the interaction of nested faculties in the task in a manner similar to Cowan (1988; 1995). Within that theoretical account, sensory memory is assumed to operate through the activation of features within long-term memory (an assumption that has been strengthened through electrophysiological studies of the role of reactivation in automatic sensory memory comparisons; see Cowan et al. 1993). The nesting relation implies that some, but not all, sensory memory information also is in the focus of attention at a particular moment in this task. In either the whole report condition or the partial report condition, the limited capacity store (i.e., the focus of attention) can be filled with as many of the items from sensory memory as the limited capacity will allow; but in the partial report condition, most of these items come from the cued row in the array. Because the display is transient and contains a large amount of information, subjects have little chance to increase the amount recalled through mnemonic strategies such as maintenance or elaborative rehearsal. Part A of the figure represents whole report and shows that a subset of the items can be transferred from activated memory to the capacity-limited store. Part B of the figure, representing partial report, shows that the items transferred to the capacity-limited store are now confined to the cued items (filled circles), allowing a larger proportion of those items to be reported.

A word is in order about the intent of this simple model shown in Figure 1. It is not meant to deny that there are important differences between more detailed structures such as the phonological store and the visuospatial sketch pad of Baddeley (1986). However, the model is meant to operate on a taxonomically inclusive level of analysis. It seems likely that there are other storage structures not included in Bad-

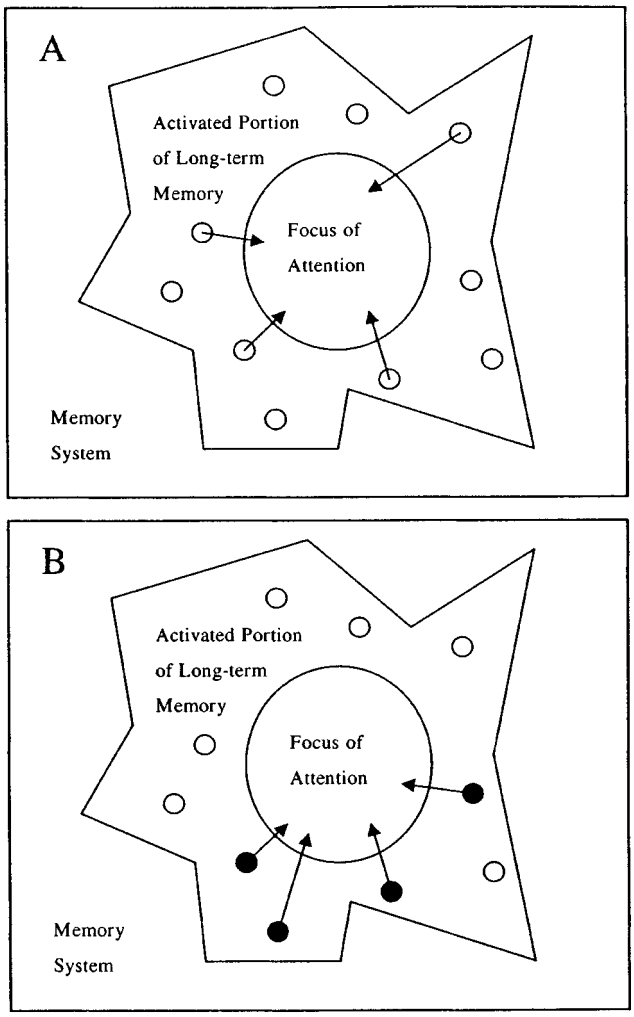


Figure 1. Illustration of the processing in (*A*) whole report procedures and (*B*) partial report procedures according to the nested processes framework suggested by Cowan (1988; 1995). In each figure, information is elevated from activated storage (jagged lines) to the limited-capacity store, which is the focus of attention (large circle), until the latter is full. Small circles represent items in the array, and those with arrows enter the focus of attention. With partial report (*B*), the cued items (filled circles) preferentially enter the limited-capacity focus of attention.

deley's model, such as memory for nonverbal sounds and for tactile stimuli. In principle, moreover, these separate structures could share important functional properties (e.g., incoming stimuli requiring a particular kind of coding interfere with the short-term retention of existing representations using similar coding) and could operate based on similar neural mechanisms of activation. Given that we do not know the taxonomy of short-term stores, they are represented together in the model according to their common principle, that they are activated portions of the long-term memory system. This activated memory includes both physical features and conceptual features. What is critical for the present purposes is that all of the storage structures making up activated memory are assumed not to have capacity limitations. Instead, they are assumed to be limited because of memory decay, interference from subsequent stimuli, and/or some other basis of temporary accessibility. *Only the focus of attention is assumed to have a fixed capacity limit in chunks,* and it is that capacity limit that is of primary concern here.

There are also a number of other theoretical suggestions that are consistent with the present approach but with different terminology and assumptions. For example, the approach appears compatible with a model proposed recently by Vogel et al. (1998). Their "conceptual short-term memory" would correspond to the activated portion of long-term memory in the model of Cowan (1988), whereas their "visual working memory" would correspond to the focus of attention. Potential differences between the approaches appear to be that what they call conceptual memory could, according to Cowan (1988), include some physical features; and what they call visual working memory would, according to Cowan (1988), prove to be one instance of the focus of attention, a central structure that represents conscious information from all modalities. The most critical similarity between the models for present purposes is that the capacity limit shows up in only one place (the visual working memory or focus of attention), not elsewhere in the model.

With these theoretical points in mind, we can return to a consideration of Sperling's study. The observed limit to about four items in whole report theoretically might be attributed to output interference. However, studies by Pashler (1988) and Luck and Vogel (1997), in which output interference was limited, militate against that interpretation. In one experiment conducted by Luck and Vogel, for example, subjects saw an array of 1 to 12 small colored squares for 100 msec and then, after a 900-msec blank interval, another array that was the same or differed in the color of one square. The subject was to indicate whether the array was the same or different. Thus, only one response per trial was required. Performance was nearly perfect for arrays of one to three squares, slightly worse with four squares, and much worse at larger array sizes. Very similar results were obtained in another experiment in which a cue was provided to indicate which square might have changed, reducing decision requirements. Some of their other experiments clarify the nature of the item limit. The four-item limit was shown to apply to integrated objects, not features within objects. For example, when objects in an array of bars could differ on four dimensions (size, orientation, color, and presence or absence of a central gap), subjects could retain all four dimensions at once as easily as retaining any one. The performance function of proportion correct across increasing array size (i.e., increasing number of array items) was practically identical no matter how many stimulus attributes had to be attended at once. This suggested that the capacity limit should be expressed in terms of the number of integrated objects, not the number of features within objects. The objects serve as the chunks here.

Broadbent (1975) noted that the ability to recall items from an array grows with the visual field duration: "for the first fiftieth of a second or so the rate of increase in recall is extremely fast, and after that it becomes slower." He cites Sperling's (1967) argument that in the early period, items are read in parallel into some visual store; but that, after it fills up, additional items can be recalled only if some items are read (more slowly) into a different, perhaps articulatory store. Viewed in this way, the visual store would have a capacity of three to five items, given that the performance function rapidly increases for that number of items. However, the "visual store" could be a central capacity limit (assumed here to be the focus of attention) rather than visu-

ally specific as the terminology used by Sperling seems to imply.

A related question is what happens when access to the sensory memory image is limited. Henderson (1972) presented 3 × 3 arrays of consonants, each followed by a masking pattern after 100, 400, 1,000, or 1,250 msec. This was followed by recall of the array. Although the number of consonants reported in the correct position depended on the duration of the array, the range of numbers was quite similar to other studies, with means for phonologically dissimilar sets of consonants ranging from about 3 with 100-msec exposure times to about 5.5 with 1250-msec exposure times. This indicates that most of the transfer of information from sensory storage to a limited-capacity store occurs rather quickly.

A similar limit may apply in situations in which a scene is changed in a substantial manner following a brief interruption and people often do not notice the change (e.g., Simons & Levin 1998). Rensink et al. (1997) proposed that this limit may occur because people can monitor only a few key elements in a scene at one time.

3.1.2. Auditory whole report of spatiotemporal arrays. Darwin et al. (1972) carried out an experiment that was modeled after Sperling's (1960) work, but with stimuli presented in the auditory modality. On each trial, subjects received nine words in a spatiotemporal array, with sequences of three spoken items (numbers and letters) presented over headphones at left, center, and right locations simultaneously for a total array size of nine items. The partial report cue was a visual mark indicating which spatial location to recall. The results were quite comparable to those of Sperling (1960). Once more the partial report performance declined across cue delays until it was equivalent to the whole report level of about four items though, in this experiment in the auditory modality, the decline took about 4 sec rather than 1 sec as in vision, and the last item in each sequence was recalled better than the first two items. In both modalities, the whole report limit may suggest the limited capacity for storage of item labels in a consciously accessed form.

3.1.3. Whole report of ignored (unattended) spoken lists. In all of the partial report studies, the measure of short-term memory capacity depended upon the fact that there were too many simultaneously presented items for all of them to be processed at once, so that the limited-capacity mechanism was filled with items quickly. If items were presented slowly and one at a time, the subject would be able to use mnemonic processes such as rehearsal (e.g., Baddeley 1986) to expand the number of items that could be held, and therefore would be able to exceed the constraints of the limited-capacity store. If a way could be found to limit these mnemonic processes, it could allow us to examine pure capacity in a test situation more similar to what is ordinarily used to examine STM (presumably yielding a compound STM estimate); namely immediate, serial verbal list recall.

Cowan et al. (1999) limited the processing of digits in a spoken list by having subjects ignore the items in the spoken list until after their presentation. Subjects played a computer game in which the name of a picture at the center of the screen was to be compared to the names of four surrounding pictures to indicate (with a mouse click) which one rhymed with the central picture. A new set of pictures then appeared. As this visual game was played repeatedly, subjects ignored lists of digits presented through headphones. Occasionally (just 16 times in a session), 1 sec after the onset of the last spoken word in a list, the rhyming game disappeared from the screen and a memory response screen appeared shortly after that, at which time the subject was to use the keypad to report the digits in the spoken list. Credit was given for each digit only if it appeared in the correct serial position. Relative to a prior memory span task result, lists were presented at span length and at lengths of span-1 (i.e., lists one item shorter than the longest list that was recalled in the span task), span-2, and span-3. A control condition in which subjects attended to the digits also was presented, before and after the ignored-speech session. In the attended-speech control condition, the number of digits recalled was higher than in the unattended condition, and it increased with list length. However, in the ignored-speech condition, the mean number of items recalled remained fixed at a lower level regardless of list length. The level was about 3.5 items in adults, and fewer in children. This pattern is reproduced in Figure 2. It is important that the number correct remained fixed across list lengths in the ignored-speech condition, just as the whole-report limit remained fixed across array sizes in Sperling (1960). It is this pattern that is crucial for the conclusion that there is a fixed capacity limit.

It is important also to consider individual-subject data. Sperling's (1960) data appeared to show that his very few, highly trained individuals had capacity limits in the range of about 3.5–4.5. In the study of Cowan et al. (1999), results from 35 adults are available even though only the first 24 of these were used in the published study. Figure 3 shows each adult subject's mean number correct in the unattended speech task, as well each subject's standard error and standard deviation across unattended speech trials. It is clear from this figure that individuals did not fit within a very narrow window of scores; their individual estimates of capacity ranged from as low as about 2 to as high as almost 6 in one participant. One might imagine that the higher estimates in some individuals were owing to residual attention to the supposedly ignored spoken digits, but the results do not support that suggestion. For example, consider the subject shown in Figure 3 who had the best memory for ignored speech. If that subject attended to the spoken digits that were to be ignored, then the result should have been a positive slope of memory across the four list lengths, similar to the attended-speech condition shown in Figure 2. In fact, however, that subject's scores across four list lengths had a slope of −0.35. Across all of the adult subjects, the correlation between memory for ignored speech and slope of the ignored speech memory function was $r = -.19$, n.s. The slight tendency was thus for subjects with better recall to have less positive slopes than those with poorer recall. The slopes were quite close to zero ($M = 0.05$, $SD = 0.32$) and were distributed fairly symmetrically around 0. Another possible indication of attention to the supposedly ignored speech would be a tradeoff between memory and visual task performance during the ignored speech session. However, such a tradeoff did not occur. The correlation between memory for unattended speech and reaction times on the visual task was −.33, n.s., the tendency being for subjects with better memory for ignored speech also to display slightly shorter reaction times on the visual task. The same type of result was obtained for the relation between memory and visual task reaction times on a trial-by-trial basis

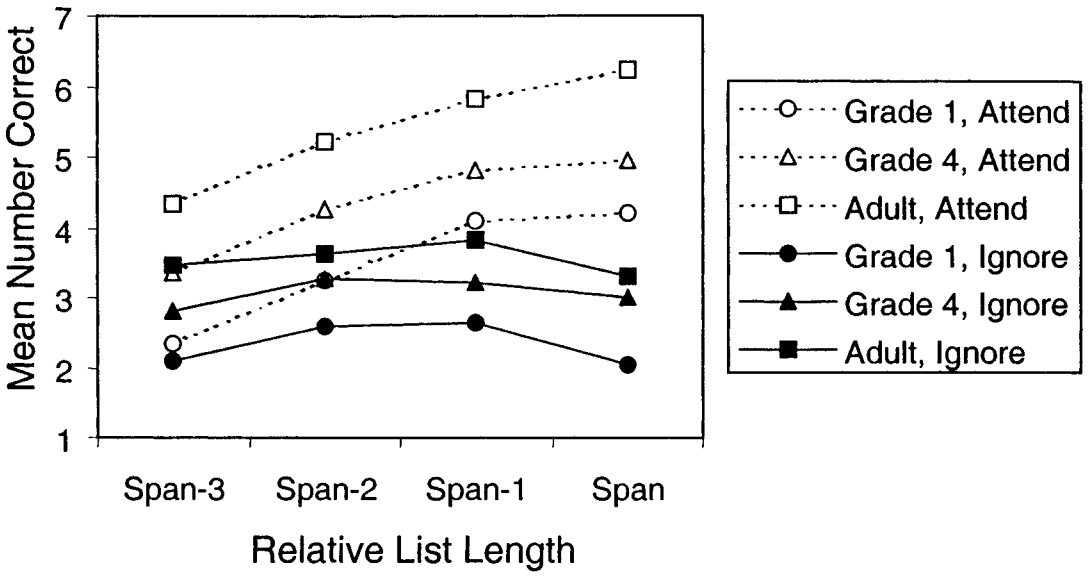


Figure 2. The number of items recalled in attended and unattended lists of digits. The flat pattern of recall of unattended digits as a function of list length is taken as evidence of a capacity limit analogous to what Sperling (1960) reported. After Cowan et al. (1999).

within individuals. The mean within-subject correlation was −.08 (SD = .25), showing that the slight tendency was for a subject's trials that produced better memory to be accompanied by shorter mean reaction times on the preceding visual task. Thus, the memory capacity of up to six items in certain individuals as measured in this technique and the individual differences in capacity seem real, not owing to attention-assisted encoding. Figure 3 shows that individuals' standard errors (rectangles) were relatively small, and that even the standard deviations (bars) of the best versus the worst rememberers did not overlap much.

The study of Cowan et al. (1999) is not the only one yielding individual difference information. For example, the data set reported as the first experiment of Luck and Vogel (1997), on visual storage capacity, resulted from individual subject estimates of storage capacity ranging from 2.2 to 4.7, and a graduate student who spent months on the capacity-estimation tasks developed a capacity of about six items (Steven Luck, personal communication, January 18, 1999). These estimates are quite similar to the ones shown in Figure 3 despite the great differences in procedures. Similar estimates can be obtained from the study of Henderson (1972), in which each consonant array was followed by a mask. For example, with a 400-msec field exposure duration (long enough to access sensory memory once, but probably not long enough for repeated access) and no supplementary load, the six subjects' mean number correct ranged from 3.0 to 5.1 items.

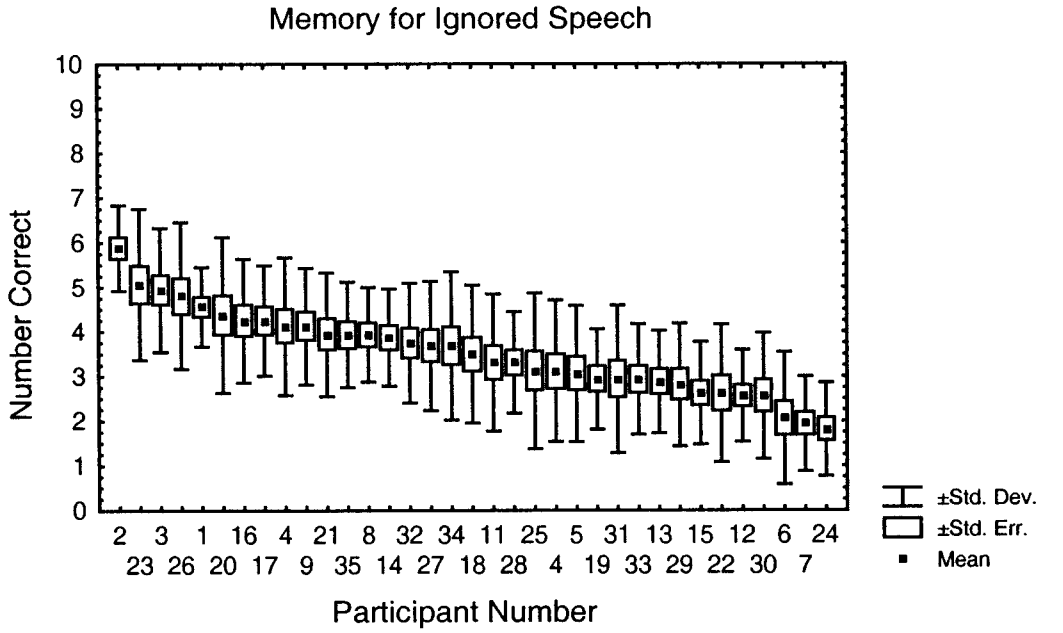


Figure 3. Number of items recalled in a memory for unattended speech task for each adult used in the procedure reported by Cowan et al. (1999). Error bars depict trials within an individual.

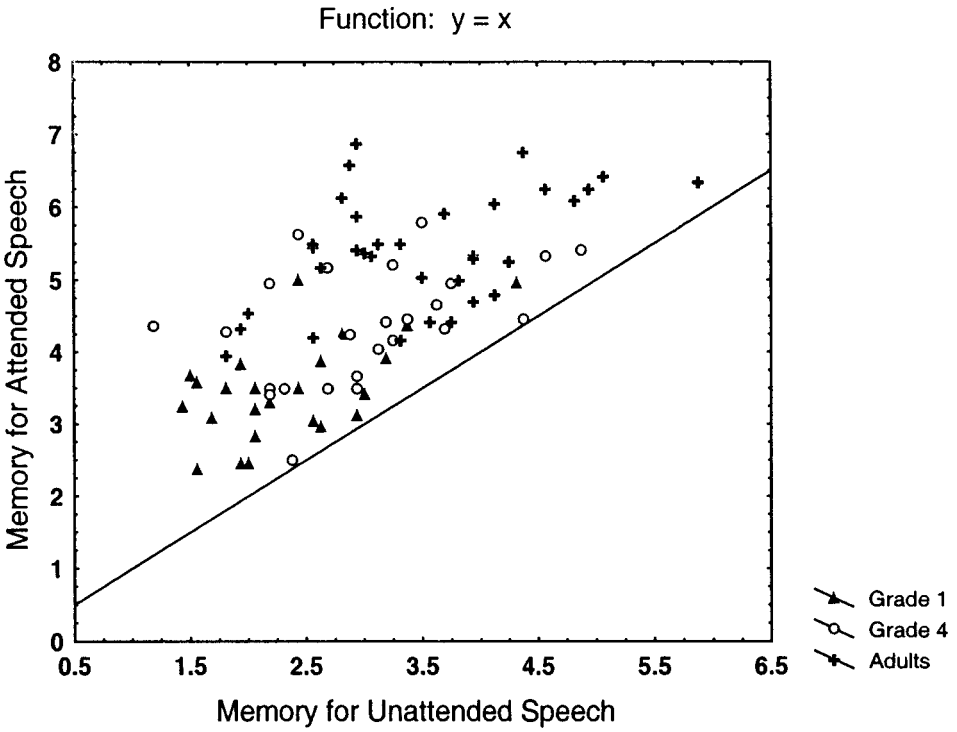


Figure 4. Scatterplot of memory for unattended versus attended speech in each age group in the experiment of Cowan et al. (1999).

All of these results appear to require a modification of conclusions that could be drawn from the previous literature. In his ground-breaking review of memory span, Dempster (1981, p. 87) concluded that "there is little or no evidence of either individual or developmental differences in capacity." In the previous literature only processing speeds were found related to span, but none of the previous developmental investigations examined memory with strategic processing during reception of the list minimized so as to examine capacity. There do appear to be individual and developmental differences in capacity.

Figure 4 illustrates another intriguing point about Cowan et al. (1999). In this scatterplot of memory for unattended versus attended speech in individuals within each age group, the equation line represents the case in which memory was equal in the two tasks. What the plot shows is that memory was always better in the attended speech task, but that the amount of improvement in the attended speech task relative to the unattended speech task was independent of the level of performance on the unattended task. In the attended condition the means (and SDs) were: for 35 adults, 5.43 (0.78); for 26 fourth graders, 4.31 (0.88); and for 24 first graders, 3.48 (0.69). In the ignored speech condition the comparable means were: for the adults, 3.51 (0.94); for the fourth graders, 2.99 (0.86); and for the first graders, 2.34 (0.69). Notice that, among all groups, the ratio of mean attended to unattended numbers correct fell within a narrow range, between 1.4 and 1.6. This pattern suggests that attention at the time of reception of the list may add a process that is independent of the processes involved in memory for unattended speech. That process presumably is independent of the pure capacity limit and could reflect the use of attention to form larger chunks.

It should be noted that the main scoring procedure used by Cowan et al. (1999) credited correct recall of a digit only if it appeared in the correct serial position. Cowan et al. also examined results of a scoring procedure in which credit was given for any correct digit, regardless of the serial position. Such results cannot be compared across list lengths because the probability of guessing correctly increases dramatically with list length (given that each digit could occur only once per list). Nevertheless, it is noteworthy that recall at all ages was more like a constant proportion correct across lists lengths in this free scoring, not a constant number correct as in the serial position scoring. Adults and fourth-grade children were over 90% correct on lists of length 4 through 6, the lengths examined with this scoring procedure, and first-grade children were correct on 83, 80, and 83% of the lists at these three lengths. The item scoring raises the question of what it is that is held in a capacity-limited mechanism. It cannot simply be the items that are held, as the free scoring does not show the limited-capacity pattern of a constant number correct across list lengths. The digits themselves may be stored in activated memory (e.g., auditory sensory or phonological memory) and drawn from it into the focus of attention as needed. Instead, it might be the mapping between the digits in memory and the serial positions in the list that would have to be held in capacity-limited storage.

3.2. Capacity limits estimated by blocking long-term memory recoding, passive storage, and rehearsal

Verbal materials can be used under conditions that discourage recoding and rehearsal, or materials that are intrinsically difficult to recode, store, and rehearse can be used. These methods force subjects to rely primarily on capacity-limited storage of chunks that were learned out of the laboratory or, at least, before the experimental trial in question.

3.2.1. Short-term, serial verbal retention with articulatory suppression. The contribution of long-term memory can be minimized by drawing the stimuli from the same, small set on every trial and requiring the correct recall of serial order. Because the same items recur over and over, it is difficult to retain long-term associations that help in the retention of serial order of the items on a particular trial. That is, in fact, the nature of the stimuli in most immediate, serial recall experiments that have been conducted. Further, the contribution of rehearsal can be minimized by imposing articulatory suppression (Baddeley 1986; Murray 1968), a secondary task in which the subject repeats, whispers, or mouths a rote utterance over and over during the presentation of items (e.g., "the, the, the . . . ") and sometimes throughout recall itself if a nonspeech recall mode is used.

Cowan et al. (1998) offered an account of what these variables do when used together. They proposed that when new words are presented on every trial in a serial recall task, the phonological portion of activated memory includes a phonological representation of the word sequence. However, when the same words are used over and over on every trial, all of the representations of items from the memory set become active in memory, so that the memory items in the current list cannot necessarily be distinguished from items used in previous trials. Rehearsal may allow a special representation of the to-be-recalled list to be constructed in active memory even under these circumstances in which a small set of items is used over and over. Cowan et al. offered these assumptions to explain why articulatory suppression has a much larger effect on performance for a small set of words than for large sets of words that are not repeated from trial to trial. A small set of words used over and over, along with articulatory suppression, may minimize the contribution of articulatory and passive phonological storage factors in recall. It is only under these conditions, for example, that the "word length effect," or advantage for lists composed of short words, is eliminated (LaPointe & Engle 1990). Word length effects that remain even with articulatory suppression when a large set of items is used can be explained on the grounds that phonological representations of these items are generated from long-term memory (Besner 1987) and remain active despite articulatory suppression. (An alternative interpretation of articulatory suppression effects would state that suppression works by taking up processing capacity rather than by blocking rehearsal. However, if that were true, suppression should impair performance even when a large set of words is used. Given that it does not, the alternative interpretation seems wrong.)

Before describing results of serial recall experiments with spoken stimuli and articulatory suppression, it is necessary to restrict the admissible serial recall data in a few other ways. Memory for multisyllabic words was excluded because these often might be retained as separate segments rather than integrated units (e.g., *fire-man* if morphemic segments are used; *um-brel-la* if syllabic segments are used). Memory for nonwords also was excluded because one might retain them in terms of separate phonemic or syllabic series even if they are monosyllabic. Only spoken words were included because articulatory suppression seems to interfere with the retrieval of the phonological representation of printed words, but not of spoken words. For example, articulatory suppression during the presentation of a list eliminates phonological similarity effects for printed words, but not for spoken words (Baddeley et al. 1984). Finally, conditions with highly unusual stimulus parameters were eliminated. Unusually slow stimulus presentations (> 4 sec per word) were excluded because it might be possible to insert rehearsals despite the articulatory suppression, as were unusually fast presentations (< 0.5 sec per word) because of encoding difficulty; and grouped presentations were omitted because they encourage long-term recoding of the list.

Table 2 shows the results for all studies meeting these constraints. I was able to find nine studies that included at least one experimental condition involving the immediate recall of spoken, monosyllabic words from a small set in the presence of articulatory suppression. Among these studies I was able to derive 17 independent estimates of memory storage. There appears to be a striking degree of convergence among the 17 estimates. All but one of the estimates fell within the range of three to five items, and most fell in the three- to four-item range. The only outlier was an estimate of 2.4 items from Longoni et al. (1993). That low estimate is difficult to understand because the stimulus conditions were almost identical to another experimental condition in Longoni et al. that yielded an estimate of 3.4 items.

The methods of estimation are described briefly in Table 2. The most commonly applicable method was to take the proportion correct at each serial position (or, when necessary, an estimate of this proportion based on a figure) and add the proportions across serial positions to arrive at the number correct. In a probed recall experiment (e.g., Murray 1968) there is an initial list item for which the procedure produces no memory estimate; based on past research on primacy effects, the available proportion at this first serial position always was estimated at 0.8. For some studies, alternative assumptions led to alternative estimates of storage. For example, in the study of Peterson and Johnson (1971), the dependent measure reported was the number of lists recalled correctly, and to estimate items recalled one must make assumptions about the number of errors within the lists recalled incorrectly. Estimates of capacity are given in the table under a "high" assumption that at least four items were recalled within each 5-item list, and under a more "moderate" assumption that erroneous lists contained 1 or 2 errors (i.e., 4 or 3 correct items) equally often. It is the more moderate estimate that appears in the rightmost column of the table. When the measure was memory span, the estimate was taken as the span in conditions in which the articulatory suppression task can be presumed to have been most effective (e.g., in Cowan et al. 1987).

Waugh and Norman (1965) impeded rehearsal in a different way, through instructions to the subjects not to rehearse. In their experiment, each list contained 16 spoken digits and the last digit was accompanied by a tone. It was to serve as a probe, the same digit having occurred once before somewhere in the list. The subject was to respond with the digit that had followed the probe digit when it was presented earlier, in the list. Results with an ordinary, 1-per-sec presentation rate (e.g., Waugh & Norman 1965, p. 91) showed that performance levels were much higher with 3 or fewer items intervening between the target pair and the response (> .8) than it was with 4 or more intervening items (< .6). The transition between 3 and 4 intervening items was abrupt. Note that with 3 intervening items in this task, successful performance would require that the subject's

Table 2. *Estimates of capacity from studies of immediate verbal memory with auditory presentation and articulatory suppression*

Reference	Data Source	Method of Calculating Items in Storage	Est.
Murray (1968)	Figure 1, p. 682. Cued recall; auditory presentation with suppression. 6 letters.	Add the proportions correct across probed serial positions and assume recall of the first, unprobed item in the list = .8. Thus, as List Length 6: .4 + .2 +.4 + .5 + .8 = 2.3; + 1 = 3.3.	3.1
same	7 letters	same	3.2
same	8 letters	same	3.0
same	9 letters	same	3.1
Peterson & Johnson (1971)	Table 2, p. 349 (5 letters, serial recall; count during presentation, low-similarity condition) Data = proportion of lists recalled correctly.	5 items, 45% of lists correct. High assumption is that on the other 55% of trials, subjects get 4 correct, for a mean of 5(.45) + 4(.55) = 4.45. A more moderate assumption is 5(.45) + 4(.28) + 3(.27) = 4.18.	4.2
Levy (1971)	Table 1, p. 126. Neutral articulation condition cued recall, simultaneous auditory presentation. 7 serial positions.	7 ser. posit. 3 .39 items/s.p. = 2.73 items correct. Add 0.8 items for the first, unprobed position = 3.53 items.	3.5
same	Table 2, p. 130. 9 serial positions	9 ser. posit. × .34 = 3.06, add 0.8 for first, unprobed position = 3.86 items.	3.9
Baddeley et al. (1975b)	Figure 6, p. 585. Serial recall, monosyllabic words, auditory presentation with suppression.	5 serial positions × .7 items/s.p. correct = 3.5 items.	3.5
Baddeley et al. (1984)	Table 1, p. 236. Serial recall, monosyllabic words, dissimilar items with suppression. Fast presentation	5 serial positions × .64 items/s.p. correct = 3.2 items in each case.	3.2
same	Slower presentation	Same	3.2
Cowan et al. (1987)	Table 1, p. 514. Monosyllabic words, serial recall, span procedure, dissimilar items with articulatory suppression. Artic. Task: Whisper alphabet	Span = estimate. (Omitted conditions in which articulatory suppression task was presumably ineffective; whisper same letter once after each item, span = 4.81; whisper same letter continuously throughout study, span = 4.86.)	4.0
same	Task: Whisper next letter on each trial.	same	4.0
Longoni et al. (1993)	Table 3, p. 17. Serial recall, distinct items, suppression task = Whisper "the." Presentation rate of 0.5 sec per item.	6 serial positions × .57 correct = 3.42 (Additional data from a very slow presentation rate of 5 sec per item condition were omitted because rehearsal was possible; for that cond., 6 serial positions × .78 correct = 4.68.)	3.4
same	Table 4, p. 19. Whisper "hiya" during presentation and recall	6 serial positions × .40 correct = 2.40. It is not clear why such discrepant results obtained in these two experiements.	2.4
Avons et al. (1994)	Table 1, p. 215. Short words, immediate recall (all had suppression). Serial recall condition	5 serial positions × .69 items/s.p. correct = 3.45 items.	3.5
same	Probed recall condition. (Probed by the serial position of the item.)	5 serial positions × .72 items/s.p. correct = 3.60.	3.6
Hitch et al. (1996)	Figure 4, p. 125. Auditory presentation of items, suppression, recall in correct serial positions.	Dependent measure = items correct = about 4.0. (Omitted results for grouped lists, about 6.0).	4.0

memory extend back far enough to remember 4 items: the target pair and two intervening items. (The last intervening item was the probe, which did not have to be remembered.) Thus, this task leads to an estimate of four items in capacity-limited short-term storage. (Performance levels with a very fast, 4-per-sec presentation decreased rather more continuously as a function of the number of intervening items, which possibly could reflect the heavier contribution of a time-limited source of activation, such as sensory memory, that was the most vivid for more recent items and faded gradually across items.)

Another way to limit rehearsal is to use a "running memory span" procedure, in which a long list of items is presented and the subject is unaware of the point at which the test is to begin. Pollack et al. (1959) devised such a procedure. In their Experiment 1, lists of 25, 30, 35, and 40 digits were presented. When the list ended, the task was to write down as many of the most recent items as possible,

making sure to write them in the correct serial positions with respect to the end of the list. Under these conditions, the list was too long and continuous for rehearsal to do any good, and the obtained mean span was 4.2 digits. (Theoretically, it might be possible for the subject continually to compute, say, what were the last 5 items; but there is no task demand that would encourage such difficult work even if it were feasible. The absence of on-line task requirements makes this task very different from the *n*-back tasks, which, as discussed earlier, do not meet the criterial for inclusion.)

It is possible to prevent rehearsal in yet another way, by requiring processing between items rather than during the presentation of items. Consider, for example, the working memory span task of Daneman and Carpenter (1980) in which the subject must read sentences and also retain the final word of each sentence. The reading should severely limit rehearsal of the target words. Daneman and Carpenter (1980, p. 455) reported a mean span of 3.15 words in this circumstance. It is at first puzzling to think that subjects could do this well, inasmuch as they might need some of the capacity-limited storage space for processing the sentences (unless storage and processing demands are totally separate as suggested by Daneman & Carpenter 1980 and by Halford et al. 1998). Notice, however, that the word memory load does not reach 3 until after the third sentence has been processed. This might well leave some of the limited storage capacity available for sentence processing until the very end of the trial.

3.2.2. Short-term retention of unrehearsable material. A second way that time-limited stores can be eliminated from a measure of storage is with materials that, by their nature, cannot be rehearsed and thereby refreshed in active memory. It is unlikely that items that cannot be rehearsed lend themselves easily to long-term recoding, either. An analysis of one early study illustrates this distinction. Some verbal materials are too long to be rehearsed (Baddeley 1986). Simon (1974) examined this in an informal study using himself as a subject, and tried to remember well-known expressions such as "four score and seven years ago," "To be or not to be, that is the question," and "All's fair in love and war." He concluded that "lists of three such phrases were all I could recall with reliability, although I could sometimes retain four." Of course, the number of words and syllables contained in these phrases was much larger. Elsewhere in the article, for example, it was noted that seven one-syllable words could be recalled. The present theoretical assumption is that, in the recall of phrases, each phrase served as a previously learned chunk and also was too long to allow effective rehearsal; thus, by aiming the focus of attention at the phrase level, four such phrases could be recalled despite their inclusion of many more units on a subchunk level. In the recall of isolated words, in contrast, given that each word was much shorter than a phrase, it was presumably possible to use rehearsal to reactivate memory (and possibly to form new chunks larger than a single word) and therefore to increase the number of words recalled above what would be expected if each word were a separate chunk. This reasoning is supported by the fact that about four unconnected spoken words can be recalled when rehearsal is blocked, as shown in Table 2.

Jones et al. (1995) carried out an experiment that reveals a capacity limit, though that was not the purpose of the experiment. On each trial, a series of dots was presented one at a time at different spatial locations on the computer screen. After a variable test delay, the response screen included all of the dots and the task was to point to them in the serial order in which they had been presented. There was very little loss of information over retention intervals of up to 30 seconds. The authors suggested that this stability of performance across test delays indicates that some sort of "rehearsal" process was used. I would suggest that the so-called rehearsal process used here does not contaminate the estimate of storage because it is not a true rehearsal process. Instead, it may be a process in which some of the items, linked to serial position or order, are held in the capacity-limited store. Each list presented by Jones et al. included 4, 7, or 10 dots. It can be estimated from their paper (Jones et al., Fig. 2, p. 1011) that these three list lengths led to means of 3.5, 3.8, and 3.2 items recalled in a trial, respectively. These estimates were obtained by calculating the mean proportion correct across serial positions and multiplying it by the number of serial positions.

Several studies of the memory for unrehearsable material produce estimates lower than 3.0. Glanzer and Razel (1974) examined the free recall of proverbs and estimated the short-term storage capacity using the method developed by Waugh and Norman (1965), based on the recency effect. The estimate was 2.0 proverbs in short-term storage on the average. Glanzer and Razel also estimated the contents of short-term storage for 32 different free recall experiments, and found a modal value of 2.0–2.4 items in storage, very comparable to what they found for the proverbs. However, there is a potential problem with the Waugh and Norman (1965) method of estimating the contents of short-term storage. They assumed that the most recent items are recalled from either of two sources: short-term storage or long-term storage. The estimate of short-term storage is obtained by taking the list-medial performance level to reflect long-term memory and assuming that the recency effect occurs because of this same memory plus the additional contribution of short-term memory. This assumption is problematic, though, if the items in the recency positions are not memorized in the same way but are more often recalled only with the short-term store and not with the same contribution of long-term storage that is found for the earlier list items. This possibility is strengthened by the existence of negative recency effects in the final free recall of lists that previously had been seen in immediate recall (Craik et al. 1970). Glanzer and Razel consequently may have overcorrected for the contribution of long-term memory in the recency positions.

Another low estimate was obtained for unrehearsable material by Zhang and Simon (1985) using Chinese. In their Experiment 1, the mean number of items recalled was 2.71 when the items were radicals without familiar pronounceable names, and 6.38 (like the usual English memory span) when the items were characters with pronounceable, rehearsable names, within which radicals were embedded. A lower estimate for unrehearsable items is to be expected. However, the fact that it was lower than three would not be expected if, as the authors asserted, there are over 200 such radicals and "educated Chinese people can recognize every radical" (p. 194), making each radical a single visual chunk. It seems possible that there are visual similarities among three or more radicals that tend to make them interfere with one another in memory when radicals are presented in a meaningless series, preventing them from serving as in-

dependent chunks. Although this analysis is speculative, the basis of the discrepancy between these few estimates below 3.0 and the estimates obtained in the many other experiments taken to reflect a capacity limit (in the 3–5 chunk range) is an important area for future research.

3.3. Capacity limits estimated with performance discontinuities

Although subjects in some procedures may be able to perform when there are more than four items, the function describing the quality or speed of performance sometimes shows a discontinuity when one reaches about four items (e.g., a much longer reaction time cost for each additional item after the fourth item). Presumably, in these circumstances, some optional processing mechanism must be used to supplement the capacity-limited store only if the stimuli exceed the capacity. This can occur in several ways as shown below.

3.3.1. Errorless performance in immediate recall. Broadbent (1975) noted that we usually measure span as the number of items that can be recalled on 50% of the trials. However, he cites evidence that the number of items that can be recalled reliably, with a very high accuracy, is about three or four and is much more resistant to modifications based on the nature of the items (Cardozo & Leopold 1963; see also Atkinson & Shiffrin 1968). That is, there is a flat performance function across list lengths until three or four items. It stands to reason that when items beyond four are remembered, it is through the use of supplementary mnemonic strategies (such as rehearsal and chunking), not because of the basic storage capacity.

3.3.2. Enumeration reaction time. The ability to apprehend a small number of items at one time in the conscious mind can be distinguished from the need to attend to items individually when a larger number of such items are presented. This point is one of the earliest to be noted in psychological commentaries on the limitations in capacity. Hamilton (1859) treated this topic at length and noted (vol. 1, p. 254) that two philosophers decided that six items could be apprehended at once, whereas at least one other (Abraham Tucker) decided that four items could be apprehended. He went on to comment: "The opinion [of six] appears to me correct. You can easily make the experiment for yourselves, but you must be aware of grouping the objects into classes. If you throw a handful of marbles on the floor, you will find it difficult to view at once more than six, or seven at most, without confusion; but if you group them into twos, or threes, or fives, you can comprehend as many groups as you can units; because the mind considers these groups only as units, – it views them as wholes, and throws their parts out of consideration. You may perform the experiment also by an act of imagination." When the experiment actually was conducted, however, it showed that Hamilton's estimate was a bit high. Many studies have shown that the time needed to count a cluster of dots or other such small items rises very slowly as the number of items increases from one to four, and rises at a much more rapid rate after that. Jevons (1871) was probably the first actual study, noting that Hamilton's conjecture was "one of the very few points in psychology which can, as far as we yet see, be submitted to experiment." He picked up handfuls of beans and threw them into a box, glancing at them briefly and estimating their number, which was then counted for comparison. After over a thousand trials, he found that numbers up to four could be estimated perfectly, and up to five with very few errors.

Kaufman et al. (1949) used the verb "subitize" to describe the way in which a few items apparently can be apprehended and enumerated in a very rapid fashion (as if these items enter the focus of attention at the same time). In contrast, when there are more items, the reaction time or the time necessary for accurate counting increases steeply as the number of items increases (as if these items must enter the focus of attention to be counted piecemeal, not all at once). Mandler and Shebo (1982) described the history of the subitizing literature. As they note, subitizing has been observed via two main procedures: one in which the duration of an array is limited and the dependent measure is the proportion of errors in estimating the number of items in the array, and another method in which the array stays on and the primary dependent measure is the reaction time to respond with the correct number. The results from the first of these methods seem particularly clear. For example, in results reported by Mandler and Shebo (1982, p. 8), the proportion of errors was near zero for arrays of 1–4 items (or for 1–3 items with a presentation duration as short as 200 msec) and increased steeply after that, at a rate of about 15% per additional item until nearly 100% error was reached with an array size of 11. The reaction time increased slowly with array sizes of 1–3 and more steeply for array sizes of 5–8. After that it leveled off (whereas it continued to increase at the same rate, for much higher array sizes, in procedures in which the array stayed on and the dependent measure was the time to produce the number). The average response was identical to the correct response for array sizes of 1–8, with an increasing degree of underestimation as array size increased from 9 to 20. From the present viewpoint, it would appear that three or four items were subitized initially and about three or four more could be added to the subitized amount without losing track of which items had been counted.

Alternative hypotheses about enumeration and related processes must be considered. Trick and Pylyshyn (1994a) put forth a theory of subitizing suggesting that it is capacity-limited (hence the limit to four items), but still not attention-demanding, and that it takes place at a point in processing intermediate between unlimited-capacity automatic processes and serial or one-at-a-time attentive processes. It was called the FINST (finger of instantiation) theory in that there are a limited number of "fingers" of instantiation that can be used to define individual items in the visual field. This theory is specific to vision, and it was contrasted with a working memory theory in which subitizing is said to occur because of a limit in the number of temporary memory slots.

The evidence used by Trick and Pylyshyn (1994a) to distinguish between the theories is open to question. First, it was shown that items could be subitized only if they were organized in a way that made them "pop out" of the surroundings (the evidence of Trick & Pylyshyn 1993). Certainly, this suggests that there is a pre-attentive stage of item individuation, but perhaps the subitization occurs only afterward, contingent not only on this rapid item individuation as Trick and Pylyshyn said, but contingent also on the availability of slots. One reason to make this distinction is

that the phenomenon of popout clearly is not limited to four items; it obviously occurs for much larger numbers of items. For example, when one looks inside a carton of eggs, all of the eggs appear to pop out against the surrounding carton. It is the inclusion of individuated items in the enumeration routine that is limited to about four. Another type of evidence used by Trick and Pylyshyn (1994a) was that there was said to be no effect of a memory load on subitization, unlike counting. Logie and Baddeley (1987) were the main authors cited in this regard, though subitization was not the focus of their study. Logie and Baddeley did find that two distractor tasks (articulatory suppression from repetition of the word "the," and tapping) had little effect in the subitizing range, whereas articulatory suppression had an effect in the counting range. However, these tasks can be carried out relatively automatically and would not be expected to require much working memory capacity (Baddeley 1986). For example, unlike counting backward as a distractor task, which causes severe forgetting of a consonant trigram over an 18-sec distractor-filled period (Peterson & Peterson 1959), articulatory suppression causes almost no loss over a similar time period (Vallar & Baddeley 1982). Interference with articulatory processing can explain why articulatory suppression interfered with counting, for items over four in the array task and also for every list length within another task that involved enumeration of sequential events rather than simultaneous spatial arrays. The data of Logie and Baddeley thus do seem to support the distinction between subitizing and counting, but they do not necessarily support the FINST theory over the working memory limitation theory of subitizing.

Another type of evidence (from Trick & Pylyshyn 1994b) involved a cue validity paradigm (a variation of the procedure developed by Posner et al. 1980). On each trial in most of the experiments, two rectangles appeared; dots were to appear in only one rectangle. The task was to count the dots. Sometimes, there would be a cue (an arrow pointing to one rectangle or a flashing rectangle) to indicate slightly in advance which rectangle probably would contain the dots. The cue was valid (giving correct information) on 80% of the cued trials and invalid (giving incorrect information) on 20% of the cued trials. On other trials, no informative cue was given. The validity of the cue affected performance in the counting range more than in the subitizing range, leading Trick and Pylyshyn (1994a) to view the results as supportive of the FINST theory. However, there was still some effect of cue validity in the subitizing range, so the result is less than definitive in comparing the FINST and working memory accounts of subitizing.

Atkinson et al. (1976a) and then Simon and Vaishnavi (1996) investigated enumeration within afterimages so that subjects would be unable to shift their gaze in a serial fashion using eye movements. Both studies found that the subitizing limit remained at four items, with errors in enumeration only above that number, even though subjects had a long time to view each afterimage. Therefore, it seems that a focal attention strategy involving eye movements is important for visual enumeration of over four items, but not at or below four items, the average number that subjects may be able to hold in the limited-capacity store at one time.

3.3.3. Multi-object tracking. Another, more recent line of research involves "multi-object tracking" of dots or small objects that move around on the computer screen (Pylyshyn & Storm 1988; Yantis 1992; for a recent review see Pylyshyn et al. 1994). In the basic procedure, before the objects move, some of them flash several times and then cease flashing. After that all of them wander randomly on the screen and, when they stop, the subject is to report which dots had been flashing. The flavor of the results is described well by Yantis (1992, p. 307): "Performance deteriorated as the number of elements to be tracked increased from 3 to 5 [out of 10 on the screen]; tracking three elements was viewed by most subjects as relatively easy, although not effortless, while tracking 5 of 10 elements was universally judged to be difficult if not impossible by some subjects." As in subitizing, one could use either FINST or working memory theories to account for this type of finding.

3.3.4. Proactive interference in short-term memory. One can observe proactive interference (PI) in retrieval only if there are more than four items in a list to be retained (Halford et al. 1988). This presumably occurs because four or fewer items are, in a sense, already retrieved; they reside in a limited-capacity store, eliminating the retrieval step in which PI arises. Halford et al. demonstrated this storage capacity limit in a novel and elegant manner. They used variant of Sternberg's (1966) memory search task, in which the subject receives a list of items and then a probe item and must indicate as quickly as possible whether the probe appeared in the list. In their version of the task, modeled after Wickens et al. (1981), lists came in sets of three, all of which were similar in semantic category (Experiment 1) or rhyme category (Experiment 2). Thus, the first trial in each set of three was a low-PI trial, whereas the last trial in the set was a high-PI trial. Experiment 1 showed that with lists of 10 items, there were PI effects. With a list length of four, there was no PI. Experiment 2 showed that PI occurred for lists of six or more items, but not lists of four items. Presumably, the items within a list of four did not have to be retrieved because they all could be present within a capacity-limited store at the same time. Also consistent with this sort of interpretation, in 8- to 9-year-old children, PI was observed with 4 items, but not 2 items in a list. The magnitude of growth of a capacity limit with age in childhood matches what was observed by Cowan et al. (1999) with a very different procedure (see Fig. 2).

In the Halford et al. (1988) study, it was the length of the target list that was focused upon. We can learn more by examining also the effect of variations in the length of the list causing PI. Wickelgren's (1966) subjects copied a list of PI letters, a single letter to be recalled, and then a list of retroactive interference (RI) letters. The subject was to recall only the target letter. There were always eight letters in one of the interference sets (the PI set for some subjects, the RI set for others), whereas the other interfering set could contain 0, 4, 8, or 16 letters. There was a large effect of the number of RI letters, with substantial differences between any two RI list lengths. In contrast, when it was the PI set that varied in length, there was a difference between 0 and 4 PI letters but very little effect of additional PI letters beyond 4. Wickelgren suggested that PI and RI both generate associative interference, whereas RI additionally generates another source of forgetting (either decay or storage interference). Thus, associative interference would have been limited primarily to the four closest interfering items on either side of the target.

A mechanism of PI in these situations can be suggested.

It seems likely that excellent, PI-resistant recall occurs when the active contents of the limited-capacity store are to be recalled. When the desired information is no longer active, the long-term memory record of the correct former state of the limited-capacity store can be used as a cue to the recall of the desired item(s). If several former limited-capacity states were similar in content, it may be difficult to select the right one. Moreover, if the limited-capacity store serves as a workspace in which items become associated with one another (Baars 1988; Cowan 1995), then it might be difficult to select the correct item from among several present in the limited-capacity store simultaneously. The PI results described above could then be interpreted as follows. For studies in Halford et al. (1988), target lists of more than four items could not be held entirely within limited-capacity storage, so that a former state of the store had to be reconstituted. This could cause PI because some of the target items may have shared a former limited-capacity state with nearby items from a prior list, or because some of the other former limited-capacity states would have contained items resembling the correct item. For subjects in Wickelgren's (1966) study who received a variable number of PI letters, the target item would have been removed from the limited-capacity store by the presentation of eight following RI letters. Therefore, at the time of recall, the subject would have had to identify the former state of the limited-capacity store that contained the single target letter. This same former state may also have included several of the adjacent letters, which could become confused with the target letter. Only three or so letters adjacent to the target letter usually would have been in the limited-capacity store at the same time as the target letter, and thus only those letters would contribute much to associative interference. In a broader context, this analysis may be one instance of a cue-overload theory of PI (cf. Glenberg & Swanson 1996; Raaijmaker & Shiffrin 1981; Tehan & Humphreys 1996; Watkins & Watkins 1975) asserting that recall is better when fewer test items are associated with the cue used to recall the required information.

3.4. Capacity limits estimated with indirect effects

So far we have discussed effects of the number of stimulus items on a performance measure directly related to the subject's task, in which recall of items in the focus of attention is required. It is also possible to observe effects that are related to the subject's task only indirectly by deriving a theoretical estimate of capacity from the presumed role of the focus of attention in processing.

3.4.1. Chunk size in immediate recall. The "magical number 4" lurks in the background of the seminal article by Miller (1956) on the magical number 7 ± 2, which emphasized the process of grouping elements together to form larger meaningful units or "chunks." The arrangement of telephone numbers with groups of three and then four digits would not appear to be accidental, but rather an indication of how many elements can be comfortably held in the focus of attention at one time to allow the formation of a new chunk in long-term memory (Baars 1988; Cowan 1995). Several investigators have shown that short-term memory performance is best when items are grouped into sublists of no more than three or four (Broadbent 1975; Ryan 1969; Wickelgren 1964).

The grouping limit seems to apply even for subjects who have learned how to repeat back strings of 80 or more digits (Ericsson 1985; Ericsson et al. 1980). These subjects did so by learning to form meaningful chunks out of small groups of digits, and then learning to group the chunks together to form "supergroups." At both the group and the supergroup levels, the capacity limit seems to apply, as described by Ericsson et al. (1980, p. 1182) for their first subject who increased his digit span greatly: "After all of this practice, can we conclude that S.F. increased his short-term memory capacity? There are several reasons to think not . . . The size of S.F.'s groups were almost always 3 and 4 digits, and he never generated a mnemonic association for more than 5 digits . . . He generally used three groups in his supergroups and, after some initial difficulty with five groups, never allowed more than four groups in a supergroup." Ericsson (1985) reviewed details of the hierarchical grouping structure in the increased-digit-span subjects and he reviewed other studies of memory experts, which also revealed a similar grouping limit of 3–5 items. This limit to the grouping process would make sense if the items or groups to be further grouped together must reside in a common, central workspace so that they can be linked. A limited-capacity store might be conceived in this way, as a workspace in which items are linked together (Baars 1988; Cowan 1995). The focus of attention, the presumed locus of limited-capacity storage, presumably must shift back and forth from the super-group level to the group level.

3.4.2. Cluster size in long-term recall. As Shiffrin (1993) pointed out, in a sense every cognitive task is a STM task because items must be active in a limited-capacity store at the time of recall, even though sometimes that can come about only through the reactivation of long-term memories. Assume, therefore, that it is necessary to represent items in the limited-capacity store to prepare them to be recalled. That sort of mechanism should apply not only to immediate recall, but also to long-term recall. Bursts of responses should be observed as an individual fills the limited short-term capacity, recalls the items held with that capacity, and then refills it with more information retrieved from long-term storage.

Broadbent (1975) used similar reasoning to motivate a study of grouping in long-term recall. He asked subjects to recall members of a learned category: the Seven Dwarfs, the seven colors of the rainbow, the countries of Europe, or the names of regular television programs. There were some important differences in the fluency of recall for these categories. However, measurement of the timing of recall showed bursts of 2, 3, or 4 items clustered together, and occasionally 5 items in a cluster. One of the 10 subjects produced a run of 6 rainbow colors, but otherwise the cluster sizes were below 6. Thus, this study provides evidence for a short-term memory capacity limit even as applied to recall from a category in long-term memory. Wilkes (1975) reviewed similar results in a more detailed study of pausing within recall. Bower and Winzenz (1969, cited in Wilkes 1975) showed that repetitions of a digit string within a longer series resulted in improved recall over time only if the grouping structure did not change each time the repetition occurred.

Mandler (1967) suggested that the size of the limited-capacity store is 5 ± 2 items. He focused on a number of experiments in which the items to be recalled could be di-

vided into a number of categories (e.g., fruits, clothing, vehicles). Recall is superior when a list is recalled by categories; in a free recall task, subjects tend to recall clusters organized by category even when the items were not ordered by category in the stimulus list. Mandler suggested that subjects left to their own devices typically could recall 5 ± 2 categories and 5 ± 2 items from each category that was recalled (e.g., "apple, orange, banana, grape; shirt, pants, hat," and so on). (Mandler also relied on Tulving & Patkau 1962; and Tulving & Pearlstone 1966.)

The recall of separate items from a higher-level category in memory might be considered analogous to the retrieval of separate items from an array represented in sensory memory. In the case of the sensory array, the items are related in that they all are part of a common visual field, but are nevertheless separate perceptually (assuming they are not organized into larger perceptual objects such as a cluster of rounded letters surrounded by angular letters). In memory, the items are related in that they are associated to a common semantic category, but are nevertheless separate conceptually (assuming they are not organized into clusters such as salt-and-pepper within the spice category). The clustering of items in recall presumably depends on the absence of an automatized routine for recall. Thus, one should not expect clustering into groups of about four items for a task like recitation of the alphabet. The same applies to extensive intra-category knowledge that can result in the recall of large chunks structures or templates (Gobet & Simon 1996; 1998).

When one focuses on flawless recall, the number is closer to three or four. Thus, Mandler's (1967, Fig. 7) summarization of recall per category shows that when there were only 1–3 items in a category, these items were recalled flawlessly (provided that the category was recalled at all). The number recalled within the category declined rather steadily thereafter, from about 80% recalled from categories with 4–6 items to about 20% recalled from categories of about 80 items. The growing absolute number of items recalled from larger categories might be attributed to covert subcategorization or to long-term learning mechanisms, but these will not allow the recall of all items.

Why should similar constraints apply to the recall of items within a category and to the recall of categories? One explanation is that the limited-capacity store can be used in more than one iteration. At one moment the categories are drawn into mind, and at the next moment the capacity is consumed with items within the first category to be recalled. An obvious question about such an account is how the categories are retained while the limited capacity is being used for items within a category. Presumably the immediate consequence of the limited capacity store is to form a better-organized long-term representation of the stimulus set. Thus, once a set of categories is brought into mind, these categories are combined into a long-term memory representation that can be accessed again as needed in the task. Ericsson and Kintsch (1995) provided a detailed account of that sort of process.

Finally, it should be noted that the capacity limit provides only an upper bound for the clustering of items in recall. If the rate of retrieval from memory is too slow, it might make sense for the individual to recall the items deposited in a capacity-limited store before that capacity has been used up. Gruenewald and Lockhead (1980) obtained a pattern of results in which the long-term recall of category exemplars occurred (according to their criteria) most often without clustering and in clusters of two, three, four, or five in decreasing order of frequency. In contrast, Graesser and Mandler (1978, p. 96) observed a limit in items per cluster hovering around 4.0 throughout a long recall protocol. A prediction of the present analysis is that if a particular procedure results in a limit smaller than about four items, extended practice with the task should eventually lead to a plateau in performance at about four items per cluster.

3.4.3. Positional uncertainty in recall. In a theoretical mechanism discussed above and used earlier by Raaijmakers and Shiffrin (1981), items held simultaneously in the limited-capacity store become associated with one another. Additionally, their serial positions may become associated with one another. In free recall, the associations can be helpful because the thought of one item elicits the associated items. However, in serial recall, the associations can present a problem because it may be difficult to retrieve the order of simultaneously held items. This type of account might explain positional uncertainty in serial recall. It predicts that an item typically should be recalled no more than three positions away from its correct position (assuming that sets of at most four items are present in limited-capacity storage at any moment). This prediction matches the data well. For example, Nairne (1991) presented words aloud with 2.5-sec onset-to-onset times between the words. Five lists of 5 words each were followed by a 2-min distractor task and then a test in which the words were to be placed into their correct lists and locations within lists. The results showed that when an item was placed within the correct list, its serial position was confused with at most three other serial positions in the list. Nairne (1992) found a flattening of the error functions with delayed testing. This change over time is compatible with increased difficulty of accessing information from a particular prior state of the limited-capacity store, but with no evidence of a spread of uncertainty to a larger range of serial positions. Within-list confusions still occurred across about three items.

3.4.4. Analysis of the recency effect in recall. Watkins (1974) reviewed research in which a long list of verbal items was presented on each trial and the subject was to recall as many of those items as possible, without regard to the order of items. In such studies, recall is typically best for the most recent items. This recency effect has been viewed as the result of the use of dual memory mechanisms, with a short-term memory mechanism used only for the last few items (which typically are recalled first). Underlying this view is the finding that the recency effect is quickly eroded if a distractor-filled delay is imposed between the list and the recall cue (Glanzer & Cunitz 1966; Postman & Phillips 1965), whereas the rest of the recall function is unaffected. Several investigators have reasoned that it would be possible to estimate the contents of short-term memory by subtracting out the contribution of long-term memory, but it is not clear exactly what assumptions one should make in order to do so. Watkins ruled out some methods on the basis of logical considerations, and compared the results of several favored methods (Tulving & Colotla 1970; Tulving & Patterson 1968; Waugh & Norman 1965). Under a variety of test situations, these methods produced estimates of short-term memory capacity ranging from 2.21 to 3.43 (Watkins 1974, Table 1). For the method judged most ade-

quate (Tulving & Colotla 1970) the estimates ranged from 2.93 to 3.35.

One apparent difficulty for this interpretation of the recency effect is that one can obtain it under filled test delays that are too protracted to permit the belief that a short-term store is still in place. Bjork and Whitten (1974) presented a series of printed word pairs for immediate free recall, with the pairs separated by silent intervals of 12 sec and with up to 42 sec following the last item pair. These silent intervals were filled with a distracting arithmetic task to prevent rehearsal. Under these conditions, a recency effect of 3–5 items emerged. The theoretical framework for understanding these results, further developed by Glenberg and Swanson (1986), was one having to do with the ratio between the inter-item interval and the retention interval, which could influence the temporal distinctiveness of the items at the time of recall. The temporal distinctiveness is higher under Bjork and Whitten's conditions for the last few item pairs than for previous item pairs, and the same can be said of the last few list items within the immediate recall conditions that Watkins (1974) had considered. Although the long-term recency effect challenges a time-limited memory explanation of the recency effect, it need not challenge a temporary memory capacity limit. A capacity-limited store could work in combination with the distinctiveness principles. The recall process could proceed in phases, each of which may involve the subject scanning the memory representation, transferring several items to the capacity-limited store, recalling those items, and then returning to the representation for another limited-capacity "handful" of items. It would make sense for the subject to recall the most recent, most distinctive items in the first retrieval cycle so as to avoid losing the distinctiveness advantage of those items. Assuming that a capacity-limited store (presumably the focus of attention) must intervene between a memory representation and recall, it is consistent with the long-term recency effect.

3.4.5. Sequential effects in implicit learning and memory. Implicit learning is a process in which information is learned without the awareness of the subject, and implicit memory is learned information that is revealed in an indirect test, without the subject having been questioned explicitly about the information. The role of a limited-capacity store in implicit learning and memory exists, though its nature is not yet clear (see Frensch & Miner 1994; Nissen & Bullemer 1987; Reber & Kotovsky 1997). It is possible that the role of a limited-capacity store depends on whether learning and memory require associations between items that go beyond a simple chain of association between adjacent items. There are data supporting this conjecture and suggesting that implicit learning can take place if one need hold no more than four items in the capacity-limited store at one time.

Lewicki et al. (1987) examined one situation in which contingencies were spread across seven trials in a row, but the capacity-limited store need not encompass all of those items. Lewicki et al. presented sets of seven trials in a row in which the subject was to indicate which quadrant of the screen contained the target item, the digit 6 (using a keypress response). The first six trials in the set included only the target, but the seventh trial also included 35 foils distributed around the screen. Moreover, unbeknownst to the subject, the locations of targets on Trials 1, 3, 4, and 6, taken together, indicated where the target would be on the seventh, complex trial. Under these circumstances, subjects succeeded in learning the contingencies and there was a drop in performance when the contingencies were changed. However, as Stadler (1989) pointed out, any three of the four critical trials were enough to determine the location of the target on Trial 7 in a set. If subjects remembered the outcomes of Trials 3, 4, and 6 and considered them together, they theoretically could predict the outcome of Trial 7. Given that subjects probably did not know which trials were predictive, they might only be able to use Trials 4, 5, and 6, the last three trials, for prediction of Trial 7. These trials by themselves were predictive on a probabilistic, though not an absolute, basis. Thus, a limited-capacity store of four items could serve for this purpose. Stadler (1989) extended the result and verified that the learning was implicit and not available to subjects' awareness.

Cleeremans and McClelland (1991) demonstrated more precisely that a sequence of four items in limited-capacity storage at one time may be enough to allow learning of the contingencies between those items. The task was to press one of six keys corresponding to stimuli at six screen locations. These locations were stimulated according to a finite state grammar on 85% of the trials, of which subjects were unaware. However, on the remaining, randomly selected 15% of the trials, the expectations according to the grammar were violated. The nature of the grammar was such that a prediction could be made only if one took into account a sequence of several previous stimuli. A sequential analysis of the reaction times showed that, after 13 experimental sessions, subjects became able to use a series of three stimuli to predict the location of the next, fourth stimulus. Even after 20 sessions, though, they remained unable to use a series of four stimuli to predict the location of the next, fifth stimulus. Thus, four seems to be the asymptotic value. (The authors presented a different theoretical account based on the diminishing predictive value of remote associations. However, the two accounts may not be mutually exclusive.)

McKone (1995) demonstrated a capacity limit in the sequences that can contribute to repetition priming in lexical decision or word naming. Series of words or nonwords were presented and there were repetitions of items with a variable number of different items intervening between the two instances of the repeated word (ranging from 0 to 23 intervening items). The measure of priming was a decrease in reaction time for the repeated word, suggesting that the representation of the first instance of that word was still active in memory. McKone concluded (p. 1108) that "for words, a large short-term priming component decayed rapidly but smoothly over the first three items" intervening between the instances of the repeated word, and then reached "a stable long-term value." This appears to be evidence of a series of about four consecutive items present in a limited-capacity store at any time, though a long-term store also contributes to priming as shown by the asymptotic level of residual priming at longer repetition lags.

An unresolved question stemming from McKone (1995) is why the priming declined smoothly over the last few items. When the presentation is sequential and there is no deliberate effort to retain any but the current item, as in this study, it is possible that the more recent items tend to be more strongly represented in the limited-capacity store. It also is possible that some of the most recent four items

sometimes are replaced in the limited-capacity store by items from elsewhere in the list or from extraneous thoughts.

3.4.6. Influence of capacity on the properties of visual search. Fisher (1984) observed what appears to be a limit in the ability to examine items in a visual array in search of a well-learned target item. In previous work, Shiffrin and Schneider (1977) and Schneider and Shiffrin (1977) distinguished between variably-mapped searches (in which a foil on one trial might become the target on another trial) and consistently-mapped searches (in which the items that serve as targets never serve as foils). On each trial in their experiments, the subject knows what target or targets to search for, and indicates as rapidly and accurately as possible the presence or absence of the target item(s) in a visual array of items. It was proposed that variably-mapped searches require capacity-limited or "controlled" processing, whereas over many trials, a consistently-mapped search task comes to be performed automatically, without using controlled processing. Under those circumstances, it was shown that processing took place on all items in parallel, so that the reaction time to detect a target item was nearly unaffected by the number of items in the array.

Fisher (1984) proposed that capacity limits still might appear if the required rate of perceptual processing were fast enough. He re-examined this assumption using a task in which there were 20 stimulus arrays in rapid succession on each trial (with a pattern mask preceding the sequence of 20 and another following it). The duration of each array varied across trial blocks (with 10 durations between 40 and 200 msec per array), as did the array size (with 4 or 8 stimuli per array). The arrays were letters except for a single target item within one array, which was always the digit "5." The task was to indicate the spatial location of this target item out of eight possible locations. This is a type of consistently-mapped task situation in which little practice is needed to achieve an automatic search because the digit and letter categories have been learned before the experiment. The results were analyzed in light of a mathematical model, the "steady-state limited-channel model," based on the following defining assumptions (Fisher 1984, p. 453):

> (1) Encoded stimuli in the visual cortex are scanned once for placement on a comparison channel; (2) the time between arrivals of stimuli to the comparison channels is exponentially distributed with rate parameter *l;* (3) the time to compare a stimulus with a prespecified target is exponentially distributed with rate parameter *m;* (4) at most *k* comparison channels can execute in parallel; (5) stimuli in iconic memory are equally likely to be replaced by the characters or masks which appear next to the input streams; (6) masks are not placed on the comparison channels; and (7) the system is in a steady state. Note that it is assumed that the two dimensional coordinates of a stimulus are retained in the visual cortex.

The critical idea is that the representation of a stimulus is lost if that stimulus is offered to the comparison process at a time when there are no comparison channels free. Thus, the capacity limit in this situation is defined by the number of comparison channels, *k.* Within the field of queuing theory, Erlang's Loss Formula describes the problem and, using that formula, Fisher found that the data fit the formula best with $k = 4$. Assuming that the comparison channels are actually slots within some short-term storage mechanism, this result serves as another indication of its limited-capacity nature (also see Schweickert et al. 1996).

3.4.7. Influence of capacity on mental addition reaction time. Logan (1988, Experiment 4) developed a task in which subjects had to verify equations like "B + 3 = E" (true in this example because E is 3 letters after B in the alphabet). The addend could be 2, 3, 4, or 5. Practice effects in this novel task followed a power function for addends of 2 through 4. However, for an addend of 5, the fit was much worse, and there was a discontinuity in the learning curves after about 24 presentations in which the reaction times for this addend dropped sharply. This discontinuity was linked to a strategy shift that many subjects reported. They reported that problems with an addend of 5 were much harder and led to a more deliberate learning strategy in which particular instances were memorized. Logan and Klapp (1991) replicated this finding. It might be further speculated that the discontinuity could occur because the numbers 1–4 can be visualized more clearly during the problem, serving as a place-holder during the adding process. Numbers of 5 and above may be difficult because the visualization of items to be added is hindered by the capacity limitation.

3.4.8. Mathematical modeling parameters. Attesting to the potential importance of the pure capacity limit, various articles presenting mathematical models of various complex cognitive processes have used the assumption that four items can be saved in a short-term store. These include, for example, the Kintsch and van Dijk (1978) model of text comprehension, the Raaijmakers and Shiffrin (1981) model of memory search (SAM), and the recent model of processing capacity by Halford et al. (1998). These models presumably use a limit of four because it maximizes the ability of the model to fit real data.

3.5. Empirical summary

In this review, care was taken to exclude situations in which the chunking of items was unclear (yielding compound STM estimates). The results of the surviving experimental situations, a wide variety of situations in fact, suggest that about four chunks can be held in a pure capacity-limited STM (presumably the focus of attention). The experimental means for groups of adults generally range from about 3 to 5 chunks, whereas individual subject means range more widely from 2 to 6 chunks.

3.6. Testability of the theoretical analysis

There are several ways in which, in future research, one could invalidate the capacity estimates of 3 to 5 chunks that have been derived from the many theoretical phenomena described above. First, one could show that performance in these studies results from a grouping process in which multiple items contribute to a chunk. In many cases the argument against this was limited to a theoretical rationale why such chunking should be absent in particular circumstances (see sect 1.2); few studies actually have provided direct evidence of chunk size. Second, one could develop conditions in which more is done to limit chunking and find smaller capacity estimates. Third, one could find that there are hidden storage demands and that, when they are eliminated, substantially larger capacity estimates arise (e.g., $>$ 6 chunks). Fourth, in a different vein, one could find low or zero correlations between different estimates of storage

capacity despite large individual differences in the estimates. This last finding would not necessarily challenge the notion that there are fixed capacity limits in a given domain, but it would challenge the concept of a central capacity mechanism (e.g., the focus of attention) that is the seat of the capacity limit across all domains.

4. Theoretical account of capacity limits: Unresolved issues

Below, I will address several fundamental theoretical questions about capacity limits. (1) Why does the capacity limit occur? (2) What is the nature of this limit: Is there a single capacity limit or are there multiple limits? (3) What are the implications of the present arguments for alternative theoretical accounts? (4) Finally, what are the boundaries of the central-capacity-limit account? An enigma in Miller (1956), regarding absolute judgments, will be touched upon to examine the potential breadth of the present framework.

4.1. Why the capacity limit?

Future research must establish why there is a limit in capacity. One possible reason is that the capacity limit has become optimized through adaptive processes in evolution. Two relevant teleological arguments can be made on logical grounds. Recently, as well, arguments have been made concerning the physiological basis of storage capacity limits. Any such account must consider details of the data including the individual variability in the capacity limit estimates that have been observed as discussed in section 3.1.3. These issues will be addressed in turn.

4.1.1. Teleological accounts. Several investigators have provided mathematical arguments relevant to what the most efficient size of working memory would be. Dirlam (1972) asked if there is any one chunk size that is more efficient than any other if it is to be the basis of a memory search. He assumed that STM is a multi-level, hierarchically structured system and that the search process is random. The nodes at a particular level are searched only until the correct one is found, at which time the search is confined to subnodes within that node. This process was assumed to continue until, at the lowest level of the hierarchy, the searched-for item is identified. In other words, the search was said to be self-terminating at each level of the hierarchy. Dirlam then asked what rule of chunking would minimize the expected number of total node and item accesses regardless of the number of items in the list, and calculated that the minimum would occur with an average chunk size of 3.59 items at each level of the hierarchy, in close agreement with the capacity of short-term memory that has been observed empirically in many situations (see above).

MacGregor (1987) asked a slightly different question: What is the maximal number of items for which a one-level system is more efficient than a two-level system? The consequences of both self-terminating search and exhaustive search assumptions were examined. A concrete example would help to explain how. Suppose that one received a list that included eight items. Further suppose that one had the option of representing this list either in an unorganized manner or as two higher-level chunks, each containing four items. With a self-terminating search method, if one had to search for a particular letter in the unorganized list, one would search on the average through 4.5 of the items (the average of the numbers 1 through 8). If one had to search through the list organized into two chunks, one would have to search on the average through 1.5 chunks to find the right chunk and then an average of 2.5 items within that chunk to find the right item, or 4.0 accesses in all. On the average, the hierarchical organization would be more efficient. With an exhaustive search method, if one had to search for a particular letter in the unorganized list, one would have to search through eight items. For the organized list, one would need two searches to find the right chunk and then four searches to find the right item within that chunk, or six accesses in all. On the average, again, the organized list would be more efficient. In contrast, consider a self-terminating search for a list of four items that could be represented in an unorganized manner or as 2 chunks of 2 items each. The unorganized list would require an average of 2.5 searches whereas the organized list would require that 1.5 clusters and 1.5 items in that cluster be examined, for a total of 3.0 searches. In this case, the unorganized list is more efficient on average. MacGregor calculated that organizing list items into higher-level chunks is beneficial with an exhaustive or a self-terminating search when there are more than 4 or 5.83 items, respectively.

Although these theoretical findings depend on some untested assumptions (e.g., that the difficulty of search is the same at every level of a hierarchy), they do provide useful insight. The empirically observed capacity limit of about four chunks corresponds to what has been predicted for how many items can be advantageously processed in an ungrouped manner when the search is exhaustive (MacGregor 1987). These theoretical and empirical limits may correspond because very rapid searches of unorganized lists are, in fact, exhaustive (Sternberg 1966). However, slower, self-terminating searches along with more elaborate mental organization of the material also may be possible, and probably are advantageous if there is time to accomplish this mental organization. That possibility can help to explain why the empirically observed limit of about four chunks is close to the optimal chunk size when multiple levels of organization are permitted in a self-terminating search (Dirlam 1972).

Another teleological analysis can be formulated on the basis of Kareev (1995). He suggested that a limited working memory is better than an unlimited one for detecting imperfect correlations between features in the environment. To take a hypothetical example, there could be a population with a 70% correlation between the height of an individual and the pitch of his or her voice. In a statistical procedure, when one uses a limited sample size to estimate the correlation (e.g., an observation of 4–8 individuals), the modal value of the observed correlation is larger than the population value. The smaller the sample size, the higher the modal value. Thus, a smaller sample size would increase the chances that a moderate correlation would be noticed at all. In human information processing, the limit in the sample size could be caused by the capacity limit of the observer's short-term memory; more samples may have been observed but the observer bases his or her perceived estimate of the correlation on only the number of examples that fit into the focus of attention at one time. Kareev et al. (1997) showed that, in fact, low-working-memory subjects were more likely to notice a population correlation of .2–.6.

In this regard, it bears mention that in the statistical sampling procedure, the modal value of the sample correlations for sample sizes of 6 and 8 were shown to be only moderately greater than the true population value (which was set at .6 or .7); but for a sample size of 4, the modal value of the sample correlations was almost 1.0. Here, then, is another reason to believe that a basic capacity limit of four could be advantageous. It could take a moderate correlation in the real world and turn it into a perceived strong correlation. At least, this could be advantageous to the extent that decisiveness in decision-making and definiteness in the perception of stimulus relationships are advantageous. For example, it makes sense to walk away from someone displaying traits that are moderately correlated with injurious behavior, and it makes sense to perceive that people usually say please when they are asking for a favor.

There is a strong similarity between the theoretical analysis of Kareev and earlier proposals that a large short-term memory capacity can be a liability rather than a strength in the early stages of language learning. Newport (1990) discussed a "less is more" hypothesis to explain why language learners who are developmentally immature at the time of initial learning have an advantage over more mature learners for some language constructs. An alternative to the nativist theory of language learning, this theory states that immature language learners grasp only small fragments of language at a time, which helps them to break up a complex language structure into smaller parts. Consistent with this proposal, Elman (1993) found that a computer implementation of a parallel distributed processing model of cognition learned complex language structure more easily if the short-term memory capacity of the model started out small and increased later in the learning process, rather than taking on its mature value at the beginning of learning.

Below, neurophysiological accounts of capacity limits will be reviewed. The teleological arguments still will be important to the extent that they can be seen as being consistent with the physiological mechanisms underlying capacity limits (or, better yet, motivating them).

4.1.2. Neurophysiological accounts. In recent years, a number of investigators have suggested a basis of capacity limits that can be traced back to information about how a single object is represented in the brain. In a theoretical article on visual shape recognition, Milner (1974, p. 532) suggested that "cells fired by the same figure fire together but not in synchrony with cells fired by other figures . . . Thus, features from a number of figures could be detected and transmitted through the network with little mutual interference, by a sort of time-sharing arrangement." In support of this hypothesis, Gray et al. (1989), in an experiment on cats, found that two columns of cortical cells that represented different portions of the visual field were active in a correlated manner only if they were stimulated by different portions of the same object, and not if they were stimulated by different objects. This led to the hypothesis that the synchronization of activity for various features represents the binding of those features to form an object in perception or STM. More recently, these findings have been extended to humans. Tiitinen et al. (1993) found that the 40-Hz oscillatory cycle, upon which these synchronizations are thought to ride, is enhanced by attention in humans. Rodriguez et al. (1999) reported electrical synchronizations between certain widely separated scalp locations 180–360 msec after a stimulus was presented when an object (a silhouetted human profile) was perceived, but not when a random field (actually an upside down profile not detected as such) was perceived. The scalp locations appeared to implicate the parietal lobes, which Cowan (1995) also proposed to be areas involved in the integration of features to form objects. Miltner et al. (1999) further showed that the binding can take place not only between perceptual features, but also between a feature and an activated mental concept. Specifically, cyclic activity in the gamma (20–70 Hz) band was synchronized between several areas of the brain in the time period after the presentation of a conditioned stimulus (CS+), a color illuminating the room, but before the presentation of the unconditioned stimulus (UCS), an electric shock that, as the subjects had learned, followed the conditioned stimulus. No such synchronization occurred after a different color (CS-) that did not lead to electric shock.

If objects and meaningful events can be carried in the synchronized activity of gamma wave activity in the brain, then the question for STM capacity becomes, "How many objects or events can be represented simultaneously in the brain?" Investigators have discussed that. Lisman and Idiart (1995) suggested that "each memory is stored in a different high-frequency ('40 Hertz') subcycle of a low-frequency oscillation. Memory patterns repeat on each low-frequency (5 to 12 Hertz) oscillation, a repetition that relies on activity dependent changes in membrane excitability rather than reverberatory circuits." In other words, the number of subcycles that fit into a low-frequency cycle would define the number of items that could be held in a capacity-limited STM. This suggestion was intended by Lisman and Idiart to motivate the existence of a memory span of about seven items (e.g., [40 subcycles/sec] / [5.7 cycles/sec] = 7 subcycles/cycle). However, it could just as well be used to motivate a basic capacity of about 4 items (e.g., [40 subcycles/sec]/[10 cycles/sec] = 4 subcycles/cycle). This proposal also was intended to account for the speed of retrieval of information stored in the capacity-limited STM but, again, just as well fits the 4-item limit. If 40 subcycles occur per second then each subcycle takes 25 msec, a fair estimate of the time it takes to search one item in STM (Sternberg 1966). Luck and Vogel (1998) made a proposal similar to Lisman and Idiart but made it explicit that the representation of each item in STM would involve the synchronization of neural firing representing the features of the item. The STM capacity limit would occur because two sets of feature detectors that fire simultaneously produce a spurious synchronization corrupting memory by seeming to come from one object.

Other theorists (Hummel & Holyoak 1997; Shastri & Ajjanagadde 1993) have applied this neural synchronization principle in a way that is more abstract. It can serve as an alternative compatible with Halford et al.'s (1998) basic notion of a limit on the complexity of relations between concepts, though Halford et al. instead worked with a more symbolically based model in which "the amount of information that can be represented by a single vector is not significantly limited, but the number of vectors that can be bound in one representation of a relation is limited" (p. 821). Shastri and Ajjanagadde (1993) formulated a physiological theory of *working memory* very similar to Lisman and Idiart (1995), except that the theory was meant to explain "a limited-capacity dynamic working memory that temporarily holds information during an episode of reflexive reasoning" (p. 442), meaning reasoning that can be car-

ried out "rapidly, spontaneously, and without conscious effort" (p. 418). The information was said to be held as concepts or predicates that were in the form of complex chunks; thus, it was cautioned, "note that the activation of an entity together with all its active superconcepts counts as only one entity" (p. 443). It was remarked that the bound on the number of entities in working memory, derived from facts of neural oscillation, falls in the 7 ± 2 range; but the argument was not precise enough to distinguish that from the lower estimate offered in the present paper. Hummel and Holyoak (1997) brought up similar concepts in their theory of thinking with analogies. They defined "dynamic binding" (a term that Shastri & Ajjanagadde also relied upon to describe how entities came about) as a situation in which "units representing case roles are temporarily bound to units representing the fillers of those roles" (p. 433). They estimated the limit of dynamic binding links as "between four and six" (p. 434). In both the approaches of Shastri and Ajjanagadde (1993) and Hummel and Holyoak (1997), these small limits were supplemented with data structures in long term memory or "static bindings" that appear to operate in the same manner as the long-term working memory of Ericsson and Kintsch (1995), presumably providing the "active superconcepts" that Shastri and Ajjanagadde mentioned.

One problem for the interpretation of synchronous oscillations of nervous tissue is that they can be observed even in lower animals in situations that appear to have little to do with the possibility of conscious awareness of particular stimuli (e.g., Braun et al. 1994; Kirschfeld 1992). This, in itself, need not invalidate the role of oscillations in binding together the features of an object or the objects in a capacity-limited store in humans. It could be the case that mechanisms already present in lower animals form the basis of more advanced skills in more advanced animals, just as the voice apparatus is necessary for speech but is present even in non-speaking species. Thus, von der Malsburg (1995, p. 524) noted that "As to the binding mechanism based on temporal signal correlations, its great advantage [is] being undemanding in terms of structural requirements and consequently ubiquitously available and extremely flexible."

4.1.3. Reconciliation of teleological and neurophysiological accounts. One concern here is whether the teleological and physiological accounts of capacity limits are consistent or inconsistent with one another. The process of scanning through the items in STM has been employed theoretically by both the teleological and the physiological theorists. For example, the teleological argument that MacGregor (1987) built using an exhaustive scan resulted in the conclusion that the scan would be most efficient if the number of items per group were four. This conclusion was based on the assumption that the amount of time it takes to access a group to determine whether a particular item is present within it is equal to the amount of time it then takes to access each item within the appropriate group once that group is selected, so as finally to identify the probed item. This concept can be mapped directly onto the concept of the set of items (or chunks) in capacity-limited STM being represented by a single cycle of a low-frequency oscillation (5 to 12 Hz) with each item mapped onto a different cycle of a 40-Hz oscillation, riding on top of the 5 to 12 Hz oscillation. These figures are in line with the teleological data and memory capacity data reviewed above if the rate for the slow oscillation is close to about 10 Hz, so that four items would fit in each of the slower cycles. As suggested by Shastri and Ajjanagadde (1993) and others, the cyclic search process could be employed recursively. For example, at one point in a probed recognition process there could be up to four chunks in the capacity-limited store. Once the correct chunk is identified, the contents of STM would be replaced by the items contained within that chunk, now "unpacked," so that the contents of the chunk can be scanned in detail. In present theoretical terms, the focus of attention need not focus on multiple levels of representation at the same time.

4.1.4. What is the basis of individual differences? We will not have a good understanding of capacity limits until we are able to understand the basis of the marked developmental and individual differences in measured capacity that were observed by Cowan et al. (1999) and comparable individual differences observed in other procedures (Henderson 1972; Luck & Vogel, personal communication, January 18, 1999). One possible basis would be individual differences in the ratio of slow to fast oscillatory rhythms. Miltner et al. (1999) found most rapid oscillatory activity at 37–43 Hz, but some residual activity at 30–37 and 43–48 Hz. One can combine a 12-Hz slow cycle with a 30-Hz rapid cycle to predict the low end of the range of memory capacities (12/30 = 2.5 items), or one can combine an 8-Hz slow cycle with a 48-Hz fast cycle to predict the high end of the range (8/48 = 6 items). According to these figures, however, one would not expect the slow cycle to go below 8 Hz, given the capacity limits observed empirically. Here, then, is a physiological prediction based on a combination of existing physiological and behavioral results. An important next step may be the acquisition of data that can help to evaluate the psychological plausibility of the theoretical constructs surrounding this type of theory. As one promising example, the finding of Tiitinen et al. (1993) that the 40-Hz neural cycle is enhanced by attention is consistent with the present suggestion that the fundamental storage capacity limit of about four items is based on the 40-Hz cycle and is in essence a limit in the capacity of the focus of attention. It is easy to see how research on this topic also could clarify the basis of individual differences in capacity. Specifically, one could determine if individual differences in oscillatory rates mirror behavioral differences in the limited storage capacity.

It remains to be explained why attended speech shows such an intriguing, simple relationship to unattended speech (Fig. 4) in which attended speech is increased above the unattended speech limit by a variable amount. This figure makes it apparent that individuals use the same processes in both conditions, plus supplementary processes for attended speech. This difference might be accounted for most simply by the process of chunking (formation of inter-item associations) during attended list presentations.

It is important not to become too reductionistic in the interpretation of biological effects. It is possible that stimulus factors and/or behavioral states modulate biological cycle frequencies under some circumstances. Some studies with an automatized response or a rapid response have resulted in smaller individual differences. The highly trained subjects of Sperling (1960) appeared to produce capacity (whole report) estimates deviating from the population mean by no more than about 0.5 items, although there were few subjects. In an enumeration task in which a reaction

time measure defined the subitizing range, Chi and Klahr (1975) found no difference between 5- and 6-year-olds versus adults in the subitizing range. Perhaps there is an intrinsic, baseline capacity of the focus of attention that shows few differences between individuals, and perhaps under some circumstances but not others, the level and direction of effort at the time of recall modulate that capacity. Further study of individual differences in memory capacity is thus likely to be important theoretically.

4.2. Central capacity or separate capacities?

In most of the research that I have discussed, the capacity limit is examined with a coherent field of stimulation. I have not directly tackled the question of whether there is one central capacity limit or whether there are separate limits for domains of cognition (e.g., separate capacities for the visual versus auditory modalities; for verbal versus spatial representational codes; or, perhaps, for two collections of items distinguished by various other physical or semantic features). According to the models of Cowan (1988; 1995) and Engle et al. (1999), the capacity limit would be a central one (the focus of attention). Some fine points must be kept in mind on what should count as evidence for or against a central limit.

Ideally, evidence for or against a central capacity limit could be obtained by looking at the number of items recalled in two tasks, A and B, and then determining whether the total number of items recalled on a trial can be increased by presenting A and B together and adding the number of items recalled in the two tasks. For example, suppose that one can recall three items in Task A and four items in Task B, and that one can recall six items all together in a combined, A + B task. Although performance on the component tasks is diminished when the tasks are carried out together, the total number of items recalled is greater than for either task presented alone. This savings would serve as initial evidence for the existence of separate storage mechanisms (with or without an additional, central storage mechanism). Further, if there were no diminution of performance in either task when they were combined, that would serve as evidence against the central storage mechanism or capacity limit.

This type of reasoning can be used only with important limitations, however. As discussed above, several different mechanisms contribute to recall, including not only the capacity-limited focus of attention, but also the time- or interference-limited sources of activation of long-term memory (sensory stores, phonological and spatial stores, etc.). *If the focus of attention could shift from examining one source of activation to examining another dissimilar source,* it would be possible to recall items from Task A and then shift attention to activated memory representations of the items in Task B, bringing them into the focus of attention for recall in turn. If all of the information need not be entered into the focus of attention at one time, performance in the combined task would overestimate central storage capacity. This possibility contaminates many types of evidence that initially look as if they could provide support for multiple capacity-limited stores. These include various studies showing that one can recall more in two tasks with different types of materials combined than in a single task, especially if the modalities or types of representations are very different (Baddeley 1986; Frick 1984; Greene 1989; Henderson 1972; Klapp & Netick 1988; Luck & Vogel 1997; Martin 1980; Penney 1980; Reisberg et al. 1984; Sanders & Schroots 1969; Shah & Miyake 1996).

Theoretically, it should be possible to overcome methodological problems in order to determine if there are true multiple capacity limits. One could make it impossible for the subject to rehearse items during presentation of the materials by using complex arrays of two types concurrently; perhaps concurrent visual and auditory arrays. It would also be necessary to make sure that the focus of attention could not be used recursively, shifting from one type of activated material to the next for recall. If the activated representations were sensory in nature, this recursive recall might be prevented simply by backward-masking one or both of the types of materials. These requirements do not seem to have been met in any extant study. Martin (1980) did use simultaneous left- and right-sided visual and auditory lists (4 channels at once, only 2 of them meaningful at once, with sequences of 4 stimuli presented at a fast, 2/sec rate on each of the 4 channels). She found that memory for words presented concurrently to the left and right fields in the same modality was, on the average, 51.6% correct, whereas memory for pairs containing one printed and one spoken word was 76.9% correct. However, there was nothing to prevent the shifting of attention from visual to auditory sensory memory in turn.

Another methodological possibility is to document the shifting of attention rather than preventing it. This can be accomplished with reaction time measures. One enumeration study is relevant. Atkinson et al. (1976b) presented two sets of dots separated by their organization into lines at different orientations, by two different colors, or by their organization into separate groups. Separation by spatial orientation or grouping was capable of eliminating errors when there was a total of five to eight dots. Color separation reduced, but did not eliminate, errors. However, the grouping did not reduce the *reaction times* in any of these studies. It seems likely that some sort of apprehension process took place separately for each group of four or fewer dots and that the numbers of dots in each group were then added together. Inasmuch as the reaction times were not slower when the items were grouped, one reasonable interpretation is that subitizing in groups and then adding the groups is the normal enumeration process for fields of five or more dots, even when there are no physical cues for the groups. The addition of physical cues simply makes the subitizing process more accurate (though not faster). This study provides some support for Mandler's (1985) suggestion that the capacity limit is for sets of items that can be combined into a coherent scheme. By dividing the sensory field into two coherent, separable schemes, the effective limit can be increased; but different schemes or groups can become the limit of attention only one at a time, explaining why perceptual grouping cues increase accuracy without altering the reaction times.

Physiological studies also may help if they can show a reciprocity between tasks that do not appear to share specific processing modes. One study using event-related potentials (ERPs) by Sirevaag et al. (1989) is relevant. It involved two tasks with very little in common, both of which were effortful. In one task, the subject controlled a cursor using a joystick, in an attempt to track the movement of a moving target. The movement could be in one or two dimensions, always in discrete jumps, and the cursor could be controlled

by either the velocity or the acceleration of the joystick, resulting in four levels of task difficulty. In the second task, administered concurrently, the subject heard a series of high and low tones and was to count the number of occurrences of one of the tones. The P300 component of ERP responses to both tasks was measured. This component is very attention-dependent. The finding was that, across conditions, the P300 to the tracking targets and the P300 to the tones exhibited a reciprocity. The larger the P300 was to the tracking targets, the smaller it was to the tones, and vice versa. The sum of the P300 amplitudes was practically constant across conditions. The simplest interpretation of these results is that there is a fixed capacity that can be divided among the two tasks in different proportions, and that the relative P300 amplitudes reflect these proportions.

In sum, the existing literature can be accounted for with the hypothesis that there is a single capacity-limited store that can be identified with the focus of attention. This store is supplemented with other storage mechanisms that are not capacity limited although they are limited by the passage of time and/or the presentation of similar interfering material. The focus of attention can shift from one type of activated memory to another and will recoup considerable information from each type if the materials represented are dissimilar.

4.3. Implications for alternative accounts of information processing

In light of the information and ideas that have been presented, it is important to reconsider alternative accounts of information processing and the question of their continued viability and plausibility.

4.3.1. The magical number seven, plus or minus two. Although Miller (1956) offered his magical number only as a rhetorical device, the number did serve to characterize performance in many tasks. It has been taken more literally as a memory limit by many researchers (e.g., Lisman & Idiart 1995). The present stance is that the number seven estimates a commonly obtained, compound capacity limit, rather than a pure capacity limit in which chunking has been eliminated. It occurs in circumstances in which the stimuli are individually attended at the time of encoding and steps have not been taken to eliminate chunking. What is needed, however, is an explanation of why this particular compound limit crops up fairly often when rehearsal is not prevented. One possibility is that this number reflects a certain reasonable degree of chunking. Most adults might be able to learn at most three chunks of information rapidly, each with perhaps three units, leading to a span of nine. The slightly lower estimates that are often obtained could result from the inability to learn the chunks quickly enough. However, these speculations are intended only to provoke further research into the basis of commonly obtained compound capacity limits. What is essential to point out in the present account is that these compound limits are too high to describe performance in the situations in which it can be assumed that the combination of items into higher-order chunks was severely limited or prevented.

4.3.2. The time-limitation account. The view that working memory is limited by the duration of unrehearsed information in various short-term buffers is exemplified by the model of Baddeley (1986). The research reviewed in the present article leaves open the question of whether time limitations exist (as explained in sect. 1). Whereas some have assumed that time limits can take the place of capacity limits, the evidence described in this article, however, cannot be explained in this manner. In Baddeley's theory, memory span was said to be limited to the number of items that could be rehearsed in a repeating loop before their representations decay from the storage buffer in which they are held (in about 2 sec in the absence of rehearsal). If rehearsal is always articulatory in nature, though, this notion is inconsistent with findings that the memory span for idioms would imply a much longer rehearsal time than the memory span for individual characters (e.g., see Glanzer & Razel 1974). Something other than just the memory's duration and the rate of articulatory rehearsal must limit recall.

The time-based account might be revived if a different means of rehearsal could be employed for idioms than for words. For example, subjects might be able to scan semantic nodes, each representing an idiom, and quickly reactivate them in that way without articulatory rehearsal of the idioms. According to a modified version of Baddeley's account, this scanning would have to be completed in about 2 sec to prevent decay of the original memory traces. However, even that modified time-based theory seems inadequate to account for situations in which the material to be recalled is presented in an array so quickly that rehearsal of any kind can contribute little to performance (e.g., Luck & Vogel 1997). Also, any strictly time-based account has difficulty explaining why there is an asymptotic level of recall in partial report approximating four items with both auditory and visual presentation of characters, even though it takes much longer to reach that asymptote in audition (at least 4 sec: Darwin et al. 1972) than in vision (at most 1 sec: Sperling 1960). The only way to preserve a time-based account would be to assume that the rate of extraction of information from sensory storage in the two modalities is a limiting factor and is, for some mysterious reason, inversely proportional to the duration of sensory storage, resulting in an asymptotic limit that does not depend on the duration of storage. It seems far simpler to assume a capacity limit.

4.3.3. The unitary storage account. Some theorists (e.g., Crowder 1993) have assumed that there is no special short-term memory mechanism and that all memory may be explained according to a common set of rules. In one sense the present analysis is compatible with this view, in that the capacity limit applies not only to the recall of recently presented stimuli, but also to the recall of information from long-term memory (see sect. 3.4.2). However, any successful account must distinguish between the vast information potentially obtainable from an individual, on one hand, and the small amount of information that can be obtained from that individual, or registered with the individual, in a short segment of time; the capacity limit. The focus of attention, which serves as the proposed basis of the capacity limit in the present approach, has not played a major role in unitary accounts that have been put forward to date, though it could be added without contradiction.

Given a unitary memory view expanded to consider the focus of attention, one could account for the 4-chunk limit on the grounds that every chunk added to the focus diminishes the distinctiveness of all the chunks. Such a mechanism of indistinctiveness would be analogous to the one that

has been used previously to account for the recency effect in serial recall (e.g., Bjork & Whitten 1974); except that the dimension of similarity between chunks would be their concurrent presence in the focus of attention, not their adjacent serial locations within a list. One article written from a unitary memory view (Brown et al. 2000) does attempt to account for the number of chunks available in one situation. Specifically, their account, based on oscillatory rhythms that become associated with items and contexts, correctly predicted that the serial recall of a nine-item list with overt rehearsal is optimal when the list is rehearsed in groups of three items. The explanation offered was that "This represents the point at which the optimal balance between across-group errors and within-group errors is reached in the model." The account of the 4-chunk limit offered earlier in this target article on the basis of neural oscillatory rhythms (sect. 4.1.2) is similar (albeit on a neural level of analysis). It states that, with the neural representation of too many chunks simultaneously, the representations begin to become confusable (cf. Luck & Vogel 1998). The critical difference between explanations is that the neural account offered by Luck and Vogel and the present article refers to particular frequencies of neural oscillation, whereas Brown et al. allowed various oscillators and did not make predictions constrained to particular frequencies of oscillation.

4.3.4. The scheduling account. It has been proposed that supposed capacity limits might be attributable to limits in the rate at which subjects can produce responses in a multitask situation without risking making responses in the incorrect order (Meyer & Kieras 1997). That theory appears more applicable to some situations than to others. In situations in which the limit occurs during reception of materials and fast responding is not required (e.g., Luck & Vogel 1997), the theory seems inappropriate. That seems to be the case with most of the types of phenomena examined in the present article. It is unclear how a scheduling account could explain these phenomena without invoking a capacity notion.

4.3.5. The multiple-capacity account. Some theorists have suggested that there is not a single capacity limit, but rather limits in separate capacities (e.g., visual and auditory or spatial and verbal; see Wickens 1984). I would suggest that, although there may well be various types of distinct processes and storage facilities in the human brain, there is no evidence that they are limited by *capacity* per se (as opposed to other limitations such as those imposed by decay and interference). Sections 1 and 3 of the present article should illustrate that strict conditions must apply in order for chunk capacity limits to be clearly observed at all, free of other factors. Moreover, the finding of Sirevaag et al. (1989) of a tradeoff between tasks in the the P300 response magnitude in event-related potentials (discussed in sect. 4.2) seems to indicate that very disparate types of processes still tap a common resource. Even the left and right hemispheres do not appear to operate independently. Holtzman and Gazzaniga (1992) found that split brain patients are impeded in responses made with one hemisphere when a concurrent load is imposed on the other hemisphere, despite the breakdown in informational transmission between the hemispheres through the corpus collosum. There thus appears to be some central resource that is used in disparate tasks, and by both hemispheres.

4.3.6. The storage versus processing capacities account. Daneman and Carpenter (1980), like many other investigators, have noted that a working memory storage load does not interfere with processing nearly as much as would be expected if storage and processing relied upon a common workspace. Halford et al. (1998) noted the storage limit of about four items but also proposed, parallel to that limit but separate from it, a processing limit in which the complexity of relations between items being processed is limited to four dimensions in adults (and to fewer dimensions in children). Thus, within processing, "complexity is defined as the number of related dimensions or sources of variation" (p. 803). For example, transitive inference is said to be a ternary relation because it can be reduced to such terms: "the premises 'Tom is smarter than John, John is smarter than Stan' can be integrated into the ternary relational instance monotonically-smarter (Tom, John, Stan)" (p. 821), an argument with three fillers. The parallel between processing and storage was said to be that "both attributes on dimensions [in processing] and chunks [in storage] are independent units of arbitrary size" (p. 803). However, the model did not explain why there was the coincidental similarity in the processing and storage limits, to about four units each.

According to the present view, both processing and storage would be assumed to rely on a common capacity limit. The reason is that, ultimately, what we take to be stored chunks in short-term memory (and what I have, for simplicity, described as such up to this point) actually are relations between chunks. It is not chunks per se that have to be held in short-term memory (as they in fact are part of long-term memory), but rather chunks in relation to some concept. For example, "in-present-array (x, q, r, b)" could describe the quaternary relation leading to a whole report response in Sperling's (1960) procedure. "Monotonically-later (3–7, x, 2, 4–8)" could describe a quaternary relation leading to partially correct serial recall of an attended list of digits for which 3–7 is a memorized initial chunk; x represents a placemarker for a digit that cannot be recalled; 2 represents an unchunked digit; and 4–8 represents another memorized chunk.

If this analysis is correct, there is no reason to expect a separation between processing and storage. The reason why a storage load does not much interfere with processing is that the storage load and the process do not have to be expanded in the focus of attention at the same time. Although both are activated at the same time, there is no capacity limit on this activation, only with its use (cf. Schneider & Detweiler 1987). The subject might only hold in the focus of attention a pointer to the activated, stored information while carrying out the processing, and then the subject could shift the focus to the stored information when necessary to recall the memory load.

4.3.7. The task-specific capacities account. A skeptic might simply assume that although there are capacity limits, they vary from situation to situation for reasons that we cannot yet understand. This type of view probably cannot be answered through reasoned discourse as it depends on a different judgment of the presented evidence. Further assessment of the view that there is a fixed underlying capacity could be strengthened by subsequent research in which new conditions are tested and found to conform to or violate the capacity limit. Numerous examples of novel condi-

tions leading to the predicted limit can be given, but two of them are as follows. First, in the research by Cowan et al. (1999), a capacity limit for ignored speech was expected to be similar to those that have been obtained for attended visual arrays (Luck & Vogel 1997; Sperling 1960) on the grounds that the task demands were logically analogous, even though the materials were very different. That expectation was met. Second, it was expected that one could observe the capacity limit by limiting rehearsal for spoken lists, and that expectation provided a very similar limit in numerous published experiments (as shown in Table 2). A third example has yet to be tested. Specifically, it was predicted (in sect. 1.2.1) that the capacity limit could be observed in a modified *n*-back task in which subjects must only indicate, as rapidly as possible, if a particular item has been included in the stimulus set previously, and in which some items would be repeated in the set but other, novel items also would be introduced.

4.4. Boundaries of the central-capacity-limit account

The boundaries of the present type of analysis have yet to be examined. For example, Miller (1956) indicated that absolute judgments in perception are limited in a way that is not clear; apparently not in chunks as for other types of phenomena. For unidimensional stimuli the limit appears to be up to about seven categories that can be used consistently, but the limit in the number of total categories is considerably higher for multidimensional stimuli (e.g., judgments of tones differing in both intensity and pitch). One possibility is that the subject need only retain, in short-term memory, pointers to the dimensions while accessing category divisions one dimension at a time. Because faculties that are not specifically capacity-limited, such as sensory memory, can be used for supplementary storage, the focus of attention is free to shift to allow the sequential use of the capacity limit to judge the stimulus on different dimensions, one at a time. This analysis might be tested with absolute judgments for backward-masked stimuli, as backward masking would prevent sensory storage from holding information while the focus of attention is shifted from one dimension to another. Thus, as in this example, the capacity concept potentially might have a broad scope of application indeed.

5. Conclusion

In this target article I have stressed several points. The first is the remarkable degree of similarity in the capacity limit in working memory observed with a wide range of procedures. A restricted set of conditions is necessary to observe this limit. It can be observed only with procedures that allow assumptions about what the independent chunks are, and that limit the recursive use of the limited-capacity store (in which it is applied first to one kind of activated representation and then to another type). The preponderance of evidence from procedures fitting these conditions strongly suggests a mean memory capacity in adults of three to five chunks, whereas individual scores appear to range more widely from about two up to about six chunks. The evidence for this pure capacity limit is considerably more extensive than that for the somewhat higher limit of 7 ± 2 stimuli; that higher limit is valid nevertheless as a commonly observed, compound STM limit for materials that allow online rehearsal, chunking, and memorization, for which the exact number of chunks in memory cannot be ascertained. The fundamental capacity limit appears to coincide with conditions in which the chunks are held in the focus of attention at one time; so it is the focus of attention that appears to be capacity-limited.

When the material to be remembered is diverse (e.g., some items spoken and some printed; some words and some tones; or some verbal and some nonverbal items), the scene is not coherent and multiple retrievals result in considerably better recall. This all suggests that the focus of attention, as a capacity-limited storage mechanism, can shift from one type of material to another or from one level of organization to another, and that the individual is only aware of the handful of separate units of a related type within a scene at any one moment (Cowan 1995; Mandler 1985).

ACKNOWLEDGMENTS

This project was supported by NICHD Grant R01 21338. I thank Monica Fabiani, Gabriele Gratton, and Michael Stadler for helpful comments. Address correspondence to Nelson Cowan, Department of Psychology, University of Missouri, 210 McAlester Hall, Columbia, MO 65211, USA. Electronic mail: cowanN@missouri.edu.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

The dangers of taking capacity limits too literally

S. E. Avons, Geoff Ward, and Riccardo Russo

Department of Psychology, University of Essex, Colchester CO7 6SJ, United Kingdom. **{savons; gdward; rrusso}@essex.ac.uk**
www.essex.ac.uk/psychology/people/{sa; gw; rr}.html

Abstract: The empirical data do not unequivocally support a consistent fixed capacity of four chunks. We propose an alternative account whereby capacity is limited by the precision of specifying the temporal and spatial context in which items appear, that similar psychophysical constraints limit number estimation, and that short term memory (STM) is continuous with long term memory (LTM).

In his target article, Cowan makes a brave attempt to unify attention and short term memory (STM) by stressing the similarity of capacity estimates across various paradigms. His evidence is drawn from four broadly defined sources: (1) the number of correct reports in whole report paradigms; (2) the size of memory span when rehearsal and recoding are discouraged; (3) performance discontinuities in various attentional and STM tasks; (4) a miscellany of indirect measures including recency in free recall, implicit memory, and visual search. The capacity estimates cited by Cowan are less consistent than he claims. If four of the most recently presented items exist within the attentional focus, then recall of about four of the most recent items would be expected.

This is true for enumeration, whole report, and alphanumeric

span tasks under certain conditions. However, performance fails to reach this level in other studies, such as recall of familiar visual stimuli (Zhang & Simon 1985), multi-object tracking (Yantis 1992), free recall of the most recently rehearsed items (Tan & Ward, in press), and, as Cowan acknowledges, in repetition priming (McKone 1995).

Our main objection is more fundamental and takes the following form: (a) these data indicate memory for events rather than items, (b) the notion of a fixed capacity limit, restricted to three or four memory elements is misleading, and (c) that these considerations, taken together, question the qualitative distinction between STM and LTM that Cowan advocates.

Memory for events, not items (or chunks). Cowan propose that the capacity of STM is not 7 chunks, as proposed by Miller (1956), but about 4 chunks. The earlier estimate was inflated because it was based on rehearsal-dependent immediate serial recall, and also because the span measure estimated the 50% probability of recall. In fact, many STM tasks require recall not only of recent items, but also of their temporal or spatial context. This is true for studies of whole report, serial recall, and serial-order reconstruction (sects. 3.1 and 3.2 of the target article), and applies also to the Steinberg paradigm (sect. 3.3.4). The apparent limit of four items could arise in two ways. It may be the case that only four items may be represented in immediate memory (i.e., within the focus of attention), or that contextual information can only be preserved for about four items. Since order and position errors are common in span tasks, and memory span is more closely related to order memory than item memory (e.g., Martin 1978), it is likely that the demands of the contextual component, rather than the number of activated items, set a limitation on capacity.

Against magical numbers. It has long been known that in dot enumeration studies, RT increases linearly as the display size increases above 4 items, suggesting that some serial counting process is taking place. One account of the processing discontinuity is that the error in number estimation increases systematically with array size (Vos 1982). Hence, there is following Weber's Law (Van Oeffelen & Vos 1982). Hence, there is no discontinuity in number estimation between small values within the subitizing range, and larger values. Rather, when estimating small numbers of discrete items, the estimation error is sufficiently small to allow a response of the nearest corresponding integer. With larger numbers, the error range may encompass several integers, and thus to maintain accurate performance at larger array sizes requires recruitment of a new algorithm, counting. To treat error-prone performance as an indication that certain items have not been processed results in the imposition of an artefactual capacity limit. This concern haunts all models which propose a fixed integer number of slots, pointers or entities which govern STM capacity (e.g., Conrad 1965; Trick & Pylyshyn 1994). Capacity is clearly limited, because increasing the information load degrades performance. But this does not imply dedicated mechanisms in 1:1 correspondence with small arrays.

Continuity of STM and LTM. This argument can be extended, given the case made above, that the limitation of span-type studies arises from temporal and spatial ordering of items. The precision with which discrete values can be classified on any dimension is limited, as was demonstrated by many of the early psychological applications of information theory.

With small numbers of categories, classification is precise, but as the categories increase in number, errors will occur. There is no reason here to consider that there is an abrupt discontinuity, and that beyond some limit, events are discarded from STM. An alternative conception proposes that over small ranges, ordering is precise and retrieval rapid, whereas, with longer lists temporal precision breaks down, and alternative encoding and retrieval processes are employed. If the requirement for temporal precision is relaxed, for example, as in the studies of Jahnke et al. (1989), then performance improves. This could not occur if items were discarded from lists exceeding some arbitrary limit. If this position is accepted, then the need to define a rigid partition between STM and LTM (one consequence of Cowan's theory) is reduced, although the strategies and processes useful for memory may indeed change as a function of elapsed time, and the labels may be convenient and pragmatically useful.

A biocognitive approach to the conscious core of immediate memory

Bernard J. Baars
The Wright Institute, Berkeley, CA 94704
baars@cogsci.berkeley.edu

Abstract: The limited capacity of immediate memory "rides" on the even more limited capacity of consciousness, which reflects the dynamic activity of the thalamocortical core of the brain. Recent views of the conscious narrow-capacity component of the brain are explored with reference to global workspace theory (Baars 1988; 1993; 1998). The radical limits of immediate memory must be explained in terms of biocognitive brain architecture.

I am pleased that Cowan finds my global workspace theory useful in understanding the limited item capacity of immediate memory (Baars 1988; 1997; Baars & Newman, in press). Without proposing a specific solution to the complex arguments adduced in this paper, I would like to suggest a somewhat different way of thinking about the problem, based on an emerging understanding of the brain basis of consciousness (e.g., Baars 1993; 1998; Crick 1984; Damasio 1989; Destexhe et al. 1999; Engel et al. 1999; Hobson 1997; Newman et al. 1997; Steriade 1993; Tononi & Edelman 1998). The "magic number 7 plus or minus 2," which Cowan persuasively maintains is really about half that size, has become so familiar to scientists that we rarely pause to think how extraordinarily small it is, given a brain with hundreds of billions of neurons, each firing at 10 Hz or faster, and each densely connected to all the others in only a few steps.

Limited capacity, which is always associated with consciousness, presents a great paradox: How could it make sense for it to be so small? Humans and animals must often run into danger because of their tiny capacity for immediately retrievable information. It is simply implausible to think that a larger memory capacity could not evolve over hundreds of millions of years of brain evolution. When a giraffe bends down to drink at a water hole, it cannot at the same time keep track of its young, pay attention to possible predators, check for competing giraffes in the herd's dominance hierarchy, and see if some sexual competitor is making eyes at its mate. If we add the need to make choices among action alternatives, which also loads limited capacity, real-life situations like this must be informationally overwhelming at times. Being so easily overwhelmed with information must reduce our ability to respond rapidly and effectively to predators and other dangers. Like sleep, which also exposes animals to danger, limited capacity is a biological puzzle, because it appears to increase the risk to survival and reproductive fitness.

Indeed, the capacity for unrelated items in focal consciousness is even smaller than the size of immediate memory. I have suggested that conscious capacity is limited to only one single internally consistent event at any given instant – one coherent "chunk" at a time (Baars 1988; 1998). The evidence comes from numerous studies of ambiguous figures and words, in which only one interpretation can be conscious at any instant. This is bolstered by evidence from dual-input tasks such as dichotic listening and binocular rivalry, and more recent experimental paradigms like inattentional blindness (see Baars & Newman, in press). Other factors like visual scene complexity, which involves scenes that are hard to organize into coherent conscious units, also degrade our ability to detect and respond to events.

It should be noted that very few items in short term memory are conscious at any moment. Immediate memory has both con-

scious and unconscious components, as becomes obvious simply by considering how many items are reportable at any given instant as conscious. In a rehearsed set of unrelated items, the Sternberg task suggests that only one can be retrieved and reported as conscious at any given moment (Sternberg 1966). Thus, the momentary capacity of consciousness may be even smaller than that of immediate memory. It can be argued that immediate memory "rides" on the function of consciousness in the brain. Both have radical limits, in sharp contrast to such things as longterm episodic memory or linguistic knowledge. In humans, immediate memory is enhanced by internal rehearsal and the "visual sketchpad" (Baddeley 1993). But the apparatus of consciousness is not limited to humans; indeed, the most basic substrate of conscious states may exist in all vertebrates, and certainly in mammals.

Such a built-in capacity limit suggests a biological tradeoff. In exchange for this radical limit, vertebrate species must gain some compensating advantage. Such trade-offs are routine in living organisms and engineered machines alike. But what could be the advantage? The only plausible answer, it seems to me, is to consider the brain as a system architecture, a society of vast numbers of complex, specialized neurons, which cluster in anatomical columns, arrays, and pathways with both fixed and variable connectivities controlled by neurochemical modulation (Hobson 1997). The most obvious circadian modulation of systemwide connectivity has to do with the changes due to conscious waking, dreaming, and unconscious slow-wave sleep. During conscious states, massive reentrant adaptive resonance takes place in the thalamocortical core, yielding a distinctive electrical field signature (Destexhe et al. 1999; Steriade et al. 1993).

Wolf Singer et al. have made a persuasive case in recent years that conscious perceptual states are associated with synchronized oscillatory activity in the 40 Hz range (e.g., Engle et al. 1999), while others suggest that thalamocortical firing is shaped by an information-theoretic parameter called complexity (Tononi & Edelman 1998). These integrative features of the massive "dynamic core" presumably underlie conscious limited capacity; but they also make possible global access between conscious contents and unconscious processes involved in learning, memory, motor control, executive processes, and the like. Further, the thalamocortical core shows remarkable moment-to-moment adaptive changes as a function of attentionally selected incoming stimulation. Thus consciousness combines limited focal contents with massive brainwide adaptation to selected input. These general features command wide agreement (see Baars & Newman, in press).

What could be the functional basis for a narrow coordinative capacity in a massively parallel society of specialized neuronal arrays? A long theoretical tradition in psychology suggests an answer. Cognitive architectures were first developed in the 1950s by scientists including Alan Newell and Herbert A. Simon, and have since been developed by many others, notably John R. Anderson (see Baars 1988). All cognitive architectures combine a parallel-interactive set of memory elements with a narrow-capacity channel, which allows the memory elements to interact and be coordinated for the purpose of problem solving. Some artificial intelligence researchers have suggested that such "global workspace" systems provide a general purpose problem-solving architecture. When predictable problems are encountered, such an architecture can rapidly retrieve a "canned" response; and when an unpredictable situation must be faced the architecture allows coalitions of specialized knowledge sources to generate potential solutions. It is striking that all our integrative models of human cognition have these basic features; A massive parallel-interactive memory with a narrow-capacity channel used for interaction, coordination and control. I have suggested that in fact the brain has found a similar architectural solution to its distinctive biological challenges (Baars 1988; 1993; 1998). The dynamic activity of the thalamocortical core described above may be the brain version of a cognitive architecture.

While these points do not directly yield a numerical estimate for the size of immediate memory, they suggest a framework for such an answer. First, the primary function of the limited capacity component associated with consciousness is to allow coordination, integration, and global access between elements of the massively parallel-distributed society of neural nets. Second, I would suggest that a crucial aspect of the problem is the way in which conscious cues may be used to index and retrieve unconscious chunks of information in immediate memory. Only within such a system architecture can we begin to make sense of the radically small size of immediate memory in a brain of extraordinary size, complexity, and moment-to-moment adaptive flexibility.

The magical number 4 = 7: Span theory on capacity limitations

Bruce L. Bachelder

Psychological and Educational Service, Morganton, NC 28655-3729.
brucebachelder@hci.net

Abstract: According to span theory, a behavioral theory of the magical numbers, Cowan's 4 and Miller's 7 are simply two different points on the same ogive describing the relation between performance and span load, a fundamental task characteristic. Span theory explains the magical numbers in terms of a unitary limited span ability, a mathematical abstraction from that ogive.

With this paper, cognitivist Cowan becomes something of a "strange bedfellow" in efforts to assert span theory (Bachelder & Denny 1977a; 1977b). Cowan is right, there is a magical number, it is unitary, and developmental and individual differences are important (sect. 4.1.4); but he is wrong (sect. 4.3.1) that Miller's idea was mere rhetoric. Miller (1956) evaluated a unitary channel capacity hypothesis but his analyses failed (Bachelder 2000). The notions of span ability and a capacity-limited focus of attention promise to fare better.

Figure 1 plots performance in the magical number tasks as a function of numbers of items in a stimulus set. The curve, an inverse ogive, suggests there is no fundamental difference between Cowan's 4 and Miller's 7; both are simply different points on the same curve.

The similar performance limitations in these tasks are usually presumed to be coincidental. Cowan's notion of a capacity-limited focus of attention undermines that presumption (sect. 4.3.3). Section 3.3.2 extends the theory to span of apprehension (known at

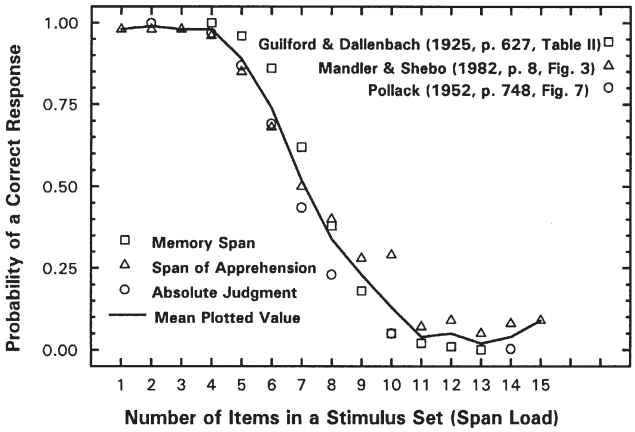


Figure 1 (Bachelder). The probability of a correct response as a function of size of stimulus set in the magical number tasks. Miller's 7 corresponds to a 50% criterion, Cowan's 4 to a 100% criterion. Pollack's data have been corrected. Mean performances on single stimuli were raised to the power of span load to estimate performance on a full stimulus set considered collectively.

one time as span of attention). Section 4.4 sketches an extension to absolute judgment which would complete a unitary account.

Span theory accounts for all three limits via a mathematical proposition, an empirical generalization from data such as shown in Figure 1: In any task the probability of a correct response is an inverse ogival function of span load. Span load is the number of stimuli jointly relevant for the target response. Span ability, operationalized as the 500 threshold of the Performance X Span Load ogive, is the ability to cope with span load. This definition closely parallels Miller's definition of channel capacity as "the upper limit on the extent to which the observer can match his responses to the stimuli we give him" (p. 82).

Span theory analyses may help Cowan avoid a cul-de-sac in an extension of his theory to uni- and multidimensional absolute judgment. First, Figure 1 shows that at Cowan's 100% criterion the magical number in absolute judgment is 4, not 7. Second, the principle that increasing the number of dimensions overcomes the magical number may be an artifact of the information metric. Miller relied largely on a study by Pollack and Ficks (1954). Bachelder (1978, Part 2) translated their data from information terms back to probability of a correct response, then modeled their task as a combination of serial recall and unidimensional absolute judgment. Subjects judged the values in each of multiple dimensions then retained and reported the values as in serial recall. Mathematical models, presuming unitary magical number limitations typical for college students, generate the published data accurately (mean error = 2.7%).

Span ability is a g-like trait construct. Figure 1 presents just one of a family of parallel curves starting at the left for the developmentally young (retarded and nonretarded). Memory span and span of absolute judgment have been found to correlate .78 (60 adults: college students, average, and retarded [IQ = 47.8]; summarized in Bachelder & Denny 1977a, pp. 139–42). All three magical numbers of mildly retarded adults are smaller (5 ± 2) than those of college students (Spitz 1973).

Cowan's concepts map to some extent onto span theory. A capacity-limited focus of attention is not that different from the notion of span ability, which can be characterized as the ability to attend to several stimuli simultaneously. Chunks usually correspond to responses. Recoding corresponds to changes in response repertoire.

Span theory tightly integrates psychometric concepts and methods into S-R style analyses of cognitive tasks. A true competition of cognitive and behavioral approaches in the analysis of the same tasks promises to enrich both traditions.

The magic number and the episodic buffer

Alan Baddeley

Department of Experimental Psychology, University of Bristol, Bristol BS8 1TN, United Kingdom. **alan.baddeley@bristol.ac.uk**
www.psychology.psy.bris.ac.uk/psybris

Abstract: Cowan's revisiting of the magic number is very timely and the case he makes for a more moderate number than seven is persuasive. It is also appropriate to frame his case within a theoretical context, since this will influence what evidence to include and how to interpret it. He presents his model however, as a contrast to the working memory model of Baddeley (1986). I suggest that this reflects a misinterpretation of our model resulting in a danger of focusing attention on pseudo-problems rather than genuine disparities between his approach and my own.

In proposing their model of working memory, Baddeley and Hitch split the earlier concept of the unitary short-term memory into a tripartite system: a sound or language-based phonological loop, a visuo-spatial sketchpad, and an attentionally limited controlling system, the central executive (Baddeley 1986; Baddeley & Hitch 1974). Within this model, the central executive (CE) maps most closely onto Cowan's capacity-limited attentional system. In the original model, as in Cowan's, the system was assumed to have both attentional and storage capacity. This issue will be discussed below. An important difference however, is that Cowan's model concerns short-term memory (STM), and as such is concerned with explaining how people remember small amounts of material over brief time intervals, while the working memory model tries to emphasise the functional role of a complex memory system across a wide range of cognitive tasks.

It is probably this narrower focus that has caused Cowan to concentrate in his contrasts between the two systems on one component of working memory, the phonological loop. The loop does indeed, as Cowan emphasizes, have a time-limited component, the phonological store. Cowan does, however, appear to ignore the overall attentional control system, the CE, the component that resembles his limited capacity model most closely.

Cowan himself assumes a range of other systems, which need to incorporate trace decay or interference, but dismisses those of our own model as "dubious" on the grounds that "one can always find examples that don't fit in" giving as an instance "spatial information conveyed through acoustic stimulation." Given that the core studies on which the concept of a visuo-spatial sketchpad were based involved the auditory presentation of visuo-spatially codeable material (Baddeley et al. 1975), this is somewhat surprising. It presumably stems from his confounding the sketchpad, an integrative system which explicitly combines visual and spatial information from a range of modalities, with modality-specific sensory memory. Cowan's claim that the components of our model are "situation-specific and therefore arbitrary" is also puzzling, given the extensive use of converging operations to define the working memory model, and its very wide application outside the memory laboratory (Baddeley & Logie 1999; Gathercole & Baddeley 1993).

Cowan's claim that apart from his attentional control system "no other mental faculties are capacity limited" is also puzzling. Can he really mean that every other cognitive system has infinite capacity? Surely not. Perhaps he means that they are limited by factors other than attentional capacity. However, if one takes this claim with a pinch of salt, and identifies Cowan's limited capacity system with the central executive, then his model resembles the 1974 working memory model, with under-specified slave systems. As Miyake and Shah (1999) suggest, most current models of working memory, whether explicit, or like Cowan's, implicit, have a great deal in common. There are however, differences between my own current model, and that implied by Cowan, including:

1. Cowan, in common with a number of other theorists, regards short-term memory as simply the currently activated areas of long-term memory. While this proposal has the advantage of simplicity, it is inconsistent with neuropsychological data. Deficits in long-term memory are found which have no apparent impact on immediate memory, or on working memory more generally (Wilson & Baddeley 1988), while both patient data and neuroradiological evidence indicator-specific short-term storage capacities (Basso et al. 1982; Smith & Jonides 1995).

2. Conceptualization of Cowan's limited capacity system differs somewhat from our view of the CE, although both are relatively loosely specified. Cowan appears to regard his attentional system as unitary, involved in most aspects of memory, and anatomically as depending principally on the parietal region. I regard the CE as a fractionable system that is less involved in retrieval than in encoding (Baddeley 1996; 1999; Baddeley et al. 1985; Craik et al. 1996). I also assume that it depends principally on the frontal lobes (Baddeley & Logie 1999).

3. Cowan seems to assume that his attentional system has storage capacity, otherwise, how are the sources of information integrated and maintained? While this assumption was made by our 1974 model, it was later abandoned, with the CE regarded as a control system using storage from elsewhere in WM or LTM (Baddeley & Logie 1999). Very recently, I have become convinced of the value of postulating a fourth component of working memory,

the *episodic buffer* which serves specifically as a limited capacity store in which information can be integrated from the slave systems and long-term memory and represented in a multi-feature code. As in Cowan's model, this is assumed to have a limited capacity determined by number of chunks, and to involve a retrieval component that is associated with conscious awareness. In addition to serving as a store, it is assumed to provide the modelling space where information can be combined and manipulated in order to plan future actions or interpret recollected experience (Baddeley 2000). It is thus assumed to be a temporary memory storage system. As such, it differs from the CE, an attentional control system that operates across many tasks other than memory. The episodic buffer is assumed to combine and store information from different sources, allowing active manipulation so as to create new representations that serve to solve novel problems. As such it provides the temporary storage needed for mental modelling.

I therefore welcome Cowan's review of evidence relating to the limited capacity of immediate memory. I prefer to attribute it however to the limited capacity of the central executive and its related storage system, the episodic buffer.

The size and nature of a chunk

C. Philip Beaman

Department of Psychology, University of Reading, Whiteknights, Reading RG6 6AL,United Kingdom. **c.p.beaman@reading.ac.uk**
www.rdg.ac.uk/acadepts/sx/Psych/people/beaman.html

Abstract: The data presented in the target article make a persuasive case for the notion that there is a fundamental limit on short term memory (STM) of about four items. Two possible means of further testing this claim are suggested and data regarding scene coherence and memory capacity for ordered information are reviewed.

Given the current state of knowledge, it is difficult (but not impossible) to argue against the idea that there is some fundamental, architectural constraint on processing capacity (Just et al. 1996; Young 1996; Young & Lewis 1999). It may be even harder, however, to point to exactly what that constraint is. The compound" STM limits Cowan refers to in the target article reflect functional rather than architectural constraints on what constitutes processing capacity, leaving Cowan with the unenviable job of defining the fundamental "pure" STM capacity limit in terms of chunks of information. As with Miller's classic article on the same topic, however (Miller 1956), there is no satisfactory, independent definition of a chunk.

The circularity of the logic becomes apparent thus: chunks can be identified with individual stimulus items when information overload or other experimentally induced conditions prevent coding of the stimulus items into larger or, arguably, higher-level chunks. How can one identify a chunk in any particular situation? When steps have been taken to prevent coding of an item to a higher level, the item will be the chunk. How do we know whether the experimental conditions have been successful at preventing recoding? Because the item will act like a chunk. In other words an item will act as a chunk when it has not been recoded. We know it has not been recoded because it is acting like a chunk.

In fact, Cowan provides a way round this problem by his analysis of the conditions under which a chunk size of 4 emerges, his four basic conditions. His argument therefore stands. As an argument. It is, and should be, open to empirical test. There appear to be two obvious ways of going about this. The first way is to implement the constraint in some formal model that also includes the four basic conditions for identifying a chunk. Are chunk sizes of 4 really observed when information overload in some (as yet unspecified) way prevents higher level recoding in simulation studies? Or when other steps are taken specifically to block the coding of items into larger chunks? The emergence of the chunk size will be identifiable by performance discontinuities and various indirect effects of the capacity limits (Cowan's final two basic conditions). Performance discontinuities and other effects of a limited capacity should then be predictable on the basis of the ways in which recoding has been impeded.

What is clear from this analysis is that Cowan's argument would benefit from a stricter definition of how to prevent higher level recoding than he currently provides. This may also require further work in investigating the nature and limitations of the processes contributing to compound STM limits in order to definitively rule out their involvement.

The second way of testing Cowan's argument is, at first blush, the simplest. In addition to studies cited by Cowan there are numerous others in the literature that address the question of processing limitations. Any that conform to Cowan's basic conditions but do not demonstrate signs of a 4-chunk capacity limitation will, in effect, falsify Cowan's thesis. Once again, there is the problem of ensuring that high-level recoding has been adequately impeded but if a procedure is available that impeded recoding in an earlier experiment it is reasonable to assume that recoding was equally well impeded even if measured capacity exceeds 4 chunks.

Finally, consider the following intriguing scenario suggested by Cowan in relation to scene coherence. Given a 4-chunk capacity-limited STM we must assume that scenes may have up to 4 separate parts, or events, in awareness at any given moment that, although associated with a common higher-level node, do not necessarily have any common level of association beyond this. To connect them will require shifting attention from one event to another in a manner sub-optimal for later serial recall. Within the auditory domain experiments of this nature have in fact been conducted. Performance is worse when untrained listeners are required to report the order of a recycled sequence of four unrelated sounds than when the sequence contains sounds of a similar type (Warren & Obusek 1972; Warren et al. 1969). These studies suggest that order information is indeed impaired when items are not associated.

A further study from a different domain, that of thought suppression (Wegner et al. 1996) also concludes that memory for order requires some level of association. Wegner et al. found that if participants were required to suppress memory for the sequence of events in a film sequence, information was lost but not item information (as measured by free and cued recall and by recognition). These studies, although they do not speak directly to the issue of capacity, confirm Cowan's intuition that shifts of attention between different events are necessary to set up sequence information for events which otherwise are not associated. Such sequence information can be destroyed by active attempts to suppress the memory of the events. The studies, however, also raise questions about the nature of an association. If the processing limit on STM is 4 chunks, what makes information about the order of those chunks easier or harder to recall? The question is one that is of more importance than might immediately appear. A number of studies which Cowan takes as evidence for his 4-chunk capacity limit were serial recall studies in which items were only marked correct if they were recalled in the correct serial order (e.g., Table 2 in target article). It seems therefore that the capacity limit reflects not a limit on capacity to recall individual items but a combined limit on recall of the items and the (order) relations between them. In serial recall, correct recall of items is interdependent. If an item is misrecalled early in the list it is unlikely to be correctly recalled later in the list as well as blocking recall of the correct item at the earlier position. If capacity as measured by serial recall studies is indeed 4 chunks, then capacity is for 4 chunks plus some extra information connecting those chunks. It is then necessary to determine the nature of this "extra" information and its relation to chunk capacity.

There is no four-object limit on attention

Greg Davis

Department of Psychology, Birkbeck College, London WCIE 7HX, United Kingdom. **g.davis@psyc.bbk.ac.uk** **www.psyc.bbk.ac.uk/staff/gid.html**

Abstract: The complex relationship between attention and STM forms a core issue in the study of human cognition, and Cowan's target article attempts, quite successfully, to elucidate an important part of this relationship. However, while I agree that aspects of STM performance may reflect the action mechanisms that we normally consider to subserve "attention" I shall argue here that attention is not subject to a fixed four-object capacity limit as Cowan suggests. Rather, performance in attention tasks as well as STM may be best accounted for in terms of decay and interference.

Many previous studies have concluded that observers can track or enumerate up to four small items efficiently, but cannot do so for more than five objects, this restriction reflecting a four-object limit on our visual attention (e.g., Pylyshyn 1989; Trick & Pylyshyn 1993). However, these studies have largely employed displays such as those in Figures 1A and 1B, where several other display variables co-vary with the number of relevant objects. For example, the five objects in Figure 1B constitute a greater overall surface area than do three objects in Figure 1A, and presumably comprise a greater number of visual "features" overall. Accordingly, any costs for attending more versus fewer objects in such displays may not reflect the number of relevant objects per se, but might alternatively result from the greater amount of relevant perceptual information comprised by larger numbers of objects.

In order to remove this ambiguity in their own study, Davis et al. (submitted) have compared attention to displays of three large objects (Fig. 1C) versus six small objects (Fig. 1D), in which only the number of objects varied. Note that although the two display types comprise different numbers of objects, they are otherwise very much alike, so that the six objects comprise approximately the same surface area and number of "features" as do the three objects. If there is indeed a fixed four-object limit on attention, then attending the six-object displays should entail performance costs relative to attending the three-object displays. Conversely, if apparent four-object limits in previous studies reflected the overall surface area and number of features that co-varied with the number of objects there, the two displays should now yield equivalent performance.

To measure how efficiently the two displays could be attended, Davis et al. employed a "divided-attention" task similar to those in many previous studies of visual attention. In each trial, the six or three objects were initially presented for 2.4 seconds to give subjects ample time to focus their attention on the objects. Next, two notches were removed from the objects in the display. The notches were either square or jagged and the position of one notch in no way predicted the position of the second notch, with every possible combination of notches and notch-positions being equiprobable. The task was simply to determine whether the two notches were of the same type (i.e., both jagged or both square) or of different types (i.e., one jagged, one square), and to press the appropriate key on a computer keyboard as quickly as possible.

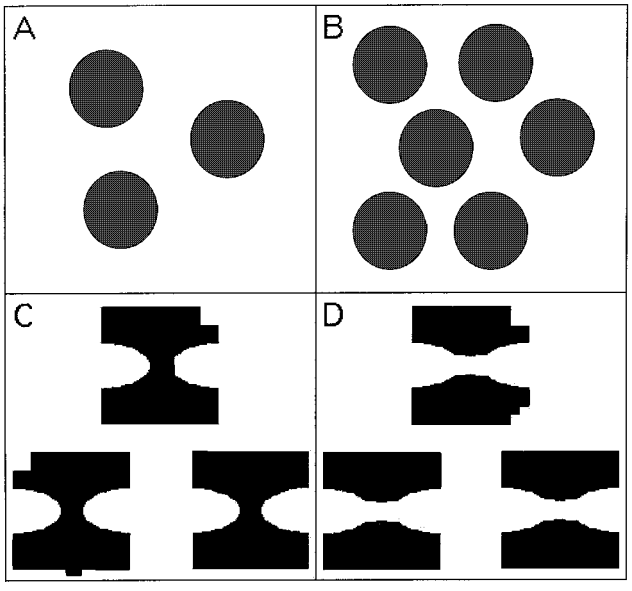


Figure 1 (Davis).

Performance was indistinguishable for the six- versus three-object displays, suggesting that no four-object limit had operated, and that six objects can be attended as efficiently as three objects, once other display variables are equated in the two cases. However, one valid objection to this new study might be that observers had not perceived the six-object displays to contain six separate objects. Rather they may have perceived each pair of objects in those displays to comprise a single object or Gestalt, and thus perceived only three objects in the six-objects displays. To preclude this possibility, Davis et al. examined performance on some specific trials within the three- and six-object displays. For six-object displays they compared RTs when the two features by chance appeared on a single small object (i.e., were horizontally-displaced from each other) versus appearing on two separate neighboring objects (vertically-separated within a single "pair" of objects as in Fig. 1D). For three-object trials they conducted an identical comparison of horizontally- versus vertically-displaced feature-pairs, except that the features now always appeared upon a single large object.

Many previous studies have demonstrated that comparison of features within pairs of objects is more efficient when they appear on the same object than on two separate objects (e.g., Davis 2000; Duncan 1984; Lavie & Driver 1996; Watson & Kramer 1999). Thus if the six-object displays had indeed been perceived to comprise six objects, and the three-object displays, three objects, the following patterns of performance should be expected. First, horizontally separated features (on the *same* small object) in the six-object displays should be responded to faster than vertically-separated features in those displays (lying on different, neighboring objects). Second, if these differences reflected whether the two features were on same versus different objects, rather than simply resulting from other differences between vertically and horizontally-separated features, this pattern of results should not hold in the single large objects of the three-object displays. Precisely this predicted pattern of results was found, validating Davis and colleagues' comparison of six versus three objects.

Accordingly, Davis et al. suggested that no four-object limit holds for visual attention. Indeed, attention may not be limited to any "magica1" number of objects, but rather may reflect the same properties of decay and interference that Cowan ascribes to STM. Recent evidence points to the existence of two distinct binding mechanisms in vision: within-object "links" that bind together features from the same perceptual objects and between-object links that bind features from separate objects. These two types of "link" appear to be coded in anatomically separate visual streams, so that there may be substantial mutual interference between links of the same type, but very little interference between the two types of link (see Humphreys 1998). Davis et al. suggested that this possible mutual interference between links might explain both the equivalent performance found for their six- versus three-object displays, and a range of other findings previously ascribed to a four-object limit. First, since their six- versus three-object displays comprised approximately equal numbers of features, presumably bound by the same number of links, interference should have been approximately the same in the two cases, correctly predicting the equivalent performance found. However, in more conventional displays, such as those in Figures 1A and 1B, where the number of features increases with the number of objects, mutual interference between links should increase, decreasing the effi-

ciency with which the objects can be attended. In fact, when Davis et al. calculated how within and between-object links might be expected to increase for 2–4 objects versus for 5–7 objects, this yielded expected performance decrements of 37 and 235 msec/item-1 respectively, closely approximating previously reported values of 40 and 240 msec/item-1. Clearly, this very simple account models only performance rather than underlying cognitive processes. However, it does provide an existence proof that apparent four-object limits in previous studies can in principle be accounted for in terms of interference alone.

I conclude that there is no four-object limit on attention, and that the Davis et al. view of attention limitations may generate far more biologically plausible models than accounts that assume a "magic" number of objects. However, the absence of a four-object limit on attention does not invalidate attempts to ascribe aspects of STM performance to attention processes. Indeed, a common interference-based framework for understanding STM and attention may assist such efforts.

The search for fixed generalizable limits of "pure STM" capacity: Problems with theoretical proposals based on independent chunks

K. Anders Ericsson and Elizabeth P. Kirk

Department of Psychology, Florida State University, Tallahassee, FL 32306-1270. **{ericsson; kirk}@psy.fsu.edu**
www.psy.fsu.edu/~ericsson

Abstract: Cowan's experimental techniques cannot constrain subject's recall of presented information to distinct independent chunks in short-term memory (STM). The encoding of associations in long-term memory contaminates recall of pure STM capacity. Even in task environments where the functional independence of chunks is convincingly demonstrated, individuals can increase the storage of independent chunks with deliberate practice – well above the magical number four.

Cowan is continuing the search for a magical number that measures the universal limit for the pure capacity-based STM ("pure STM"). Miller (1956) first proposed the magical number seven to describe the remarkable invariance in performance on laboratory tests of short-term memory (STM). He discovered that individuals encoded stimuli as familiar distinct patterns (chunks) already stored in long-term memory (LTM) and could recall roughly the same number of chunks regardless of type of presented items, such as letters or words. When no associations are formed between chunks each chunk must reside independently in transient STM to be accurately recalled. Consequently, Miller proposed that the storage capacity of STM had an invariant general limit that could be expressed as approximately seven independent chunks.

Subsequent research has shown that the fixed capacity limit of independent chunks does not constrain STM after extended practice (Chase & Ericsson 1982; Richman 1995), and for skilled and expert performance (Ericsson & Kintsch 1995; Gobet & Simon 1996). In domain-specific tasks, experts are able to encode and store presented information in LTM and expand the capacity of working memory (WM).

Ericsson and Kintsch (1995) showed that skilled individuals can store information in long-term working memory (LTWM) by integrating two encoding mechanisms. First, they are able to generate associations between different presented chunks of information and build new integrated structures in LTM. If experts can encode associative relations between virtually all chunks within their domain of expertise, then the concept of chunk independence would not apply. Second, and more important, skilled individuals acquire skills to associate chunks of presented information with internal cues organized in retrieval structures. LTM storage mediated by this mechanism would not require associations formed between presented chunks. More recently, Ericsson et al. (2000) showed how LTWM mechanisms developed to support expert performance can transfer and lead to superior performance in traditional memory tasks involving domain-related information.

Cowan acknowledges that traditional STM tests can be influenced by LTM storage and techniques such as rehearsal, and suggests that this type of performance yields compound, "not pure" estimates of STM capacity. He claims that it is possible to experimentally control for these contaminating influences, using experimental procedures that have been designed to prevent participants from rehearsing and encoding presented items and their associations in LTM. Cowan argues that memory testing with these experimental procedures will uncover pure STM with a universal capacity of four independent chunks.

We question the effectiveness of procedures advocated by Cowan to eliminate storage in LTM and restrict recall to independent chunks from pure STM. Many of the "prototypical" pure STM studies Cowan cites in Table 2 involve recall of meaningful stimuli, such as unrelated words, presented while subjects are engaged in rehearsal suppression. To test how effectively these procedures control encoding, we examined how skilled and expert participants performed under such conditions. In contrast to the studies cited by Cowan, rehearsal suppression during encoding had little or no effect on skilled and expert performers. They were able to encode and to form associations between different pieces of presented information in LTM (see Ericsson & Kintsch 1995 for a review), even when the information was auditorily presented (Chase & Ericsson 1981; Ericsson & Folson 1988).

Another approach proposed by Cowan for minimizing storage in LTM and assessment of pure STM involves examining memory for presented material that cannot be easily be rehearsed. However, studies of expert performance are replete with examples of experts who exhibit superior recall for information that cannot be easily rehearsed, such as chess positions.

The ability to rapidly encode and integrate information in LTM is not restricted to skilled performers and experts in their domains of expertise. During the initial trials of memory testing (prior to the build-up of practice interference), storage in LTM is considerable even with typical stimuli and college students (Ericsson & Kintsch 1995). With more sensitive recognition based techniques relational storage in LTM is observed even during repeated testing (Jiang et al. 2000). It is doubtful whether Cowan's recommended procedure can reliably prevent relational encoding of presented information with storage in LTM during a memory test and thus indirectly guarantee independence of recalled chunks.

To rigorously test the validity of Cowan's fixed chunk limit of "pure STM," we must identify task environments in which successful performance requires independence of chunks. When tasks prevent individuals, including highly skilled performers, from relying on associations between chunks, the evidence is inconsistent with Cowan's fixed limit of four independent chunks. For example, during mental addition of long sequences of numbers individuals must maintain the running sum as individual digits that can be updated and changed independently of each other. Skilled abacus operators hold running sums of three or four digits while performing mental computation. This number of independent chunks is consistent with Cowan's estimate of "pure STM" yet when operators engage in deliberate practice they can increase the number of digits of the running sum by one digit per year (Stigler 1984). Skilled abacus operators can expand their memory performance to 14–16 digits with minimal disruption by rehearsal suppression or preloading STM with unrelated information (Hatano & Dsawa 1983). They maintain efficient access to independent digits and can recall up to 10 digits equally fast in forward or backward order. This large expansion of the functional capacity is consistent with the second type of acquired mechanism of LTWM that involves recency-based encodings (Ericsson & Kintsch 1995).

Expert performers' ability to modify their functional "pure STM" capacity in a specific domain raises issues about the gener-

alizability of fixed limits of capacity measured by independent chunks. And our point is not restricted to experts. Expert performance results from years of gradual skill acquisition (Ericsson 1996). Similar types of LTWM mechanisms, albeit in less refined form, mediate the extended development of many types of skills. Ericsson and Kintsch (1995) showed that the same type of LTWM mechanisms mediate reading and other everyday activities mediated by comprehension. When individuals perform tasks involving comprehension, they encode relevant, associatively related information to guide their performance. Associative encoding and integration of encountered information is an essential part of this process. If most cognitive activities in ecologically valid situations do not involve storage of independent chunks, it is unlikely that "pure STM" capacity limits based on independent chunks will be relevant predictors of performance.

Working memory capacity and the hemispheric organization of the brain

Gabriele Gratton,[a] Monica Fabiani,[a] and Paul M. Corballis[b]

[a]*Department of Psychology, University of Missouri-Columbia, Columbia, MO 65203;* [b]*Center for Cognitive Neurosciences, Dartmouth College, Hanover, NH 03755.* **{grattong; fabianim}@missouri.edu**
paul.m.corballis@dartmouth.edu
www.missouri.edu/{~psygg; ~psymf}

Abstract: Different hypotheses about the mechanisms underlying working memory lead to different predictions about working memory capacity when information is distributed across the two hemispheres. We present preliminary data suggesting that memory scanning time (a parameter often associated with working memory capacity) varies depending on how information is subdivided across hemispheres. The data are consistent with a distributed model of working memory.

The Cowan target article emphasizes the limited capacity of working memory, and presents a varied and significant body of evidence indicating that working memory capacity is approximately four items. Over the last two decades, researchers have become increasingly interested in understanding the relationship between cognitive function and the brain. A way to begin addressing the issue of what type of brain mechanisms underlie the fundamental limitation of working memory capacity emphasized by Cowan and previous investigators (e.g., Miller 1956) is to consider the interaction between working memory capacity and the hemispheric organization of the brain.

This interaction may take different forms. According to one idea working memory is a single, unified resource, or "module." This suggests that working memory may be "localized" in one hemisphere only (for instance, in the left hemisphere) rather than being distributed across the two hemispheres ("unitary view"). If this were the case, distributing information to be held in working memory across the two hemispheres would not be advantageous, and may in fact be deleterious in some cases. A variant of this view is that a number of working memory systems – each specialized for different types of materials – coexist, and each of them is localized (absolutely or relatively) in one hemisphere (e.g., left hemisphere for verbal material and right hemisphere for spatial material).

A different idea is that working memory can be envisioned as a distributed system ("distributed view"). In this case, the circuit supporting working memory function could be partly implemented in one hemisphere and partly in the other, even for the same type of stimulus material. Hence, distributing information across the two hemispheres may facilitate its activation and/or maintenance, leading to an increase in working memory capacity.

Finally, as indicated by Cowan, the limitation of working memory capacity may not be due to a specific limitation in the mechanisms used to activate and/or maintain information. Rather, this limitation may result from a general property of information processing systems related to the computational requirements needed to achieve distinguishable mental representations of objects ("computational view"). According to this view, whereas each hemisphere may independently have a memory capacity of four, the combined memory capacity across the two hemispheres would still be four (at least in normal, neurologically intact young adults).

These three views of working memory (which are neither mutually exclusive nor exhaustive) lead to different predictions about what would happen if a working memory task involved presenting the information exclusively to one or the other hemifield (and therefore to one hemisphere) or distributing it across both hemifields (thus dividing the information between the hemispheres). Specifically, the distributed view would predict an increase in capacity when information is distributed across the two hemifields, whereas the unitary and computational views would not. In addition, large differences in capacity between left and right visual field presentations would be more consistent with the unitary than with the computational view.

For this logic to be applicable it is important to be able to manipulate the hemisphere in which the information is activated and/or maintained. Various approaches to this problem can be considered, including the type of material to be maintained (e.g., verbal vs. spatial) and the location in space where the stimulus is presented (e.g., left vs. right visual field–divided field paradigm). In a series of recent studies (Fabiani, in press; Gratton et al. 1998) we have obtained evidence for the hemispheric organization of visual memory. The basic paradigm used in this work involved: (1) the presentation of information to one visual hemifield at study (and therefore, at least initially, to one hemisphere of the brain); and (2) the subsequent testing of memory for the item when it is presented at different locations within the same or the opposite hemifield. Typically, in these studies we observed a reduction in performance when the stimulus was presented in a different hemifield during study and test, compared to conditions in which the stimulus was presented within the same hemifield during both study and test. Further, if the stimulus was presented centrally (and therefore to both hemispheres) at test, the brain activity elicited during the test phase was systematically lateralized (i.e., larger on the left or on the right) depending on the hemifield where the stimulus had been studied.

These data suggest that the divided field paradigm can be used

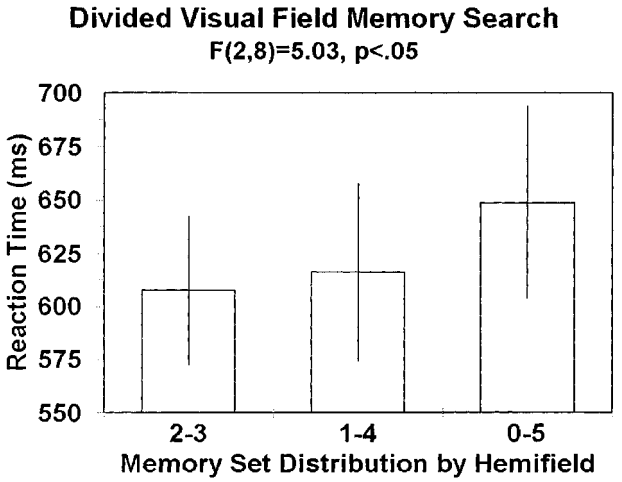


Figure 1 (Gratton et al.). Average reaction times for correct positive trials in the divided visual field memory-search paradigm, as a function of the distribution of memory set items across the left and right visual field (2–3: two items in one hemifield and three in the other; 1–4: one item in one hemifield and four in the other; 0–5: zero items in one hemifield and five in the other). Error bars indicate the standard error of the mean for each condition.

to manipulate the hemisphere that holds information to be maintained in working memory. In a pilot study, we applied this logic to the question of whether distributing information across the two hemifields may improve working memory performance. We used a variant of the memory search paradigm (Stemberg 1966) in which the memory set stimuli were presented one item at a time, either to the left or to the right of a central fixation cross, in an unpredictable fashion. The stimuli consisted of patterns of randomly oriented lines to minimize verbal rehearsal. Each item was presented for 200 msec, with 800 msec ISI.

Following the presentation of the memory set, a test stimulus was presented centrally. This could be one of the five memory-set stimuli or a similar unstudied stimulus. The average reaction times for the correct positive("yes") responses to test items are presented in Figure 1. The data indicate a performance advantage (i.e., faster reaction time) when the memory set items were divided across the two hemifields compared to those conditions in which the memory set items were all presented to the same hemifield.

In interpreting these data, it is important to note that the test stimuli were physically similar for the three conditions presented in Figure 1, and that the total memory set always consisted of five items.

Although preliminary, these data suggest that dividing information between the two hemifields may improve memory-scanning time, which has been traditionally considered as related to working memory capacity. These data are consistent with a distributed view of working memory.

A temporal account of the limited processing capacity

Simon Grondin

École de psychologie, Université Laval, Québec GIK 7P4, Canada.
simon.grondin@psy.ulaval.ca **www.ulaval.psy.ca**

Abstract: A temporal account of the mental capacities for processing information may not be relevant in a context where the goal is to search for storage capacity expressed in chunks. However, if mental capacity and information processing is the question, the time issue can be rehabilitated. A very different temporal viewpoint on capacity limit is proposed in this commentary.

Using a reasonable set of boundary conditions for testing the chunk "capacity," Cowan has helped distinguish between what belongs to a pure storage estimate and what is a compound measure of storage. His impressive review of evidence defending the 4-chunk thesis comes from direct and indirect empirical facts in cognitive psychology, along with evidence from mathematical models and physiological data. Beyond the overall convincing support for the 4-chunk notion, Cowan returns the reader to the cause of this capacity limit, which is reported to be attention. In the process however, time-based interpretations were quickly discarded. It is argued below that the temporal option should and can be used for describing the capacity limit.

What are the magical chunk number and chunking for? We are getting closer to a pure estimate of the capacity limit, as expressed by the number of chunks. From 7 ± 2 which was indeed 3 to *15*, in Miller (1956), it passed to 4 ± 1, which remains nevertheless 2 to 6, in the target article. The numbers have changed and the margin of error is narrower, but scientists eager to have a reliable *law* or a *constant* for mental life to work with will probably remain unsatisfied.

However, although the numbers have changed, the chunking notion has not. Indeed, beyond any estimation of maximum chunks that can potentially be processed simultaneously (perhaps in very rapid, but tractable, succession), we are dealing with a fundamental principle of mental processes: fragmenting information. We are looking for a magical number, an integer, perhaps because we actually have a simple mind that imposes simplicity on the understanding of phenomena, and also perhaps because it would be more useful to work on firm ground rather than rely on the capricious fluctuations of malleable chunk sizes or content.

We may not have a definite magical number yet, but we do have a magical principle to offer to answer the basic capacity question, which is: How can the connection between what is made available by the sensory systems and the organization of it by the brain (or higher processing structures) become as efficient as possible in capturing the essence of available information? Miller (1956) promoted this fundamental chunking/recoding principle: "when . . . we want to remember, we usually try to rephrase [it] 'in our own words'" (p. 95). This is not without recalling notions such as the assimilation mechanism assumed by Piaget in the adaptation process; or the use of field theory in Gestalt, where the brain's electrochemical forces, which underlie structured fields, transform sensory data and are transformed in the simplest way possible.

Representations of mental capacity. Between what is already stored in the system and new inputs, it looks like something occurs, and that an intermediate variable is needed to describe this state. The name of this variable indicates the researchers' stand. They call it memory – short-term memory, in the target article – and the question that they most often address concerns its storage capacity. In this context, a capacity hypothesis such as the duration of an item in this memory without rehearsal may simply be irrelevant. The idea here is that terms like storage and capacity often refer to the notion of volume, that is, to a spatial representation. Duration cannot express a notion having a spatial connotation. If mental abilities or mental processes are the issue, then storage space is only one of a number of potential answers, and a duration hypothesis – which may or may not involve storage time – can possibly be rehabilitated. It is indeed somewhat ironic to discard a duration explanation from a short-term memory issue, given that the word "term" itself expresses duration. This "term" should, instead, be quantified, if not with a time unit, at least with a descriptor of a fading rate.

Whether they are Working/Active/Immediate/Short-Term (WAIST?), all such varieties are in fact called memory. The word "memory" makes sense if information is viewed as being retained or stored for some duration, rather than being present, available or circulating. However, although "WAIST memory" favours permanent storage, storing within "WAIST memory" is not the purpose: processing is.

In Cowan's theoretical framework, the capacity issue enters a new phase once the limit is said to depend on attention, more specifically, on the focus of attention. There are several ways of expressing the involvement of attention in information processing. Focus is one, but one that may also carry a "spatial" connotation, that is, visuo-spatial attention, which, along with the storage capacity notion, serves to reinforce spatial representation.

Attention limits are ultimately reported to be responsible for producing chunk limits. If anything, chunking reveals more about attention than about memory. Chunking is the principle by which attention mechanisms are accommodated, and processing is made easier.

Duration as information, rather than duration of information. Another vein in the literature of experimental psychology may enliven a debate on capacity limits, where considerations about duration decay remain a persisting potential explanation. Because rehearsal provokes interference, either proactive or retroactive, Cowan reports the decay issue to be unresolvable. Nevertheless, this should not totally disregard attempts to provide a temporal account of the state occurring at some point between sensory stimulation and permanent storage.

In the literature on time perception, there is a concept called the *psychological present* for describing this state. This can be defined as a highly flexible tuning process that is dynamically fitting the temporal width of the field of attention and its phase relations

to the sequential structure of the pattern of events." (Michon 1978, p. 89); or " . . . the temporal extent of stimulations that can be perceived at one given time, without the intervention of rehearsal during or after the stimulation" (Fraisse 1978, p. 205). These authors propose that the present has an average value of 2 or 3 sec, with an upper limit of 5 (Fraisse 1978) or 7 or 8 sec (Micho 1978). The wide variability of estimates is at least as disappointing as is that surrounding a maximum chunk number. Nevertheless, a 2-sec value is at least interesting in terms of recalling the memory span estimate of Baddeley (1986).

Just as for the variability in estimating maximum chunks, variability in estimating the psychological present (in time units) depends on boundary conditions adopted. One way to look at the problem is to refer to the discrimination law of Weber (Grondin, in press). This law states that, for one given sensory continuum, difference threshold ($\Delta\phi$) should increase as a function of the magnitude of the stimulus (ϕ), with $\Delta\phi/\phi$ being a constant, k, the Weber fraction. It is well-known that for very small sensory magnitudes, this fraction is large and gradually becomes smaller, and then stabilizes.

At this point, it is worth taking a look at Weber's law for time perception, more specifically concerning duration discrimination under a condition where the only processing required is to judge the duration of a sensory signal or the empty duration between two brief sensory signals. For duration discrimination, the Weber fraction presents a degree of constancy, but at some point, with longer intervals, some discontinuity also occurs (as in boundary condition 3 in Cowan's target article): the fraction becomes higher (see for instance, Fig. 1 in Fraisse 1978). This discontinuity occurs at some point between 1 and 2 sec, and can be interpreted as a point marking the upper limit of the psychological present span, that is, as temporal factor accounting for capacity limit.

Another operational definition can also be adopted for describing a critical limit in temporal processing. One strategy for reducing difference threshold when the time intervals to be discriminated are long is to adopt a counting strategy, that is, using chunks. One way of defining the upper limit of psychological present is to look at the point from which chunking becomes a useful strategy. This point can be estimated by examining two functions relating difference threshold and time, one when counting (chunking) is permitted, and one where subjects refrain from counting. Such functions intersect at about 1.18 sec (Grondin et al. 1999). When intervals are longer than that value, using chunks is a more valuable option; and this critical point can be viewed as an upper limit of the psychological present span. With optimal chunking, this value may be lower, but this remains to be demonstrated. A formal analysis of the consequences of fragmenting intervals into subintervals, thus, of adopting chunking strategies in estimating time, is presented in Killeen and Weiss (1987).

Concluding remarks. Cowan has made a commendable contribution to current investigations of processing capacity by connecting multiple evidence of the 4-chunk storage. Nevertheless, although the amount of information expressed in stored chunks is dimensionless, it may suggest a type of representation of mental capacity that is advanced so as to rule out viewing potential temporal interpretations for processing capacity. Both views talk about a "limited span of apprehension of reality" or an "active state of consciousness." Confusion may arise from calling it a memory.

Using terminology such as short-term memory, in opposition to long-term memory, is certainly an oversimplification of physical time (*term*), which is a continuous variable. Information is in a state varying from 0 to 100% of processing availability. Attention mechanisms, which probably respond to sensory-mode requirements and particularities, activate pieces of information for making it available, present. Four such pieces of information can potentially become available without impairing performance.

Processing capacity limits are not explained by storage limits

Graeme S. Halford,[a] Steven Phillips,[b] and William H. Wilson[c]

[a]*School of Psychology, University of Queensland, Brisbane 4072, Australia;* [b]*Information Science Division, Electrotechnical Laboratory, Tsukuba 305, Japan;* [c]*Department of Computer Science and Engineering, University of New South Wales, Sydney, New South Wales, Australia.* **gsh@psy.uq.edu.au**
stevep@etl.go.jp billw@cse.unsw.edu.au
www.psy.uq.edu.au/gshm www.cse.unsw.edu.au/~billw
www.etl.go.jp/etl/ninehi/stevep@etl.go.jp/welcome.html

Abstract: Cowan's review shows that a short-term memory limit of four items is consistent with a wide range of phenomena in the field. However, he does not explain that limit, whereas an existing theory does offer an explanation for capacity limitations. Furthermore, processing capacity limits cannot be reduced to storage limits as Cowan claims.

In his excellent review, Cowan concludes that short-term memory storage is limited to four items, noting that this corresponds to the limit in processing capacity defined by Halford et al. (1998). Furthermore, his conclusion that the limit is in the number of integrated objects, independent of the complexity of each, agrees well with the observation of Halford et al. (1998) that humans are limited to relating four entities, irrespective of their complexity. However, these correspondences do not imply that processing limits can be subsumed under storage limits, as Cowan claims.

The fact that the size of the limit is four in both cases is not a strong argument for identification because, given that the limit is small, the same number could occur in both contexts by coincidence. Alternatively, storage and processing systems could be distinct but with equal capacities to facilitate transfer from one to the other. There are a number of reasons why processing cannot be subsumed under storage. To take a straightforward example, there clearly is a difference between simply holding the numbers 7 and 4 in short term store, and adding them to yield the sum, 11. In general storage, in the sense of internal representation, is a prerequisite for processing, but cognitive processing cannot be reduced to storage. Furthermore higher cognitive processes require representations that have properties beyond those required for storage, including omni-directional access and analogical mapping (Halford et al. 1998).

Cowan's position is that a concurrent short term memory load can be held in the activated portion of long term memory while other information is being processed in the focus of attention. Lack of interference between processing and short-term storage is explained because the focus of attention can be devoted to either storage or processing, but need not be devoted to both at once. However this still implies that storage and processing are distinct, and also implies there would be no tradeoff between the two. It is not fundamentally different from the position of Halford et al. (1998).

Cowan offers no explanation for the limit he observes in storage capacity, whereas Halford et al. (1998) offer a natural explanation for processing capacity limits. In this model, conceptual complexity is defined by the arity, or number of arguments that can be bound into a single relation. Human adults are typically limited to processing one quaternary relation in parallel. Each component of the relation is represented by a vector, and the binding is represented by the tensor product of the vectors. Thus, the binary relational instance larger (elephant, mouse) is represented by $v_{larger} \times v_{elephant} \times v_{mouse}$. The rank of the tensor product is one more than the arity of the relation. The more complex relations are represented by tensor products of higher rank, the greater complexity of which explains why more complex relations are associated with higher processing load. However, the size of the component vectors has much less effect on processing load, so the fact that the limit is not related to the size of the entities is also explained. Thus, in terms of our relational model, there is a limit

on tensor rank entailed by the rapid growth of the number of tensor units as rank increases. Given that a short-term memory store of capacity 4 is connected to a tensor-like system for processing, the limit of 4 on store size is a consequence of the fact that for most cognitive tasks, processing of the objects in the store is a necessity.

The links between storage and processing phenomena are worth exploring. In section 2, Cowan argues that the unity of conscious awareness implies the contents of attended channels should be integrated or combined. Similarly, category clusters (discussed in sects. 2.7 and 3.4.2) imply a link between instances of the category. Cowan further contends, in section. 3.1.3, that the short term storage limit is observed only with items recalled in correct serial positions. Given that the slots of relation are identified, serial position can be coded as a relation ordered-items (item 1, item 2, item 3, item 4). The observation of no limit with free recall would then suggest that it is ability to represent the relation, rather than the items, that is subject to the limit. This would appear to be consistent with the relational complexity theory of Halford et al. (1998). Furthermore, it clearly points to explaining storage limits in terms of complexity of relations that can be represented. This would also explain the finding of Nairne (1991, referred to by Cowan in sect. 3.4.3) that errors occur up to three positions from the correct position. The reason would be that the items are represented as a quaternary relation, which contains only four slots. The further finding, in section 3.4.5 that participants could predict the seventh item from items 3, 4, 6 may also indicate that the task is represented as a quaternary relation.

These phenomena indicate links between entities that are important, but the nature of these links is not really clear, and the issue is clouded by the lack of a well specified theory in Cowan's paper. Some properties of relational knowledge defined by Halford et al. (1998) seem to be involved in the phenomena discussed above, but it is not clear that they all are. We could define the relational instances fruit (apple, banana, orange, pear) and fruit (lychee, pineapple, passionfruit, guava), and so on. Organizing memory storage as quaternary relations in this way would account for recall of items in clusters of four. However, it would also predict a lot of other properties of relational knowledge that Cowan has not demonstrated. For example, relational knowledge has the property of omni-directional access (Halford et al. 1998) which means that, given any $n-1$ components of a relational instance, the remaining component can be retrieved. Thus, given the quaternary relation proportion (4, 2, ?, 3) we can determine that the missing component must be "6" because it is necessary to complete the proportion 4/2 = 6/3. However, it is far from clear that category clusters share this property. If given a list [apple, banana, ? pear] there is no particular reason why we should recall "orange." Thus category clusters do not entail the kind of constraints that are entailed in relations. Another property of relational knowledge is that analogical mappings can be formed between corresponding relational instances (Holyoak & Thagard 1995). Again, it is not clear that analogies can be formed between category clusters.

Storage is not a simple, unitary matter, but can take many forms. Furthermore, the form in which information is stored affects the form in which it is processed. Some of the possibilities, together with possible implementation in neural nets, are:

1. Item storage – implemented as a vector of activation values over a set of neural units.

2. Associative links between items, implemented as connection weights between units in different vectors.

3. Superposition of items – implemented as summation of item vectors. This is tantamount to a prototype.

4. Superimposed items bound to a category label, such as fruit(apple) + fruit(banana) ± fruit(orange) + fruit(pear). This is equivalent to a unary relation and can be represented by a Rank 2 tensor:

$$v_{fruit} \times v_{apple} + v_{fruit} \times v_{orange} + v_{fruit} \times v_{orange} + v_{fruit} \times v_{pear}$$

Item-position bindings: ordered-fruit (first, apple) + ordered-fruit {(second, orange) +, . . . , + ordered-fruit (fourth, pear)}. This is a binary relational instance and can be implemented by the tensor product

$$v_{ordered\text{-}fruit} \times v_{first} \times v_{apple} + v_{ordered\text{-}fruit} \times v_{second} \times v_{orange} + \ldots + v_{ordered\text{-}fruit} \times v_{fourth} \times v_{pear}$$

Binding items into n-ary relations where n has a maximum value 4. This can be implemented by a tensor up to Rank 5:

$$v_{fruit} \times v_{apple} \times v_{orange} \times v_{pear}$$

These representations have different characteristics. They permit different retrieval operations, and impose different processing loads. More important, at least some of these properties can be captured by neural net models. The Rank n tensor would explain why processing load increases with the number of entities related, and consequently suggests why the capacity limit tends to be low. However, the earlier representations are not sensitive to processing load in this way. It should be clear from these examples that storage and process are intimately related, and that a theory of capacity must include both aspects of computation. However, while their interaction may be complex, it is not arbitrary. Our theory specifies a unique set of properties for processes involving relations of different antics.

Conclusion. Cowan has done the field a great service by showing that a broad range of observations is consistent with the limit of four entities that had been proposed previously by Halford et al. (BBS, 1998). However his claim to reduce processing capacity to storage capacity is not substantiated. Furthermore he offers no explanation for the limit, and glosses over the fact that at least one existing theory offers a potential explanation as to why the limit should be small.

Pure short-term memory capacity has implications for understanding individual differences in math skills

Steven A. Hecht and Todd K. Shackelford

Division of Science, Department of Psychology, Florida Atlantic University, Davie, FL 33314. **{shecht; tshackelford}@fau.edu**
www.uni-bielefeld.de/ZIF/heiko.html

Abstract: Future work is needed to establish that pure short-term memory is a coherent individual difference attribute that is separable from traditional compound short-term memory measures. Psychometric support for latent pure short-term memory capacity will provide an important starting point for future fine-grained analyses of the intrinsic factors that influence individual differences in math skills.

Cowan presents a clear and convincing theoretical case for the fixed capacity limit of three to five chunks in the focus of attention. Cowan has significantly advanced the field of memory research by providing a cogent analysis of the conditions that must be met in order for memory storage capacity to be measured accurately. Cowan presents an organized and impressive array of empirical data corroborating his theoretical claims. Cowan's claims rely on the idea that "purer" estimates of storage capacity in adults can be derived from existing sources of evidence. These sources of evidence come from various methodologies, and converge on the conclusion that a smaller than previously thought chunk limit exists in the focus of attention.

Cowan's analysis provides researchers with a promising tool for relating capacity limits to individual differences in various kinds of human abilities; in this commentary, we address mathematical thinking. Cowan distinguishes between memory measures that do, or do not, control for noncapacity-limited mechanisms. "Compound short-term memory" tasks capture both memory storage capacity and other sources of variance, such as strategic processing. An important limitation of traditional compound short-term

memory tasks is that the relative contributions of both pure short-term memory and other noncapacity limited mechanisms to variability in math skills are usually not readily determined. Only compound memory tasks have been systematically investigated as capturing memory processes that contribute to individual differences in math skills (see Geary 1993, for a review)

There are at least three aspects of a complete account of individual differences in math skills (Hecht 1998). First, it is necessary to focus on specific subdomains of math skills (e.g., simple arithmetic, fraction estimation), because different factors may influence each subdomain of math ability. Second, the unique mathematical knowledge (e.g., counting knowledge) needed to carry out specific problems in a subdomain should be determined. Third, the contributions of intrinsic factors such as memory capacity on the efficiency with which mathematical knowledge is acquired and carried out should be investigated. Another characteristic of a complete account of variability in math skills is a description of the relative contributions of biological and environmental mechanisms that influence the development of math skills (Geary 1995). As Cowan alluded, pure short-term memory capacity may be determined solely by biological factors optimized by adaptive processes in human evolution. Quality of math instruction would be an example of an environmental mechanism.

Cowan's target article suggests important avenues of future research that might lead to a complete account of variability in math skills. The first suggested line of research is the nature of individual differences in pure short-term memory capacity. The relations between pure short-term memory capacity and variability in academic performance can be investigated only if pure estimates of capacity can be measured as a distinct individual difference attribute. Cowan notes that individual differences in capacity limits appear to exist in seemingly disparate tasks such as Sperling's (1960) full report task, Cowan et al. (1999) unattended speech task, and Luck and Vogel's (1997) visual storage capacity task. Performance on these seemingly disparate tasks should correlate if they measure the same underlying pure short-term memory construct.

Confirmatory factor analysis (CPA) can be used to support empirically pure short-term memory capacity as a unique domain of memory ability. Based on Cowan's analysis, it is likely that pure short-term memory and compound memory tasks will yield separate, though correlated, factors. The constructs should be correlated to the extent that compound short-term memory task performance is influenced by pure short-term memory capacity. An important benefit of using latent variables is that correlations among factors can be observed while controlling for sources of variance associated with task specific level and direction of effort or attention. CPA also can be used to determine whether separate versus central pure short-term memory capacity exists. For example, CPA can be used to determine if visual versus verbal pure capacity tasks yield separate or singular constructs of memory storage capacity.

If the construct validity in the psychometric sense of pure short-term memory tasks is empirically established, then another line of research could examine whether individual differences in the capacity limit of chunks in the focus of attention contributes to math ability. Although relations between memory-related latent factors and emerging variability in math outcomes have a been demonstrated (see Hecht et al., in press), these predictors are nor as fine-grained as the measures suggested by Cowan. That is, an important limitation of extant research relating memory processes to variability in math skills is that observed correlations do not indicate which aspects of memory performance influence specific aspects of math ability (Geary 1993). Once separate latent variables for pure short-term memory capacity and compound short-term memory have been identified, the relative contributions of these factors to individual differences in math skills can be assessed.

Thus, Cowan's analysis suggests important starting points for finegrained investigations of relations between intrinsic memory abilities and variability in specific aspects of mathematical ability. One place to start is suggested by Cowan's speculations regarding pure short-term capacity size and performance on Logan's (1988, experiment 4) alphabet arithmetic task. The alphabet arithmetic test is considered to be an analog measure of the acquisition of simple arithmetic knowledge (Logan 1988). Cowan speculated that problems with addends of 1–4 can be visualized (i.e., held in the focus of attention) more clearly while problem solving, because the addend sizes correspond to the number of chunks of information that can be held in the focus of attention. In contrast, performance on problems with addends of five or more may be hindered by pure short-term capacity limitations. Presumably, individual differences in pure short-term memory capacity should be associated with the effects of addend size on alphabet arithmetic performance.

Obtained relations between pure short-term memory capacity and variability in math skills also may help disentangle the influences of biological and cultural factors on math attainment. Geary (1995) makes a distinction between biologically primary abilities and biologically secondary abilities. Biologically primary abilities are found cross-culturally and are designed by natural selection in our evolutionary past. Biologically secondary abilities are not found in all cultures and require sustained formal training (e.g., reading, advanced calculus). It is likely that mean estimates of pure short-term memory capacity size, and degree of individual differences in that construct, are uniformly found across cultures. In his description of teleological accounts of a pure short-term memory capacity limit, Cowan reviews evidence from several sources suggesting a plausible evolutionary function of a very limited capacity of chunks in the focus of attention. For example, Cowan summarizes evidence by Kareev (1995) that a limited pure short-term memory capacity assists in the efficient detection of correlations between features in the physical world. A limited pure short-term memory capacity, shaped by evolutionary forces, may currently be "co-opted" for many contemporary tasks such as biologically secondary mathematical problem solving skills. Cowan's analysis suggests a line of research for investigating potential indicators of co-optation in the domain of math skills. Co-optation of biologically primary memory ability would be suggested by observed correlations between pure short-term memory capacity and biologically secondary math skills.

Cowan's target article should stimulate important avenues of future research toward demonstrating psychometric support for separate constructs of pure short-term memory and compound memory capacity. Current research focusing on individual differences in mathematical thinking has much to gain from the kind of fine-grained analyses of memory capacity suggested by Cowan. The predictive validity of latent pure short-term memory capacity would provide important progress toward understanding the biologically primary factors that influence variability in math skills.

Dual oscillations as the physiological basis for capacity limits

Ole Jensen[a] and John E. Lisman[b]

[a]*Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology, Helsinki 02015 HUT, Finland*; [b]*Department of Biology, Volen Center for Complex Systems, Brandeis University, Waltham, MA 02254.*
ojensen@neuro.hut.fi **lisman@brandeis.edu**
www.boojum.hut.fi/ojensen
www.bio.brandeis.edu/lismanlab/faculty.html

Abstract: A physiological model for short-term memory (STM) based on dual theta (5–10 Hz) and gamma (20–60 Hz) oscillation was proposed by Lisman and Idiart (1995). In this model a memory is represented by groups of neurons that fire in the same gamma cycle. According to this model, capacity is determined by the number of gamma cycles that occur within the slower theta cycle. We will discuss here the implications of recent reports on theta oscillations recorded in humans performing the Sternberg task. Assuming that the oscillatory memory models are correct, these findings can help determine STM capacity.

In reading the target article by Cowan it is evident that it is problematic to determine the STM capacity from psychophysical experiments alone. The main problem is to design experiments in which the influence of chunking and long-term memory are controlled. The framework of oscillatory memory models (Jensen & Lisman 1996; 1998; Lisman & Idiart 1995) may allow the capacity of STM to be tested more directly. In these models multiple representations are assumed to be kept active by a multiplexing network where the dynamics are controlled by nested theta (5–10 Hz) and gamma (20–80 Hz) oscillations. A memory is represented by a group of neurons that fire in the same gamma cycle. The set of memory representations is sequentially reactivated, one representation per gamma cycle, in each theta cycle. Hence, the number of gamma cycles per theta cycle determines the capacity of the memory buffer.

The retrieval time from STM is measured by the Sternberg method in which a set of *S* items is presented to a subject. After a few seconds retention the subject must press a button to indicate whether a probe item matched one of the items on the list. The models predict that the gamma period, T_{gamma}, determines the increase in reaction per item (the Sternberg slope).

In the initial model it was proposed that the Sternberg slope equaled T_{gamma} (Lisman & Idiart 1995). In later work, which attempted to account for the full distribution of the reaction time data, a correction term was introduced (Jensen & Lisman 1998). Since memory scanning cannot be initiated in the period of the theta cycle when the *S* items are activated a wait-time comes into play. The average wait-time is $\frac{1}{2}ST_{gamma}$. Hence the corrected contribution to the retrieval time is $\frac{1}{2}ST_{gamma} + T_{gamma}\,S = \frac{3}{2}T_{gamma}S$ resulting in the Sternberg slope: $\frac{3}{2}T_{gamma}$. The Sternberg slope has been determined to 35–40 msec/item, hence $T_{gamma} \approx$ 25msec ($f_{gamma} = 40\ Hz$) in multiple psychophysical studies.

While this strategy may make it possible to estimate T_{gamma} psychophysically, there are not yet clear strategies for determining the frequency of theta psychophysically. Thus, we cannot suggest how the framework of oscillatory models would provide an independent way of measuring storage capacity on the basis of purely psychophysical data.

The great advantage of the oscillatory memory models is that they can be tested by recording brain rhythms in humans performing STM tasks. So far there are no reports on ongoing gamma activity measured during the Sternberg task. However, in several recent studies it has been possible to measure ongoing theta activity in subjects performing the Sternberg task (Jensen & Tesche 2000; Raghavachari et al. 1999). In these studies the theta frequency during the retention interval of the Sternberg task was measured at 7–8 Hz. The on- and offset of the theta activity correlated with the events of the task, whereas the frequency was independent of the memory load. Applying the theta frequencies measured in these experiments and the gamma frequency estimated from the Sternberg slope, the upper limits of STM is about 5–6 items. Future studies in which both theta and gamma are measured simultaneously will allow a more precise estimate. Several studies hint that this will be possible in the future. In a recent EEG study, Tallon-Baudry et al. (1999) have reported ongoing gamma activity during the retention period of a delayed-matching-to-sample task. The frequency of the gamma activity ranged from 24 to 60 Hz which is too broad to be used for estimating the STM capacity. There have also been attempts to manipulate the frequency of the gamma rhythm. By delivering periodic auditory stimuli in the gamma range (or half the gamma frequency), Burle and Bonnet (2000) sought to entrain the gamma generators in humans performing the Sternberg task. Consistent with the oscillatory memory model they were able to increase or decrease the Sternberg slope by supposedly driving the gamma generators to higher or lower frequencies.

In conclusion, it is now possible to envision experiments which will rigorously test oscillatory models of STM. It may be possible to determine by direct measurement how many gamma cycles occur during a theta cycle while a subject is actually performing a STM task. If this number correlates with storage capacity, as measured psychophysically, it would lend credence to oscillatory models and provide insight into the absolute magnitude of capacity, independent of assumptions about chunking.

The magic number four: Can it explain Sternberg's serial memory scan data?

Jerwen Jou

Department of Psychology and Anthropology, University of Texas-Pan American, Edinburg, TX 78539-2999. **jjou@panam.edu**
www.w3.panam.edu/~jjou

Abstract: Cowan's concept of a pure short-term memory (STM) capacity limit is equivalent to that of memory subitizing. However, a robust phenomenon well known in the Sternberg paradigm, that is, the linear increase of RT as a function of memory set size is not consistent with this concept. Cowan's STM capacity theory will remain incomplete until it can account for this phenomenon.

After almost half a century of being imbued in the doctrine of magic number 7, one is reminded by Cowan's target article that this number may not be as sure a thing as most people have believed it to be. Cowan's target article poses some serious challenges to a long established tenet about memory and provides the field with new ideas and the vigor necessary for the continued advancement of our discipline. However, some well established facts in the short-term-memory (STM) literature seem to be incompatible with Cowan's characterization of the pure STM capacity limit. The concept of pure capacity limit of 4 has to account for these well known STM facts if it is to become a serious competitor of, and a possible alternative to Miller's (1956) theory of STM capacity of 7 ± 2.

Cowan argues that a pure measure of STM storage capacity, uncontaminated by rehearsal (or the use of other mnemonic devices such as chunking) or the mental processing sometimes required beyond simple retaining, can be obtained. The size of this capacity is 4. One crucial defining property of the information held in this pure STM storage capacity, according to Cowan, is its being the target of the momentary focus of attention. Any information within this window of attentional focus is fully activated and fully accessible. In other words, Cowan's concept of a pure STM storage is a memory version of the perceptual process known as subitizing (Jensen et al. 1950; Jou & Aldridge 1999; Kaufman et al. 1949; Klahr 1973; Logie & Baddeley 1957; Mandler & Shebo 1982), which is defined as a rapid, effortless, and yet highly accurate immediate apprehension of the numerosity of a small number (typically under five) of items. It is also in essence the same as

what Ebbinghaus called the window of simultaneous consciousness (cited in Slemecka 1954). Cowan implies in many parts of the target articles and indicates, with support from empirical data, that the reaction time (RT) for retrieving information from memory sets within the size of the pure STM limit remains relatively low and constant, but shows a sudden large increase as the memory set size (MMS) exceeds 4, in parallel with the finding in perceptual quantification of a discontinuity in RT functions from the subitizing range to the counting range (beyond 3 or 4 items) (Mandler & Shebo 1982). But, is there such a parallel?

The concept of a smaller-sized STM storage, which is the momentary focus of attention, or the window of simultaneous consciousness, must account for a very robust phenomenon in STM literature if it is to become a viable theory of STM. The phenomenon is the linear increase of RT as a function of MMS, first brought forth by Sternberg's seminal papers (Sternberg 1966; 1969) and later replicated by countless studies conducted in different contexts and with different test materials (Sternberg 1975). If items reside within the attention-focused window of simultaneous consciousness, why does the time required to access or retrieve an item from this range increase linearly with MMS so consistently and reliably? Cowan presents some cases where the RT for memory retrieval increases little until the MMS exceeds 4. How can the concept of memory subitizing and the data cited by the author in support of it be reconciled with the even larger body of data that is consistent with the Sternberg's concept of serial memory scan?

Furthermore, contrary to the finding cited in the target article of a shallow slope of RT functions associated with MMS of 4 or smaller, and a steeper slope for MMS past 4, a considerable number of studies have shown the opposite pattern of results. In these studies, the RT functions for MMS of 6 or smaller are typically characterized by steep linearity. But as the MMS exceeds 6, the functions become essentially flat (Baddeley & Ecob 1973; Burrows & Okada 1975; Okada & Burrows 1978) Those results were interpreted as suggesting that a serial access mode was in operation for small MSSs, but that a direct access mode is adopted when MSS exceeds 6. Jou (1998) used a fixed set version of the Sternberg task with MMSs varying from 1 to 20 and memory-set items randomly and repeatedly sampled from the 50 U.S. state names. RT increased at about 66 msec/item up to the MMS of 6 and then 1eve1ed off to an insignificant 7 msec/item past MSS 6. This was consistent with the findings of the above-noted studies. The attention focus theme emphasized throughout Cowan's target articles would have predicted results of reversed pattern.

Jou and Aldridge (1999) had subjects estimate the serial positions of letters in the alphabet and the alphabetic distances between two letters. This task can be considered to involve a form of memory quantification of some overlearned magnitude facts. The results were that, for the serial position estimation, the RT/alphabetic serial position functions were linear and steep up to the serial position of 6 or 7, past which the RT functions turned essentially flat. The RT functions for the alphabetic distance judgment showed basically the same pattern except that the steep linear portion was reduced to a magnitude of about 4. Although these results are highly counterintuitive in that memory quantification seems to operate in a serial fashion for small values, but in a parallel fashion for large values, they are consistent with Burrows and Okada (1975), Okada and Burrows (1978), and Baddeley and Ecob (1973) findings. Jou and Aldridge (1999) concluded that there is no memory subitizing, unlike in the perceptual domain. They suggested that there is a fundamental difference between quantification in memory and that in perception because in perception, stimuli are physically present whereas in memory they have to be internally represented, which consumes resources and prevents subitizing from taking place. Perceptual subitizing, according to Jou and Aldridge (1999), is a result of having an overabundance of attentional resources available. In memory, this rarely occurs because of the resource demands made by the mental representation process (the only exception perhaps being when the memory representation and retrieval processes are automatized through overlearning).

Again, Cowan's concept of a small STM window of direct and simultaneous access of information must explain the above data to be a viable theory. Specifically, it has to account for (1) the robust linear increase of RT up to a MSS of 6 (a phenomenon suggesting a lack of simultaneous consciousness for the information) and (2) the leveling off of the RT functions as MSS exceeds 6 (which is the opposite of what Cowan's theory would predict).

Also incompatible with the author's view was the serial position confusion pattern, ironically cited (Nairne 1991; 1992) by Cowan to support his concept of STM capacity. That is, the confusion occurred mostly within 3 or 4 positions of the target item. This contradicts the basic concept of subitizing. If the information processing within the small range of 4 items is like perceptual subitizing, then no errors, or at most a minimum number of errors should occur, because "perfect performance" is a hallmark of subitizing.

Finally, Cowan attempts to define a condition under which a pure STM storage capacity can be measured, that is, one in which the memory system is neither overburdened nor under-taxed. This seems to involve a very delicate balance between too much and too little resource demand. For the theory to be formalized, this delicate balanced condition has to be more clearly spelled out. It is possible that a MMS of 4 is not the STM limit, but an optimal working range of STM. Capacity is the largest possible amount of information that STM can hold. If the focus of attention can be switched, as the author suggested, between two or three blocks of three items each (assuming the chunking is not formed by relying on long-term-memory knowledge, but by using temporal proximity), and if all three chunks are available within a certain short period of time (though not necessarily simultaneously as in the case of focused attention), why can't all three chunks be considered parts of the STM? What is the a priori basis for limiting the STM capacity to the span of focused attention, or simultaneous consciousness? Grouping several numbers together into a chunk by reading these numbers faster as a unit or delivering items at an overall faster rate (as in Waugh & Norman 1965, cited in the target article) increases the total amount of information that can be held in STM by minimizing the loss of information over time. Why can't it be justified to push the STM limit to its maximum by delivering the items at a faster rate? Cowan states that the presentation rate should not be too slow, either. Again, concerning the presentation rate, there seems to be a delicate balance in order to demonstrate a capacity of 4. What this exact rate of presentation is did not seem to have been clearly specified in the target article either.

"Magical number 5" in a chimpanzee

Nobuyuki Kawai and Tetsuro Matsuzawa

Section of Language and Intelligence, Department of Behavior Brain Science, Primate Research Institute, Kyoto University, Kanrin, Inuyama City, Aichi Pref. Japan 484-8506. **{nkawai; matsuzaw}@pri.kyoto-u.ac.jp**
www.pri.kyoto-u.ac.jp/koudou-shinkei/shikou/index.html

Abstract: One of our recent studies has revealed that a numerically trained chimpanzee can memorize a correct sequence of five numbers shown on a monitor. Comparative investigations with humans show very similar patterns of errors in the two species, suggesting humans and chimpanzee share homologous memory processes. Whether or not 5 is a pure capacity limit for the chimpanzee remains an empirical question.

Cowan has proposed a new "magical number 4" in human adults from a careful reconsideration of short-term memory processes. In his theoretical account of the capacity limit, Cowan suggests that the "capacity limit might have become optimized through adaptive processes in evolution" (sect. 4.1). If this is the case, we can expect to find similar memory processes in non-human ani-

mals. Chimpanzees are good candidates for investigating the origins of our cognitive evolution, being the closest relatives of humans among living creatures.

Recently, we found that a numerically trained chimpanzee had a memory for numbers in several aspects similar to humans (Kawai & Matsuzawa 2000b). The chimpanzee called Ai, has more than 20 years of experimental experience. Prior to the memory test, Ai learned to count dots on computer monitor or real objects and to select the corresponding Arabic numerals on a touch-sensitive monitor (Matsuzawa 1985). Ai also learned to order the numbers from zero to nine in sequence, regardless of the inter-integer distance (Biro & Matsuzawa 1999; Tomonaga & Matsuzawa 2000). Utilizing her numerical skills, we set up a memory task. A set of numbers (e.g., 1, 3, 4, 6, and 9) was spatially distributed on a screen. Ai was required to touch the numbers in an ascending order. Immediately after the selection of the lowest number (i.e., 1), all the remaining numbers were masked by a white square. Hence Ai had to memorize the numbers (now masked) accurately to select the correct sequence. She reached more than 95% correct with four numbers and 65% with five, significantly above chance in each case (17 and 4%, respectively). This indicates that she could memorize the correct sequence of any five numbers (Kawai & Matsuzawa 2000b).

The most interesting result concerned Ai's response time. The longest response times were obtained for the first number of the sequence. Response times were shortest for the other numbers, and did not differ from one number to another. Thus, her mean reaction time of first response to a set size of five was 721 msec, and then 446, 426, 466, and 41, respectively, for the remaining four, (now masked) numbers. This pattern of responses is similar to that of humans. For example, mean reaction times of five adult humans in the test were 1,430, 524, 490, 741, 672 msec, for each response, suggesting that both Ai and humans memorize the numbers and their locations before the first response (Biro & Marsuzawa 1999; Kawai, in press).

One may argue that both Ai and humans might use a rehearsal strategy during the longest reaction time preceding the first choice. The accuracy of humans decreased when the numbers were masked 750 msec after the initiation of the trial. However, Ai's reaction times were almost half of those of humans, and they remained approximately the same for the masked and unmasked trials. Although rehearsal constitutes a major cause of compound STM estimate, rehearsal was impossible because Ai's fast reaction times seem incompatible with rehearsal. Thus, Ai's performance in memorizing five items may reflect a "pure capacity limit."

There is a possibility that other mnemonic strategies were involved in the task. For instance, one might suspect that Ai used the configuration of the numbers as possible spatial cues for responding. The procedure ruled out this possibility however, because locations were randomized across trials and all trials were unique in each session, thereby demonstrating that long-term memory did not contribute to the performance. In addition, a detailed analysis of error trials confirmed that neither Ai nor the humans used spatial cues. The majority of errors (84.1% for Ai, 84.5% for humans) consisted in skipping one number only (e.g., selecting 1–3–6 instead of 1–3–4–6–9). The remaining errors were also independent of spatial factors. Most (87.5% for Ai, and 82.0% for humans) consisted of selecting the highest number in the sequence (i.e., 1–9 or 1–3–9 instead of 1–3–4–6–9), regardless of the spatial arrangement of the numbers on the screen. These trials were regarded as showing a recency effect because the highest numbers were the last to be processed in the preplanned sequence. Even more interesting, the frequency of these last numbers was proportional to the size of the greatest number: more errors of this type were made when the sequence contained a 9 as the last number than when it contained an 8, and that tendency remained for the lowest numbers. All these results suggest that, like humans, Ai built up a linear representation of numbers, from 0 to 9, and referred to it in performing the task. Similarity in the error patterns for the two species moreover suggests that their memory systems may share homologous mechanisms (Kawai & Matsuzawa 2000a; in press).

Because masking occurred at the first touch, one might argue that the memory span of the chimpanzee is four instead of five. As demonstrated above, both Ai and the humans had already planned their response at the onset of each trial: there is thus the possibility that the first number was included in the memorized sequence. Even more, according to our recent test for six numbers Ai's performance was about 30%, significantly higher than the chance level (0.8%). We do not deny the possibility that the "pure capacity limit" of the chimpanzee might be less than four. Further experimental studies will be required determine this. Nevertheless, because comparable data were obtained in Ai and the humans, our study strongly suggests that if there are any "pure capacity limit" differences between the two species, they should be quantitative rather than qualitative. The essential point is that humans share quite a similar memory process with chimpanzees.

What forms the chunks in a subject's performance? Lessons from the CHREST computational model of learning

Peter C.R. Lane, Fernand Gobet, and Peter C-H. Cheng

ESRC Centre for Research in Development, Instruction and Training, School of Psychology, University of Nottingham, Nottingham NG7 2RD, United Kingdom. **{pcl; frg; pcc}@psychology.nottingham.ac.uk**
www.psychology.nottingham.ac.uk/staff/{Peter.Lane; Fernand.Gobet; Peter.Cheng}

Abstract: Computational models of learning provide an alternative technique for identifying the number and type of chunks used by a subject in a specific task. Results from applying CHREST to chess expertise support the theoretical framework of Cowan and a limit in visual short-term memory capacity of 3–4 looms. An application to learning from diagrams illustrates different identifiable forms of chunk.

Cowan's theoretical framework (sect. 2) assumes that the "focus of attention is capacity-limited," and that "deliberately recalled [information] is restricted to this limit in the focus of attention." This framework is compatible with the EPAM/CHREST family of computational models, and this commentary highlights the role that a model of learning can play in clarifying the nature of chunks. CHREST (Chunk Hierarchy and REtrieval STructure) is a computational model of expert memory in chess players (Gobet 1998; Gobet & Simon, in press), and is based on the earlier EPAM model (Feigenbaum & Simon 1984) of perceptual memory. CHREST possesses an input device (simulated eye), a visual short-term memory (STM) for storing intermediate results (equivalent to the focus of attention), and a long-term memory (LTM) based around a discrimination network for retrieving chunks of information. Each chunk is learnt from information in the visual field, using the STM to compose information across one or more eye fixations.

The classic recall task (Chase & Simon 1973; Cowan, sect. 3.4.1; De Groot 1946; 1978) has been used to show that subjects recall information in chunks. The task requires the model/subject to observe a display for a set time period, and then reconstruct the stimulus from memory; in simulations, the chunks within the model's STM are used as the reconstructed response. In a study of chess expertise, Gobet (1998) showed how the accuracy of the reconstructed position depends on the number and size of chunks which the model identifies; the size of chunk depends on the level of expertise, but the number can be systematically varied, and a value of 3 or 4 was found to best match the performance of different levels of player, providing further empirical support for the findings of Cowan. Also significant is that the better performance of experts is explained by their use of larger chunks (typically, master chess players recall chunks of twice the size of average club

players), and the number and content of these chunks may be extracted from the model (see also Gobet & Simon 1998; in press)

Chase and Simon (1973) did, however, find that expert chess players appeared to recall more chunks than novices. As discussed in Gobet and Simon (1998), these findings do not contradict the existence of a fixed capacity limit, because additional factors affect the subject's performance; in this case, the number of pieces which the player can pick up. So, are the chunks observed in the subject's performance due to previously learnt information or to other factors relating to the task or cognitive performance? This question may be answered through a simulation of the learning process. The role of learnt knowledge in producing chunks in performance is currently being explored in a problem-solving version of CHREST (Lane et al. 2000a) which learns a diagrammatic representation for solving electric circuit problems. In Lane et al. (2000b) different computational models were analysed based on their respective representational, learning, and retrieval strategies for handling high-level information. From these two studies, it is clear that chunks observed in the model's performance may arise from a number of causes. Three of the more apparent are as follows:

(1) A chunk may be observed in the output because of an explicit representation in the system's LTM, which is the underlying representation used in the EPAM/CHREST family of computational models. For example, Richman (1996) describe a chunk as "any unit of information that has been familiarised and has become meaningful."

(2) A chunk may be observed in the output because the input has matched a stored chunk based on some similarity-based criterion; this is familiar from neural network approaches.

(3) A single chunk may be observed although it is based on a functional composition/decomposition of the stimulus and its subcomponents. For example, subjects may retrieve and store multiple chunks within their STM, but the performance based on these multiple chunks may then give the appearance of a single chunk.

The presence of three distinct processes yielding chunk-like behaviour in such models clarifies how the observational characteristics of chunks inter-relate with learnt knowledge, and hence clarifies the connection between observed and learnt chunks. This connection assists in developing a deeper understanding of the capacity limit, especially in areas where the subject is continuously learning new chunks for composite objects. Most importantly, only by modelling the entire learning history of each subject can we really attempt to probe the content and format of chunks manipulated in STM, and thereby estimate STM capacity.

The focus of attention across space and across time

Brian McElree[a] and Barbara Anne Dosher[b]

[a]*Department of Psychology, New York University, New York, NY 10003;*
[b]*Cognitive Science, University of California, Irvine, CA 92717;*
bdm@psych.nyu.edu bdosher@uci.edu
www.psych.nyu.edu/dept/faculty/mcelree/research.html
www.aris.ss.uci.edu/cogsci/personnel/dosher/dosher.html

Abstract: Measures of retrieval speed for recently presented events show a sharp dichotomy between representations in focal attention and representations that are recently processed but no longer attended. When information is presented over time, retrieval measures show that focal attention and rapid privileged access is limited to the most recently processed unit or chunk, not the last 3–5 chunks that Cowan estimates from various recall procedures.

Cowan presents a diverse array of evidence to support the claim that the focal attention has a capacity of 3–4 chunks. Much of this evidence comes from studies examining processing limits in multielement displays in which all elements are simultaneously displayed. These studies may provide good evidence for the claim that there is a 3-to 4-item limit on the simultaneous coding and reproduction of elements, at least in some domains. Cowan believes the same limit holds for sequentially displayed elements, namely, elements distributed over time rather than space. However, Cowan's estimates are based largely on indirect measurements. Crucially, measures of retrieval speed from studies using sequential presentation provide direct evidence for a distinct representational state associated with the focus of attention that is limited to the most recently processed unit. These measures indicate, contra Cowan, that only one chunk is maintained across a dynamically changing environment. One possibility is that the capacity of focal attention differs for simultaneously available elements arrayed in space, and for representations encountered over time. If Cowan's analysis is correct, perhaps we can attend to more than one simultaneously presented element; however, we do not appear to be able to process more than one temporally extended event.

Cowan's evidence. Cowan forwards, as an estimate of the capacity of focal attention, findings that the number of recalled items often converges on 3–4. However, recall performance is determined by a confluence of factors other than the capacity of focal attention. Undoubtedly, these estimates partly reflect the recall of representations outside focal attention, analogous to the way that serial position functions were classically argued to reflect output from both long-term and short-term. Further, recall is limited by forgetting that occurs over the learning phase and during the recall process (e.g., Dosher & May1998). The number of items recalled, even when the preconditions enumerated by Cowan (sect. 1.2) are met, provides at best an indirect estimate of the capacity of focal attention, and is equated with focal attention primarily by assumption.

Retrieval speed. The claim that focal attention is distinct from more passive memory representations implies that information in focal attention is accessed more immediately than information in a passive state. Measures for access speed can provide direct evidence for distinct representational states if access speed can be measured for memories with different strengths (or probability of access). Unfortunately, RT does not provide pure estimates of retrieval speed because it is affected by memory strength (e.g., Dosher 1984; McElree & Dosher 1989; Wickelgren et al. 1980). However, retrieval speed can be directly measured with the response-signal speed-accuracy tradeoff (SAT) procedure. In this procedure, subjects are cued to respond at some time after the onset of a test probe. With a suitable range of cue times, the full time course of retrieval is evaluated, providing measures of when information first becomes available, the rate at which information accrues over retrieval time, and the asymptotic level of observed performance. The asymptote reflects the probability of retrieval, and provides an estimate of memory strength. When accuracy departs from chance, the rate at which it grows to asymptote jointly measure retrieval speed. More accessible information should be associated with an earlier intercept or faster rate, irrespective of differences in asymptotic accuracy.

Wickelgren et al. (1980) used a probe recognition task to examine SAT time-course profiles for accessing representations in a list of 16 sequentially-presented items. Asymptotic accuracy decreased monotonically with the decreasing recency of the tested item, indicating that memory strength systematically declines as time or activity is interpolated between study and test. Crucially, however, retrieval speed was constant across all serial positions save the last, most recently studied position. Retrieval speed was 50% faster when no items intervened between study and test. The most recently studied item received privileged access. This finding has been replicated with different procedures and materials, including a Sternberg task (McElree & Dosher 1989), a forced-choice recognition task (McElree & Dosher 1993), a paired-associate recognition task (Dosher 1981), and even when the task required judging whether a test item rhymed or was synonymous with a studied item (McElree 1996). Related effects are found in judgments of recency (McElree & Dosher 1993) and the *n*-back task (McElree, in press).

This sharp dichotomy in retrieval speed – fast access for the most recently processed item and a slower, but constant retrieval speed for all less recently processed items – provides direct evidence for two representational states, one associated with focal attention, and the other associated with memory representations outside of the focus of attention. Although rapid access is usually reserved for the most recently processed single item, McElree (1998) found that rapid access accrues to multiple items if they form a chunk or unit: the retrieval advantage extended to three items if they were members of the most recently experienced category. Further, the retrieval advantage is not bound to the last positions in a list, but instead reflects the last cognitive operation (McElree 1997): an advantage is found for a non-recent category when a category cue is used to retrieve and restore items to focal attention.

Conclusions. Collectively, direct measures of retrieval speed indicate that focal attention, associated with especially rapid retrieval, is more limited than the three or four items suggested by the indirect analysis forwarded by Cowan. Measures of retrieval indicate that we are only able to maintain one temporally extended event or epoch in focal attention.

Capacity limits in continuous old-new recognition and in short-term implicit memory

Elinor McKone

Division of Psychology, Australian National University, ACT 0200, Australia.
elinor.mckone@anu.edu.au www.psyc.anu.edu.au/staff/elinor.html

Abstract: Using explicit memory measures, Cowan predicts a new circumstance in which the central capacity limit of 4 chunks should obtain. Supporting results for such an experiment, using continuous old-new recognition, are described. With implicit memory measures, Cowan assumes that short-term repetition priming reflects the central capacity limit. I argue that this phenomenon instead reflects limits within individual perceptual processing modules.

Cowan makes a prediction (sects. 1.2 and 4.3.7) regarding a new circumstance in which the capacity of short-term explicit memory should be limited to 4 chunks. This is where subjects must only indicate, as rapidly as possible, if a particular item had been included in the stimulus set previously, and in which some items would be repeated in the set but other, novel items also would be introduced." Experiment 4 of McKone (1995) used such a design. Each trial presented a single word or pseudoword (e.g., mave), in continuous sequences of 250–300 trials. Approximately 65% of trials presented items new to the experiment; the other 35% were a repeat of an item seen earlier in the list. Repeats occurred at various "lags" (i.e., number of intervening items: range 0–23). The task was old-new recognition to every trial in the sequence, with reaction time for correct old responses as the dependent measure. This experiment appears to satisfy Cowan's general criteria for producing a pure, rather than compound, capacity measure: the fact that half the items were nonsense words and the lack of particular semantic associations between successive words, should limit chunking across items; the presentation rate (2 secs per trial) and the very long lists should limit active rehearsal.

Results (see Fig. 1) were both consistent and inconsistent, with Cowan's predictions. A basic capacity limit of 4 items was supported. In calculating the capacity, my logic is that (1) to be compared with previous items, the current stimulus must be in the focus of attention, (2) any previous item still in the focus of attention should be recognised more quickly as old than an item which has been pushed out, and (3) some sort of discontinuity in RTs should therefore appear at the point at which the capacity limit is reached. Figure 1 indicates that repeats at 0, 1, and 2 intervening items produce noticeably faster RTs than repeats at 3–9 intervening items. This discontinuity between lags 2 and 3 corresponds to a short-term memory capacity of 4 items: at lag 2, the contents of short-term memory are the current word, its first presentation, and two intervening items.

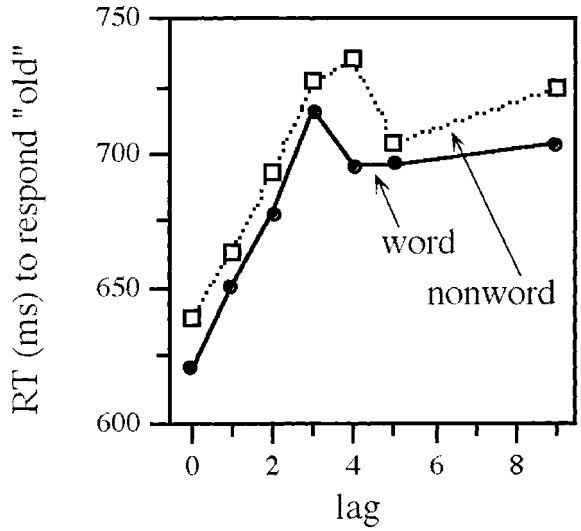


Figure 1 (McKone). Explicit short-term memory in continuous old-new recognition (data from McKone, 1995, Experiment 4; averaged across high frequency and low frequency word conditions).

Cowan's exact prediction for the old-new recognition experiment, however, was that "the mean reaction time should be much faster when the item had been presented within the most recent 3 or 4 items than when it was presented only earlier in the sequence." This suggests that the discontinuity in RTs should take the form a flat function up to 4 chunks (i.e., all items within the focus of attention show equally fast RTs), followed by decay. The data, in contrast, show smooth and rapid decay up to 4 items, followed by much slower decay. Thus, while Cowan's prediction seems to presume that all items in the focus of attention are equally strong, the RT data instead suggest that older items fade from the limited capacity mechanism (at least without active rehearsal). This would seem to reconcile a capacity limit with the demonstrated importance of time-based factors when rehearsal is allowed (e.g., effects of articulation rate on word span tasks).

Turning to implicit, rather than explicit, memory tasks, Cowan cites my demonstration of short-term implicit memory (i.e., an additional short-lived repetition priming advantage in tasks such as lexical decision or word naming) as supporting evidence for a capacity limit. However, it is not at all clear to me that short-term implicit memory is affected by the central capacity limit, rather than by module-specific processing factors. Theoretically, each item in turn in a lexical decision or naming task must fall within the focus of attention to allow a response to be made to that item. Unlike explicit memory tasks, however, no overt comparison of more than one item is ever required, and so it is not necessary that both presentations of the target are simultaneously held in Conscious awareness. Using Cowan's terminology (sect. 4.3.6), an implicit memory task requires only "chunks" (e.g., x,y,z,x), and not "chunks in relation to some concept" (e.g., recently seen [x,y,z,x]), as does explicit memory.

Empirically, McKone (1995) dissociated the apparent "capacity" of short-term repetition priming from that of short-term explicit memory. With written stimuli, the explicit recognition memory experiment described above obtained the 4 chunk capacity limit for both words and pseudowords; however, implicit memory in lexical decision and naming, produced a purported capacity limit of 5 items for words but only 2 items for pseudowords. McKone and Dennis (2000) then demonstrated that implicit "capacity" varies consistently with the demands of visual versus audi-

tory word identification. Unlike (type) written words, the ambiguity of a continuous speech signal is so severe that a word can often be understood only following the arrival of several successive speech segments. Thus, the auditory word identification system might need to retain more items simultaneously active than does the visual word identification system and, since real words often initially sound like nonsense words, pseudowords might need to be retained as long as real words. Consistent with these proposals, short-term repetition effects indicated "capacities" of 4 for words and 2 for pseudowords with written presentation, but at least 5–6 (and possibly more) for both words and pseudowords with spoken presentation. It is not obvious why a single central capacity should be responsible for these very different limits.

Overall, Cowan has provided convincing evidence of a central capacity limit (of 4 chunks) in short-term explicit memory, presumably corresponding to the focus of attention. The data he reviews, however, do not directly address the question of whether or not individual processing modules (perceptual, cognitive or motor) also have their own "capacity" limits outside the focus of attention. Clear theoretical reasons exist for limits on the number of items thatcan be simultaneously active within a single domain since left over activation from previous items will interfere with identification of the currently-presented item: in visual word recognition, many models in fact assume a "capacity limit" of only one item (the current stimulus kills all traces of previous words).

Magical attention

Peter M. Milner

Department of Psychology, McGill University, Montreal, QC, H3A 1B1, Canada. **pmilne@po-box.mcgill.ca**

Abstract: Cowan postulates that the capacity of short-term memory is limited to the number of items to which attention can be simultaneously directed. Unfortunately, he endows attention with unexplained properties, such as being able to locate the most recent inputs to short-term memory, so his theory does little more than restate the data.

Cowan's thorough review of memory-span data provides convincing evidence that in the absence of rehearsal the span is about 4 items. This information is useful, but how much does it further our knowledge to be told that the span is limited because the "focus of attention" can hold only about 4 items at a time?

Cowan adopts Lisman and Idiart's (1995) theory that attention is dependent on oscillations of cortical potentials. A cortical wave of about 10 Hz is supposed to select items from an unlimited short-term store; other wavelets at a frequency of about 40 Hz then select one item each. Assuming that such a process were neurally feasible, it still does not explain why successive waves select the same items, moreover the wave frequencies vary considerably, as Cowan points out. Apart from his uncritical recourse to this theory of attention to explain the magical figure of 4, Cowan confines his discussion to cognitive models, so it may be unfair to criticise the theory on neural grounds. Nevertheless, I believe that consideration of the constraints imposed by neurophysiology may be helpful.

In the first place, it is unlikely that sensory information can be moved to a focus of attention except in a metaphorical sense. Attention may change the intensity with which visual neurons respond to a stimulus (Fuster et al. 1985; Moran & Desimone 1985), but not their location. Furthermore, neurons fired by objects in different parts of the visual field are spatially segregated only in the primary visual cortex. At higher levels of the visual system, neurons have large overlapping fields (Gross et al. 1974; Miyashita 1993). The relative positions in space of the objects they represent do not determine the cortical location of the neurons. Thus individual items in the visual field cannot be selected by spatially directed attention. At this level of the visual path, selective attention must be delivered, presumably via learned connections, to widely scattered neurons. In other words, if attention is to be directed to a book, the attention signal must find the diffuse cloud of neurons that represent visual features common to books.

It is, thus, clear that the neural substrate of attention is not a unitary system that can be pointed like a spotlight or a camera, much less a static process into which peripatetic images can be directed. Apparently it must be a highly organized system of centrifugal paths, every bit as specific as the centripetal sensory paths that it modulates (Milner 1999). At least some of the neural activity corresponding to attention originates in the response planning mechanism where the intentions of the subject are elaborated. Neural representations of objects, related either innately or through prior experience to the task in hand, need to be selected by specific facilitation as part of the planned response.

According to the above account, what Cowan calls the focus of attention must vary widely depending on motivation, or the task. Thirst, for example, facilitates the sensory input from all sources of water known to the subject, increasing the probability that if one of them happens to be present it will gain control of the response mechanism. In the absence of a matching sensory input, the dominant thirst-quenching intention facilitates items in long-term memory with which it has acquired associations. The more objects that have been associated with slaking thirst, the more widespread the initial attentional facilitation. Any input amplified by attentional facilitation is then likely to determine the response that is released.

In most memory-span measurements, the subject's task is to recall recently heard or read words. These must be distinguished from all the other words in the subject's vocabulary by short-term changes in the thresholds or connections of corresponding neurons. At a given moment during recall only one word is being released to the motor system for pronunciation (or writing). One problem is to discover why it is not the most recently heard word (which presumably retains the strongest trace), another is to discover why the system can cope with no more than about 4 or 5 words. Are the later words of a series less effectively tagged at the time of storage or does recall of the earlier words interfere with retrieval of the later ones? How important a factor is decay of the trace with time? Can rehearsal of the earlier words ever be completely prevented? It seems to me that the limit of four words is at least as much to do with decay and interference of the short-term synaptic changes than anything related to attention.

Another datum cited by Cowan in support of his theory is that the subitizing limit is about 4. He bypasses the difficult problem of how items simultaneously present within his postulated span of attention are summed to elicit a numerical response. Of course he is not alone in not having solved that problem, but it is at least possible that the process involves discriminating a signal generated by x objects from that generated by $x \pm 1$ objects. The ratio of x to $x \pm 1$ may become too small to permit discrimination when x is greater than about 4. I am confident that most people are able to discriminate instantly between 10 objects and 20 objects without eye movements or counting, which might be interpreted as indicating that at least 20 objects can simultaneously lie within Cowan's focus of attention in some circumstances.

Nothing left in store . . . but how do we measure attentional capacity?

Sergio Morra

Università di Genova, DISA-Sezione Psicologia, 16126 Genova, Italy.
morra@nous.unige.it

Abstract: I compare the concepts of "activation" and "storage" as foundations of short-term memory, and suggest that an attention-based view of STM does not need to posit specialized short-term stores. In particular, no compelling evidence supports the hypothesis of time-limited stores. Identifying sources of activation, examining the role of activated procedural knowledge, and studying working memory development are central issues in modelling capacity-limited focal attention.

Cowan's main thesis is that short-term memory depends on attention. "The focus of attention is capacity-limited . . . Any information that is deliberately recalled, whether from a recent stimulus or from long-term memory, is restricted to this limit in the focus of attention . . . The same general principles of activation and de-activation might apply across all types of code" (sect. 1). So far, so good.

However, Cowan often calls "store" the capacity-limited focus of attention (whereas Cowan 1988, made it clear that it is an activated portion of long-term memory), and leaves as an open issue the existence of "supplementary storage mechanisms, which . . . are time-limited rather than capacity-limited" (sect. 1). I think that an attention-based view of STM entails that it is regarded as an activated part of LTM, which is in turn incompatible with the concept of short-term storage. (Connecting an activation-based central executive with specialized slave stores has been a long-standing problem for Baddeley's [1986] theory.)

"Storage" and "activation" are obviously metaphors, but not so innocent and vague as to be interchangeable. They have empirically distinguishable consequences.

(1) If there are short-term stores, performance should never be improved by occupying them with irrelevant materials (unless this prevents use of inadequate strategies: Brandimonte et al. 1992). In an activation-based system, instead, an irrelevant memory load may sometimes pre-activate structures and thus enhance performance. The latter proved to be the case (e.g., Hellige & Cox 1976; van Strien & Bouma 1990).

(2) Storage models must specify flow of information transfer, whereas activation models must specify time course of activation. Numerous priming paradigms lend themselves to an activation account. Detailed models have been proposed for the time course of activation in both positive and negative priming (e.g., Houghton & Tipper 1994; Neely 1991). I am not aware of equally powerful models of priming as information transfer among stores. As Anderson (1983, p. 21) put it: "The results on associative priming have shown us that the amount of information brought into working memory, at least temporarily, is very large."

(3) If supplementary stores exist, then one should specify either their capacity or duration. Numerous short-term stores had been proposed in the literature, but only for one was a clear capacity estimation made. This was the "articulatory loop," deemed to hold as much phonological material as can be uttered in about 2 seconds (Baddeley et al. 1975; for converging evidence see Baddeley 1986; Hulme & Tordoff 1989; Schweickert & Boruff 1986). Because this seems to be the only advantage gained by the storage view, it is worthwhile to examine it more closely.

Early claims of falsification of a time-limited articulatory store (Morra 1989, 1990; Morra & Stoffel 1991) remained unpublished. Journal reviewers discarded them, sometimes even on the grounds that the results were not credible, although some authors (e.g., Anderson & Matessa 1997) trusted them well and quoted them extensively. However, evidence has continued to accumulate indicating that, contrary to Baddeley et al.'s (1975) prediction, the ratio of verbal recall to articulation rate is not a constant (e.g., Henry 1994; Hulme et al. 1991; Morra 2000). The few available cross-linguistic estimates of the articulatory loop suggest that its capacity is larger in Chinese than in English (Cheung & Kemper 1993), and in turn, larger in English than in Italian (Morra et al. 1993) – this seems rather paradoxical, for a supposedly universal component of the mind's architecture.

Explanatory alternatives to the articulatory loop have been suggested for the word-length effect, such as output interference, proactive interference, and complexity of speech programming (see Brown & Hulme 1995; Caplan Rochon & Waters 1992; Cowan et al. 1992; Henry 1991; Nairne et al. 1997; Service 1998). In addition, various studies have found that verbal STM span is affected by variables that have little effect on articulation rate, such as word familiarity, frequency, grammar class, semantic variables, and order of stimulus words. Hence, the articulatory loop model has either been dismissed, or transformed (e.g., by Burgess & Hitch 1997) into something radically different from the original time-limited store. Furthermore, estimates of loop capacity are affected by use of span procedure versus supra-span lists (Morra 1990; Mona & Stoffet 1991; Nicolson & Fawcett 1991). Thus, the only supposedly precise estimate of a supplementary storage mechanism may have been an artifact, obtained fortuitously from supra-span lists of English words.

At this point, I think, we can abandon the idea of specialized short-term stores, and retain instead the view of working memory as the activated part of LTM. Apparently separate STM modules can be regarded as epi-phenomena of LTM modularity. Instead of searching for different supplementary stores, we can think of different sources of activation; that is, working memory must be seen as broader than the capacity-limited focus of attention, because LTM units may also be activated by other sources. These may include current perceptual input, associative learning, top-down processes from higher-order cognitive structures (e.g., Case 1974; Pascual-Leone 1987), and of course, residual activation of items that have recently been activated by the capacity-limited attentional mechanism.

If we construe in this manner working memory and its relation to the capacity-limited focal attention, then we can ask further questions. One is whether this view accounts for short-term memory phenomena traditionally explained in terms of storage. It seems so; for instance, Morra (2000) has presented a neo-Piagetian model of verbal STM that does not include any time-limited specialized store.

A second question concerns the content of working memory. Cowan implicitly suggests that it only includes declarative knowledge, as one can infer from his mention of parietal lobes. However, LTM includes procedural as well as declarative knowledge, and one may assume that procedural knowledge also needs to be activated by attentional resources. In a neuropsychological perspective, Moscovitch and Umiltà (1990) conceived working memory as "whatever processes are currently active" (without distinguishing between anterior or posterior parts of the brain) and suggested that its limits are set by the resources necessary for mantaining information and operating on it. A corollary is that, if also procedural information is considered (as some neo-Piagetian theoreticians suggested; e.g., Pascual-Leone & Johnson 1991), then the estimated size of focal attention may be more than 4 chunks.

A third question concerns individual and developmental differences and their measurement. An answer to measurement problems partly depends also on the assumptions one makes regarding whether only declarative or also procedural information is counted. These questions are central to neo-Piagetian theories. Different positions have been expressed in the debate on capacity measurement, for instance on whether an average person at the highest point of cognitive development has a capacity of 4 or 7 units (Case 1985; 1995; Ha1ford 1993; Morra 1994; Morra et al., in press; Pascua1-Leone 1970; Pascual-Leone & Baillargeon 1994). Clearly, the last word in this debate has not yet been spoken. Some results of my own research (Morra et al. 1988; 1991) suggest, however, that early adolescents have a capacity of 5 or 6

units. This sort of developmental results may suggest that an average adult's capacity possibly spans over 6 or 7 chunks of information.

Partial matching theory and the memory span

David J. Murray

Department of Psychology, Queen's University, Kingston, Ontario K7L 3N6, Canada. **murrayd@psyc.queensu.ca**

Abstract: Partial matching theory, which maintains that some memory representations of target items in immediate memory are overwritten by others, can predict both a "theoretical" and an "actual" maximum memory span provided no chunking takes place during presentation. The latter is around 4 ± 2 items, the exact number being determined by the degree of similarity between the memory representations of two immediately successive target items.

Cowan's wonderful target article suggests that there is a "pure" limit of about four items in immediate memory. He notes that we do not, as yet, have an explanation for this limit. In this commentary, I shall suggest that an explanation can be derived from partial matching theory as described by Murray et al. (1998).

This theory was developed with the intention of predicting hit and false alarm rates in immediate probed recognition tasks of the type investigated by Wickelgren and Norman (1966) and Sternberg (1966), among others. In this task, a sequence of L target items is presented, followed by a probe item that is either old or new with respect to the target list. The participant's task is to judge whether the probe is indeed "old" or "new."

In partial matching theory, m is defined as the probability that two adjacent target items share a predetermined feature of importance in memory encoding. Following Neath and Nairne (1995), it was assumed by Murray et al. (1998) that the presence of this common feature would entail that the second target would overwrite the memory representation of the first target. Murray et al. defined x to be the number of memory representations of targets available at the time of onset of the probe and predicted that:

$$x = \Sigma(1 - m)^i \text{ for } i = 0 \text{ to } i = (L - 1) \quad (1)$$

Equation 1 can be written alliteratively as

$$x = (1/m)\,[1 - (1 - m)^L] \quad (2)$$

As L tends to infinity, it can be seen from Equation 2 that x tends to $(1/m)$. According to partial matching theory therefore, the maximum possible number of memory representations available, at the time of the probe, of a very long list would tend towards $(1/m)$.

Experimental evidence that the theory can predict hit and false alarm rates in immediate probed recognition tasks was provided by Murray et al. (1989; 1999). Preliminary evidence that the theory might also predict recognition latencies in this task was provided by Boudewijnse et al. (1999) in the course of their exposition of Herbart's (1824/1890) theory of how mathematics can be applied to the prediction of how Vorstellungen (ideas) enter and leave consciousness.

However, the terms $(1/m)$ and x can also be considered to be measures of accuracy in immediate memory tasks generally. The term $(1/m)$ could represent a "theoretical maximum memory span," given a memory system subject to overwriting. If L itself were set to be $(1/m)$ the corresponding value of x would represent an "actual maximum memory span," namely, the number of target items out of $(1/m)$ target items that had left memory representations that had not been overwritten, and were therefore, still accessible to consciousness when an old probe appeared. This particular value of x, derived by letting $L = (1/m)$ in Equation 2, will be labeled A, standing for "actual maximum memory span."

An example will illustrate these numerical values. If the target material consisted of digit trigrams (for example, 2 1 8 or 6 3 9), in which each individual digit has been drawn (randomly, with replacement) from the population of the ten digits 0 to 9, then the probability, m, that a target trigram will be immediately followed by a target trigram bearing the same first digit will be .10 (one-tenth). Murray et al. (1999) argued that the memory encoding of digit trigrams by participants in this task was indeed often in terms of the first digits of those trigrams, especially if presentation were purely visual with no auditory components. According to Equation 2, even if a list were extremely long, the theoretical maximum number of (non-overwritten) memory representations of those targets at the time of the probe would be approximately $(1/m)$, that is, approximately ten. But the actual number of memory representations available to the participant at this time, as determined by constraints on conscious experience that are not yet understood, would be obtained by setting $L = (1/m) = 10$ in Equation 2, yielding an x-value of 6.5132. That is, the value of A, the actual maximum memory span, would be 6.5132.

Figure 1 shows predicted values of the theoretical maximum memory span $(1/m)$ and the corresponding actual maximum memory span (A) for values of m ranging from .1 to .5. The range of in-values from $m = .1$ to $m = .3$ has been boxed off to show that, in this range, the predicted A-values lie between approximately 2 and 6, that is, they lie in the range 4 ± 2.

For single randomly selected digits, it was shown above that $A = 6.5132$. But other measures of memory span for single digits usually provide estimates of the memory span for digits that exceed 6.5132; for example, Cavanagh (1972), on the basis of a meta-analysis of previous reports, gave the traditional memory span for digits as 7.7. But, as Cowan has documented in impressive detail, most participants, upon hearing or seeing a sequential list of digits, will bring to bear, on the process of the memorizing of that list, various techniques of chunking and associating that will expand the number correctly recalled in order, after one presentation, from 4 ± 2 to 7 ± 2 or even more.

Cavanagh's estimate that the traditional memory span for digits is 7.7 is therefore almost certainly based on data that were not free from having been grouped. The actual maximum memory span for

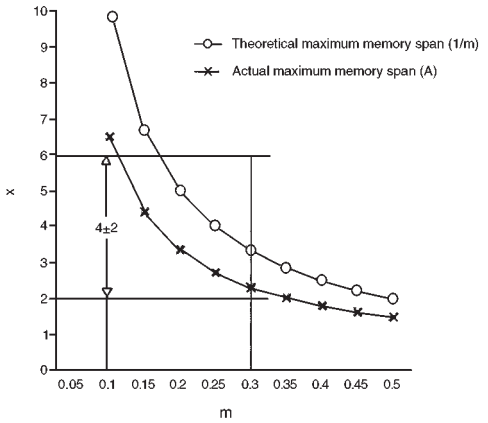


Figure 1 (Murray). The theoretical and actual maximum memory spans as predicted by partial matching theory. The variable-name x on the ordinate denotes the number of non-overwritten memory representations at the time of the probe. The variable-name m on the abscissa denotes the probability that the memory representation of a target item T shares a common feature (predetermined by the experimenter) with the memory representation of the target item immediately following. The value of the variable named A is given by $(1/m)[1 - (1 - m)^{(1-m)}]$, as explained in the text.

digits of 6.5132 estimated by partial matching theory is a measure that presumes a minimum of chunking as the participant hears or views the target list under investigation. In practice, participants usually do report more than *A* target items because the participants have succeeded in grouping or chunking the target items at the time of their presentation. To the references on grouping and chunking given in Cowan's target article, we can add contributions by Slak (1970), Thompson et al. (1993), and Murray (1995, pp. 92–105).

ACKNOWLEDGMENT

Preparation of this manuscript was supported by Social Science and Humanities Research Council of Canada grant 410–99–0831. The author is most grateful to Andrea Kilgour and Louise Wasylkiw for research assistance, and to R. Kalin, Chair of the Department of Psychology at Queen's University at the time this research was initiated, for providing the necessary computer facilities.

The nature of forgetting from short-term memory

Paul Muter

Department of Psychology, University of Toronto, Toronto, Ontario M5S 3G3, Canada. **muter@psych.utoronto.ca**
www.psych.utoronto.ca/~muter

Abstract: Memory and forgetting are inextricably intertwined. Any account of short-term memory (STM) should address the following question: If three, four, or five chunks are being held in STM, what happens after attention is diverted?

Assuming that the central thesis of Cowan's target article is correct, the question remains: Given that three, four, or five chunks are in STM, are negligibly registered in long-term episodic memory, and have been erased from sensory memory, what will be the nature of the forgetting when attention is diverted? In the third paragraph of his paper, Cowan raises the issue of the nature of forgetting from STM, and somewhat peremptorily dismisses it as "nearly intractable" (para. 3), and beset with difficulties, such as the "apparent unresolvability of the decay issue" (para. 3). Is the question of the nature of forgetting from STM any more intractable than the question of the capacity of STM?

Of course, some issues regarding the nature of forgetting from STM are covered, both explicitly and implicitly, in Cowan's paper, but the above question is largely ignored, and is a remarkable lacuna in the discussion. Memory and forgetting are always inextricably intertwined.

Over the decades there have been hundreds of attempts to answer approximations to the above question. Many of these attempts have been concerned with the rate of forgetting from STM. A study by Peterson and Peterson (1959) was quite typical: As Cowan mentions in passing, in a serial recall task Peterson and Peterson found severe forgetting of three letters after 18 seconds of distracting activity. This study is often cited as indicating the "duration" of short-term memory (e.g., Solso 1995). Muter (1980), however, argued that in the Peterson and Peterson experiments and experiments like them, participants were relying on more than STM, because they knew that they were going to be asked to recall the to-be-remembered items after an interval filled with distracting activity. Theory (e.g., Craik & Lockhart 1972) and data (e.g., Jacoby & Bartz 1972; Watkins & Watkins 1974) suggest that if participants know they are going to be tested after a retention interval filled with distracting activity, secondary memory traces are likely to be formed. When subjects expect to be tested after a filled retention interval rarely or never, there is evidence that severe forgetting occurs after approximately 2 seconds, (Marsh et al. 1997; Muter 1980; Sebrechts et al. 1989), though this finding remains controversial (Cunningham et al. 1993; Healy & Cunningham 1995; Muter 1995).

If the experiments of Peterson and Peterson and others like them did indeed tap more than STM, then many questions remain unanswered regarding the nature of forgetting from STM, and are not covered in Cowan's target article. What is the typical rate of forgetting? (This will undoubtedly depend on various circumstances, just as the capacity does, but it may tend to be at a certain level, just as capacity tends to be a certain chunk-size.) What is the shape of the forgetting curve, and how does it compare to the shape in long-term memory (Rubin & Wenzel 1996)? Does the forgetting curve depend on the nature of the information remembered (e.g., Murdock & Hockley 1989)? What are the roles of decay, displacement, and interference (Laming 1992)? What is the role of inter-chunk similarity (Posner & Konick 1966)? How important are the expectations and needs of the rememberer (Anderson et al. 1997)? What is the effect of the nature of the distracting activity (e.g., verbal vs. nonverbal, level of difficulty)? Can forgetting from STM be instant, if an extremely salient multimodal event occurs? Almost all of the research on such questions has been performed under conditions in which secondary memory contamination was likely, because the participants expected to be tested after a filled retention interval.

Cowan's paper usefully elucidates many issues regarding STM. However, a comprehensive account of STM should surely include treatment of the nature of forgetting after attention has been diverted.

ACKNOWLEDGMENTS

I thank Bennett B. Murdock, Jr. and Jay W. Pratt for helpful comments.

Long-term memory span

James S. Nairne and Ian Neath

Department of Psychology, Purdue University, West Lafayette, IN 47907-1364. **{nairne; neath}@psych.purdue.edu**
www.psych.purdue.edu/{~nairne; ~neath}

Abstract: Cowan assumes that chunk-based capacity limits are synonymous with the essence of a "specialized STM mechanism." In a single experiment, we measured the capacity, or span, of long-term memory and found that it, too, corresponds roughly to the magical number 4. The results imply that a chunk-based capacity limit is not a signature characteristic of remembering over the short-term.

Long-term memory span. As advocates of unitary approaches to memory, we applaud Cowan's efforts to identify general mnemonic principles. His heroic review of the literature has produced what appear to be remarkably consistent short-term memory capacity estimates. Although it would easy to quibble with the selective nature of his review – for example, Tehan and Humphreys (1996) report data counter to the claim that one can observe proactive interference only if there are more than four items in a list – we have chosen to focus our limited attention here on more general issues.

Cowan defines memory storage capacity operationally as the maximum number of chunks that can be recalled in a given situation. Given this definition, a few concerns arise. First, almost all the cited studies measure the number of items that can be retrieved rather than "the number of items that can be stored" (sect. 1. 1). As a result, bottlenecks in the retrieval process could well lead to an underestimation of true storage capacity. Second, in virtually every case Cowan examines, the tasks require some form of order or relational processing. For example, in the prototypical case of memory span it is necessary to remember both the presented items as well as their ordinal positions in the list. It is unclear, as a consequence, whether the capacity limits apply to item information, order information, or to some combination of both.

Third, and most relevant to the remainder of our discussion, Cowan assumes that chunk-based capacity limits are synonymous with the essence of a "specialized STM mechanism" (sect. 1.1). Although it is certainly possible that STM has such a limit, how can we be certain that the same sort of capacity limits are not characteristic of other memory systems? Does the magical number 4 apply only to the specialized STM mechanism, or might it apply as well to other memory systems, such as long-term memory? To the majority of memory theorists, Cowan included, the capacity of long-term memory is assumed to be essentially unlimited. However, to our knowledge long-term memory span has never been measured precisely, at least using the procedures and inclusion criteria adopted by Cowan. What then is the storage capacity of long-term memory?

In the typical memory span experiment, capacity is estimated by requiring subjects to recall lists of various lengths and then pinpointing the list length that produces correct performance on 50% of the trials. In principle, there is no reason why this procedure cannot be adapted to a long-term memory environment. So, we presented lists of various lengths to subjects and tested their ability to remember the lists 5 minutes later. To obtain a valid estimate of storage capacity, Cowan argues, chunk size needs to be controlled. Mnemonic strategies, such as rehearsal, can lead to hidden "higher-order" chunking that produces an overinflated estimate of capacity. To prevent such strategies, we used unrelated word lists under incidental learning conditions.

Method.

Subjects. Two hundred and five Purdue University undergraduates volunteered to participate in exchange for credit in introductory psychology courses.

Materials and design. The stimuli were 22 concrete nouns from Paivio et al. (1969) with approximately equal length and ratings of concreteness, imageability, and frequency. The subjects were assigned to one of two groups: Group 1 (N = 105) received list lengths of 2, 4, 7, and 9, whereas Group 2 (N = 100) received list lengths of 3, 5, 6, and 8.

Procedure. Subjects were tested in small groups and were informed that we were interested in obtaining pleasantness ratings for several lists of items. Each word was pronounced out loud by the experimenter at a rate of 1 word every 3 seconds, and each list was separated by 5 seconds at the end of which the experimenter said "Next list." Order of list length was partially counterbalanced and the order of the words was random and different for each group. After rating the words, a geometric filler task was performed for 5 minutes. Following this, subjects were given a sheet with all the words in their appropriate lists, with the words in alphabetical order. The subjects were asked to write down a number below each word to indicate its original presentation order. Subjects were free to work on any list at any time and were allowed much time as necessary to complete the task.

Results and discussion.

The relevant data are shown in Figure 1. Each data point represents average recall performance, collapsed across serial position, for a given list length. Note that the performance function looks very similar to what you would find in a typical short-term span experiment. Performance declines with list length in a nearly linear fashion – in fact, a straight line fit of the data accounts for over 95% of the variance. To calculate long-term memory span, we simply estimated the list length that produces 50% correct performance – in this case, the value is 5.15. It is also possible to calculate long-term span when correct recall of the entire list is required – under these scoring conditions, long-term memory span drops to 3.75. Both of these estimates are within the range of values typically seen in short-term memory span tasks (Schweickert & Boruff 1986).

What do these data mean? The major implication is that a "chunk-based capacity limit" – specifically, the magical number 4 – is not a signature characteristic of remembering over the short-term; as a result, the capacity limitations identified by Cowan say little, if anything, about whether there is a specialized short-term memory system. Instead, span limitations (the magical number 4) might simply be a characteristic of information load – there is only so much information that we can order correctly regardless of the time scales involved.

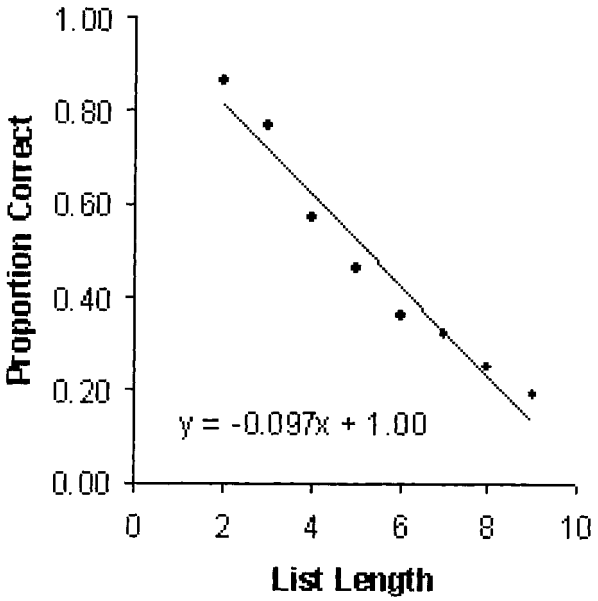


Figure 1 (Nairne & Neath). Long-term memory span: the proportion of words correctly recalled for long-term memory as a function of list length.

Where the magic breaks down: Boundaries and the "focus-of-attention" in schizophrenia

Robert D. Oades and Boutheina Jemel

Biopsychology Group, University Clinic of Child and Adolescent Psychiatry, 45147 Essen, Germany. **oades@uni-essen-de; bjemel@excite.com**
www.biopsychology.uni-essen.de

Abstract: The boundaries, the influences on, and consequences of a short-term memory (STM) capacity of 4 leads us to consider global versus local processing. We argue that in schizophrenia cognitive problems can lie partly in pre-conscious automatic selective attention and partly with the speed of processing in later controlled processes (including compound STM). The influence of automatic attentional mechanisms may be underestimated in normal psychology and explain the loss of the magic 4 in schizophrenia.

Cowan's arguments for a memory store with a capacity limited to 4, or thereabouts are persuasive: they are most eloquent in tests of serial recall, using articulatory suppression and involving articulated responses. It is a sort of explicit, short-term memory (STM). To call it the focus-of-attention does an excellent service by emphasizing that the store (and by implication attention) are terms and concepts applicable not only for information of exogenous origin but also for information with endogenous sources (recall and inter-cortical monitoring activities). However, the need to define in and define out certain methodological (e.g., use of 0.5 sec presentation times in serial recall, sect. 3.1.1, para. 7) and conceptual considerations (e.g., what makes up a chunk? Cf. Shyns et al. 1998) points to some limits and consequences (see question 4) concerning the imprint of the arrow of information flow on the arrow of time.

The limit of 0.5 sec in tests of serial recall is remarkably convenient and reminiscent of Libet et al.'s (1964) report that a stimulus should last 500 msec to enter consciousness. Yet, Libet et al. also remind us that the information is there to be used when it

reaches the cortex after less than 100 msec. Although processing at this early stage could literally be regarded as implicit, Cowan suggests that "implicit memory" stores are also limited to about 4 items. So the sensory buffer or icon remains conceptually intact. This has been clear, at least since Sperling's (1960) account that cues at short intervals (<1 sec) can elicit correct recall of *any* of 4 rows of 4 letters. The buffer may contain up to 16, or even more items. Creation of this buffer also involves selective attention as it will occur only if the subject is not concentrating on the experimenter's tie or distracted by some fluff on the floor.

The separate influences of these 16+ and 4-item stores is a matter of daily experience. Faces (250 msec) can be recognised, holistically, by an automatic process up to 300 msec later. Only then can they be decomposed into the elements of eyes, mouth, and so on (ca. 4), by a controlled process (George et al. 2000). Global processing occurs faster than that for local information, with a peak of excitation in the range of 200+ msec, led by temporo-parietal areas on the right (Sugase et al. 1999; Yamaguchi et al. 2000). The efficiency of the early (automatic) selective process, in part, determines the performance of the "focus-of-attention."

But we need to look closely at the influences of efficiency and the speed of processing. Increases of the speed of processing improve the accuracy of recall and the number of items recalled. This applies, among controlled conscious processes, to the more superficial encoding of items in pure STM (as in the digit span) as well as deeper processing in compound STM (as in word lists). This holds also for patients with schizophrenia (Brebion et al. 2000), for whom there is abundant evidence of impaired STM, controlled processing and slowed information processing (Straube & Oades 1992). However, it should not be overlooked that controlled selective attention mechanisms are also required, at the least to inhibit interfering associations (e.g., Stroop colour-word interference). It is of interest, here, to note that Brebion and colleagues reported that Stroop indicators of selective attention would predict superficial but not the deeper (compound STM) performance. Yet, Stroop interference performance is *not* disproportionately impaired in schizophrenia (except for those with predominantly disorganised symptoms), although compound (and pure) STM measures *are* reduced in size. This implies that one problem for patients with schizophrenia lies in making associations beyond the temporal and strategic bounds of pure STM and incurs the speed of processing.

But having emphasized the pure STM store and a selective attention mechanism at the level of controlled processing, we come to the sensory buffer and automatic processing of the through-put to pure STM. While latencies of later event-related potentials (ERPs) such as the P300 (around 400 msec) are usually delayed, the latencies of P50 (marking the thalamo-cortical arrival of information), N1 (excitatory cortical registration of sensory information) and mismatch negativity (sensory memory for deviance) are not consistently different among patients with schizophrenia and healthy subjects (e.g., Bender et al. 1999; Gades et al. 1996). It appears that the speed of automatic (e.g., MMN) as opposed to controlled processing (e.g., P300) is not delayed, even though the content marked by the amplitude of the ERPs recorded is often reduced in both cases.

Thus it would seem that at short latencies automatic selective mechanisms may make a larger contribution to impaired processing (exaggerated by the state of attention, Oades et al. 1997) than speeds of processing, but the opposite holds for controlled strategic processing. This idea is supported by the finding that the magnitude of sensory gating, a selective process, is impaired in schizophrenia around 100 msec post-stimulus (±50 msec: Bender et al. 1999) and may contribute to the frequently reported phenomenon of sensory overload. In contrast, local speeds of processing in different regions (frontal and temporal lobe latencies) may account for apparent deficits in the later "negative difference" marker of controlled selective attention (Oades et al. 996).

Cowan suggests that the apparent pure-STM capacity can be increased by forming inter-chunk relations, perhaps by automatic processes (e.g., priming). This suggests to us the prediction that patients with schizophrenia, renown for their "loose associations" (see Spitzer 1997) should average a larger capacity than normal: Why is there ample evidence that this is not true (e.g., Brebion et al. 2000)? We return to global percepts. These activate the right temporo-parietal junction, as emphasised by Frith and Dolan (1997) in their imaging study, especially during sustained attention with few switches of attention. Granholm et al. (1999) found not only that processing of local stimuli was impaired in patients with schizophrenia but (perhaps because of this) under conditions of divided attention (requiring switches of attention between global and local conditions in the search for a target) were actually at an advantage compared to their controls when global processing was required. Switching requires inhibition of the alternative, selective attention the inhibition of the irrelevant: its under-use resulted in the identification of fewer local items. It is this automatic mechanism of selective attention and its impoverished use that restricts the patients' STM capacity, not just to "normal" but below normal levels.

We conclude that the automatic/pre-conscious application of selective attention is not only a source of schizophrenic cognitive problems, but has a determining influence on the normal appearance of 4 in Cowan's pure STM, that he has perhaps underestimated in his worthy review.

If the magical number is 4, how does one account for operations within working memory?

Juan Pascual-Leone

York University, Toronto, Ontario M3J 1P3, Canada. **juanpl@yorku.ca**
www.yorku.ca/dept/psych/people/faculty/pasleone.htm

Abstract: Cowan fails to obtain a magical number of 7 because his analysis is faulty. This is revealed by an alternative analysis of Cowan's own tasks. The analysis assumes a number 7 for adults, and neoPiagetian mental-capacity values for children. Data patterns and proportions of success (reported in Cowan's Figs. 2 and 3) are thus quantitatively explained in detail for the first time.

Hausman and Wilson (1967) criticized Nelson Goodman (1951) for his formal-logic model of humans' construction of experience. He did so because Goodman represented objects ("entities") as collections ("sums") of predicates ("qualia") that mysteriously combined into objects of phenomenal experience without the need of a combinator or nexus, which integrates this collection into a single object. This omission, characteristic of empiricist theoreticians, is harder to imagine today; Computer science and neuroscience have recognized the need for a nexus that dynamically combines simple or compound predicates to instantiate objects. This nexus problem is currently called "the binding problem" (Kolb & Wishaw 1996; Robin & Holyoak 1995). Cowan acknowledges the binding problem, but does not seem to draw its epistemological implications for task analysis. Indeed, as many do, he estimates attentional capacity (working memory) by counting distinct schemes that mental-attentional mechanisms must hyperactivate to cause their coordination. He fails, however, to count the binding nexus: operative schemes (procedures, combinators) and parameters needed to produce integration or performance (cf. Robin & Holyoak 1995). Because of this omission, Cowan, like Halford (1993; Halford et al. 1998; Pascual-Leone 1998) and others, estimates active working memory (attentional capacity) as 3 to 5 chunks, rejecting Miller's insightful "number 7."

I have analyzed all paradigms that Cowan shows, and results are consistent with the number 7 when nexus is counted. I now illustrate these analyses using one paradigm. More refined modeling, with theoretical probability calculations, can yield more exact quantitative predictions confirmed in studies (Burtis 1982; Morra

2000; Pascual-Leone 1970; Pascual-Leone & Baillargeon 1994; Pascual-Leone et al., in preparation). There are also developmental capacity data of constructivist neoPiagetians, consistent with Miller's number when extrapolated to adults (Case 1998; Case & Okamoto 1996; Johnson et al. 1989; Morra 1994; 2000; Pascual-Leone 1970; 1995; Pascual-Leone & Baillargeon 1994; Stewart & Pascual-Leone 1992).

Consider now Cowan's experiment reported in section 3.1.3 of the target article. In this experiment, there are two ongoing subtasks: a visual-rhyme task and an auditory-digit task. One is a computer game where the picture at the center of the screen tells which of four surrounding pictures the subject must click; the one whose name rhymes with the name of the central picture. New items of this subtask are presented while participants hear (but must ignore) lists of digits presented through headphones. Occasionally the screen shifts to the digit's game, and asks for digit recall in the appropriate order. Call *pict.cent a subject's figurative scheme for the picture at the center, and *pict.surr:x (where x = 1, . . . ,4) the surround picture currently attended to. Call CLICKMATCH:RHYME the procedure for clicking the matching picture; it has a name-recalling subprocedure (operative parameter) that I call NAME, which may be chunked with the first one. Finally, the concurrent auditory subtask consists in hearing, without intentionally attending, the digit series. Because hearing and vision do not interfere with each other, auditory schemes, released by an automatic orienting reaction (OR), produce concurrent implicit hearing. The eliciting situation is facilitating, because no interfering auditory scheme is activated by the situation (Pascual-Leone 1987; 1995; in press; Pascual-Leone & Baillargeon 1994). Call *digit1, *digit2 . . . *digits, etc., figurative schemes (symbolized by a prefixed *) that encode digits being presented; and call OR:AUDIT the innately-automatic orienting reaction. With this terminology we symbolize; in a mental strategy formula the explicit visual-rhyme part as follows:

[CLICKMATCH :RHYME(NAME(*pict.cent, *pict. surr:x)) (1)
& OR:AUDIT(*digit1, *digit2, *digit3, *digit4, *digit5 . . .)]

The first segment has a mental demand of 3 or 4 symbolic schemes processed simultaneously (it can be 3 when CLICKMATCH: RHYME and NAME are chunked). The second concurrent segment is automatic; but on the assumption that adults have a processing capacity of 7 (Pascual-Leone 1970; 1987; 1998; Pascual-Leone & Baillargeon 1994), we infer that they also allocate attention to 3 digits in this implicit subtask (OR:AUDIT does not need attentional boosting). Thus, when the explicit auditory-digit part arrives, most adults should remember at least 3 digits. Indeed, Figure 3 of Cowan's article shows that 80% of his adult subjects do so. But before modeling this *explicit auditory-digit* part, I will state three principles/postulates of the neoPiagetian constructivist theory (Pascual-Leone 1987; in press; Pascual-Leone & Baillargeon 1994). Cowan's theoretical framework is fully compatible with them (see his whole sect. 2). I list them in 3 points:

(1) When mental attention (M-capacity) is allocated to schemes in the subject's repertoire (making them part of working memory), other activated schemes which are not being attended to (i.e., not placed in the M-space or focal-attentional region of WM), are generally automatically inhibited (attentionally interrupted) by the mental-attentional system; this automatic attentional interruption produces the focal "beam of attention" (Crik 1994; Pascual-Leone 1987).

(2) When the subject's executive processes are sophisticated enough, task-relevant schemes may be exempted from the attentional interruption (automatic inhibition) discussed in point (1).

(3) Unless they have been so interrupted, some activated currently in an unattended state (i.e., placed outside M-space) can be retrieved and brought into focal attention. At least one or two schemes should be so recovered, and, one by one, boosted with M-capacity. Notice that recovery is not possible when schemes have decayed too much or were previously interrupted along with misleading schemes.

These three postulates help to explain data that Cowan reports in Figure 2. I discuss only span data. During recall of *unattended list of digits,* in the explicit digit part, adults (on the assumption that their real attentional capacity is 7) will be very likely to RECALL 3 digits, as explained above, and RETRIEVE one or two schemes from outside M-space. But against this happening is the fact that overtly recalling three digits that are inside M-space demands focussing attention on linguistic schemes of the corresponding numerals, and this selective allocation of M capacity might lead, with some probability (which increases with the number of digits – Morra 2000), to interruption/inhibition of digit schemes that have remained outside M-space. This is what Figure 3 of Cowan confirms: 80% of subjects recall 3 digits, 48.5% recall 4 digits, 11.4% recall 5 digits, and only 1 subject (i.e., 2.8%) recalls 6 digits. Notice that if we assume that 3 is the demand of the explicit visual-rhyme task, 4 digits could still be attended to concurrently; and the problem would then be to recall them in order when the unattended-digits task arrives. Ordinal positions of the first and last digits are easily recalled because of perceptual saliency (this is a known anchor effect); but the relative position of digit 2 and digit 3 would have to be guessed with probability of 1/2 (=50% correct), which agrees with Cowan finding 48.5% subjects recalling 4 digits. If there were 5 digits to be recalled (i.e., 4 from inside M-space plus 1 retrieved from outside M-space), then the probability of guessing positions 2, 3, and 4 is 1/6 (6 combinations of 3 positions 16 %); close enough to Cowan's 11.4%. Finally, if there are 6 digits available for recall (4 held inside M-space and 2 retrieved from outside M-space), the probability of guessing positions 2 to 5 is 1/24 (4%) – again very close to Cowan's 2.8%.

The situation is very different in the *attended-digits task.* In this task subjects' full attention is devoted to digits; and attention may not be needed to recall digit order because forward serial order (natural to language!) was considerably practiced during the span pretest task (Cowan et al. 1999).

Consequently, when adults are ready to recall, and a RECALL operative scheme is placed inside M-space to begin to voice the numerals, 6 digit schemes will also be there. And when the corresponding 6 numerals are voiced, the digit schemes that are outside M-space will be interrupted automatically by virtue of postulate (1) – unless adults' are sophisticated and postulate (2) applies, which is unlikely. Consequently, 6 digits should be recalled on average (see Cowan's Fig. 2).

The same model explains children's data, as reported in Figure 2, when we use theoretical levels of M-capacity predicted for the different age-groups (Pascual-Leone 1970; 1987; Pascual-Leone & Baillargeon 1994). According to these levels, grade-4 children (i.e., 9- and 10-year-olds) have a magical number equal to 4; and grade 1 children have a magical number of 3 if they are 7-year-olds, as they were in Cowan's study (1999), who report their mean age in months and SD). Thus grade-4 children will be able to cope easily with the visual-rhyme part of the *unattended digits task,* and in addition they may retain 1 digit inside M-space during the visual-rhyme process. Then, during the explicit digit part, children should retrieve from outside M-space two digits at least.[1] The task-analysis representation of this step is as follows:

[*digit 1, RETRIEVE(*digit2, *digit3)] (2)

Thus we expect that grade-4 children will recall, in the unattended digit task, no more than 3 digits on average. Figure 2 shows this to be the case. As for Cowan s grade-1 children, because their "magical" number is 3, they cannot keep any digit inside M-space during the visual-rhyme part. Therefore they should recall only digits they can retrieve from outside M, that is, usually 2 digits. Figure 2 confirms this expectation.

In the case of the attended digit task, however, grade-4 children can at first consciously attend to 4 digits, but one must then be placed outside M-space to make room for the operative RECALL. Then, after the three numerals have been voiced, children are still able to retrieve the momentarily dropped (but highly activated) digit, plus two other ones from outside M-space to a total of 5 –

as shown in Figure 2. In the case of Cowan's grade-1 children, however, the same argument yields an average number of 4 digits recalled, which is in agreement with data from Figure 2.

The task analysis summarized here illustrates, in a simplified manner, why operative schemes (procedures and their parameters) should be counted to estimate the demand on working memory. In doing so, experimental psychologists might find more congenial Pascual-Leone's (1970) original proposal, which models developmental emergence of Miller's number 7. This proposal offers two distinct advantages over the magical-number-4 model of Cowan and others: (1) It can explain, in a manner congruent with modern neuroscience, the developmental growth of working memory from infancy to adulthood, and then its regression in later years; (2) It can eliminate the many anomalies, in adults' working-memory performance, that magical-number-4 theoreticians customarily explain away by vaguely appealing to "chunking." Indeed, in Cowan's paper, I count 14 occasions when the author's scrupulous review acknowledges research results that suggest a magic estimate of 6 or 7 – but chunking can be used to explain this interpretation away. I discuss this problem further in a BBS commentary (Pascual-Leone 1998) that complements the present one.

ACKNOWLEDGMENTS

Research informing methods and ideas summarized in this commentary were supported by Operating Research Grants from the Social Sciences and Humanities Research Council of Canada.

I am grateful to Dr. Janice Johnson, Dr. Sergio Morra, and Antonio Pascual-Leone for insightful comments that found their way into the final manuscript.

NOTE

1. We expect that digit-schemes from outside M-space will be recalled with higher probability whenever shorter lists are presented, because interference among digit-schemes increases with their number; and with this number so does the possibility of error in encoding and retrieval operations. For this reason, and because the length of lists changes with age in Cowan's study, we expect that younger children recall more digits from outside M-space than adults.

Linguistic structure and short term memory

Emmanuel M. Pothos and Patrick Juola

[a]*School of Psychology, University of Bangor, Bangor LL57 2DG, United Kingdom;* [b]*Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, PA 15282.* **e.pothos@bangor.ac.uk juola@mathcs.duq.edu www.bangor.ac.uk/~pss41b www.mathcs.duq.edu/~juola**

Abstract: We provide additional support for Cowan's claim that short term memory (STM) involves a range of 3–5 tokens, on the basis of language correlational analyses. If language is at least partly learned, linguistic dependency structure should reflect properties of the cognitive components mediating learning; one such component is STM. In this view, the range over which statistical regularity extends in ordinary text would be suggestive of STM span. Our analyses of eight languages are consistent with STM span being about four chunks.

Introduction. We take the view that language learning involves a major component of automatically identifying regularities in sequentially presented material. While there has been an influential research tradition arguing that language learning must be guided by a substantial innate component (Chomsky 1975; Pinker 1994; see Gold 1967 or Pinker 1979 for a mathematical analysis of this problem and Wharton 1974 for important qualifications on Gold's results), considerable evidence has accumulated recently that language can be largely inferred on the basis of experience (Elman 1996; reviews of statistical models of language and, in particular, connectionist ones are given respectively in Charniak 1993, and Chater & Christiansen 1999).

If language is (at least partly) learned, then we expect language structure to reflect the properties of the cognitive learning systems likely to be involved. The STM span can be related to language learning. For example, first, it determines the linguistic material that is immediately available to memory for comprehension. Jarvella (1971) had participants listen to various passages of normal discourse, interrupting them at various points to ask how much they could remember of what they had just heard. His main finding was that there was almost perfect recall for up to seven words before the point of interruption. Second, partly motivated by Newport's (1988; 1990) observations in developmental psychology, Elman (1993) showed that artificial languages could not be learned by a simple recurrent neural network model, unless the STM of the model started "small" and was only gradually increased to adult size (to reflect abstraction of regularities of increasing complexity).

Language learning is likely to be partly mediated via STM (alternatively, language is likely to have co-evolved with STM). Thus, we would expect the language statistical structure to reflect properties of STM, and, in particular, STM size. This is not to say that we cannot process linguistic contingencies that exceed STM size, but if the cognitive system is optimized to process automatically statistical associations only within a certain range (namely STM span), we would likewise expect language structure to be consistent with this limitation.

Cowan presents a compelling line of evidence to argue for an STM span range between three and five chunks. Consistent with this observation and relevant to language learning, Cleeremans and McClelland (1991) showed that in learning regularities in sequentially presented material, participants could not identify contingencies involving stimuli separated by more than about three other stimuli.

In this work we look at the statistical correlational properties in text from eight languages. On the basis of the above finding that language structure involves contingencies between word tokens that are separated by about four other word tokens would provide additional support for Cowan's proposal in the important area of language learning and processing.

Mutual information and linguistic structure. All the analyses presented are based on the notion of mutual information (MI), a measure of relatedness between different probability distributions. "Range" is defined as the number of words between two given word tokens, x and y. For instance, a range of 1 will indicate that words x and y are separated by only 1 other word; word tokens x and y are separated by range 0 if and only if x and y are adjacent. We ask whether our expectation of obtaining word y at a particular location is affected by the knowledge that we have word x in an earlier location. A measure of this is the mutual information (MI) between $P(x)$ and $P(y)$, the probabilities of obtaining word x and word y respectively. Mutual information indicates how much the uncertainty involved in predicting y is reduced by knowledge that we have x, and is given by Sum [all x, y] $(P(x, y)/(P(x)/P(y))$. For different ranges, $P(x, y)$ is the probability of having words x and y separated by a number of words equal to the range.

Table 1 (Pothos & Juola). *SE is the standard error of the mean sample size, for each language*

Language	Mean Words	SE	Samples
Bulgarian	1,468	256	4
Czechoslovakian	27,591	860	29
Dutch	181,407	33,483	35
English	97,272	11,805	12
Estonian	19,944	15,043	2
French	166,620	202	26
Gaelic	200,239	–	1
German	129,270	96,532	8

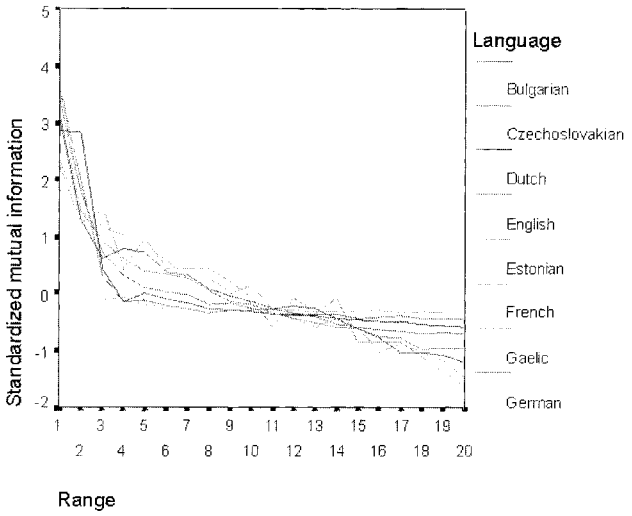


Figure 1 (Pothos & Juola). Mutual information profile when MI values were standardized for the different languages.

By MI profile, we mean the way MI varies with increasing range. This we take to be an indicator of statistical structure in language.

Analyses. We investigated samples from eight different languages, all from the CD-ROM database of the European Corpus Initiative Multilingual Corpus 1 (ECI/MCl), distributed by the Association for Computational Linguistics. In all analyses, the linguistic tokens examined are words (however, the same type of investigation can be conducted on, e.g., phonemes; see Pothos 1998). Table 1 shows the number of samples and average number of words in each language.

The MI profiles for each language were averaged and standardized; standardization enables us to compare MI values for each language regardless of sample size differences. Figure 1 shows the results of this calculation. One can see that the mutual information dependence "elbows" at about four items for all the languages. We take this observation to be compatible with the idea that linguistic dependency structure is contained primarily within a range of not more than about five word tokens, which would be consistent with the capacity of STM span being of the order of 3–5 chunks. This provides an additional source of evidence for Cowan's proposal. With additional work we aim to extend the present MI computations to take into account more directly sample size variation, and also introduce a mathematical model for the MI profiles.

A neurophysioiogical account of working memory limits: Between-item segregation and within-chunk integration

Antonino Raffone,[a] Gezinus Wolters,[b] and Jacob M. Murre[c]

[a]*Department of Psychology, University of Rome "La Sapienza," I 00185 Rome, Italy;* [b]*Department of Psychology, Leiden University, 2300 RB Leiden, The Netherlands;* [c]*Department of Psychology, University of Amsterdam, 1010 WB Amsterdam, The Netherlands.* **raffone@uniromal.it**
wolters@fsw.leidenuniv.nl **jaap@murre.org**

Abstract: We suggest a neurophysiological account of the short-term memory capacity limit based on a model of visual working memory (Raffone & Wolters, in press). Simulations have revealed a critical capacity limit of about four independent patterns. The model mechanisms may be applicable to working memory in general and they allow a reinterpretation of some of the issues discussed by Cowan.

Why the capacity limit? As Cowan points out, the reasons for a short-term memory capacity limit are not clear. Neurophysiological accounts often mention the role that could be played by neural oscillations and synchrony. For example, Lisman and Idiart's (1995) model might account for a four-item limit of short-term memory, given an appropriate frequency ratio of nested oscillations.

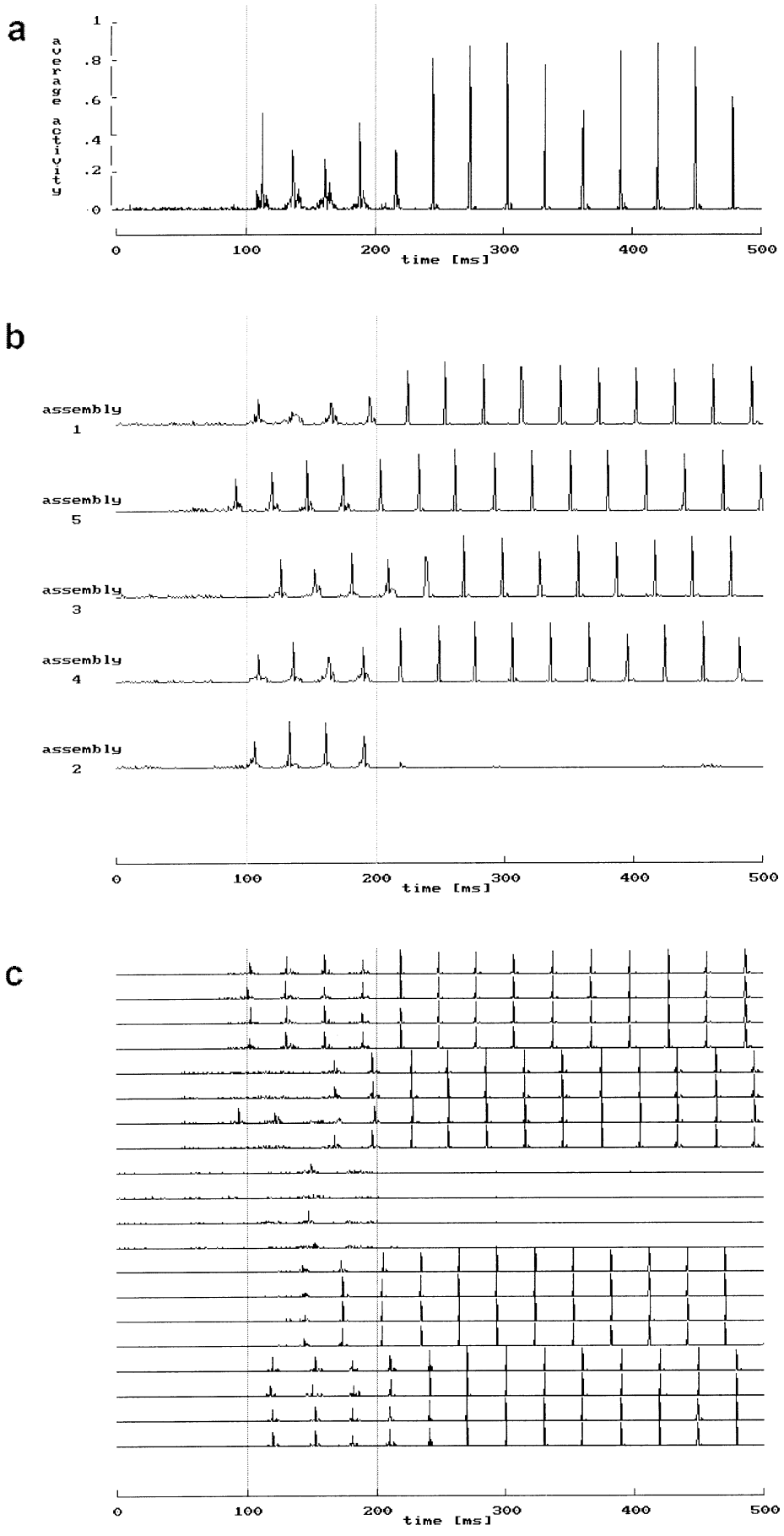
As an alternative, we present a recently developed cortical network model accounting for the limited capacity of visual working memory (Raffone & Wolters, in press). A critical capacity of about four independent patterns showed up in our simulations, consistent with the results of Luck and Vogel (1997) and with Cowan's conclusions. This capacity was proven to be independent of the number of features making up objects (Luck & Vogel 1997). The model mechanism may be applicable to any type of information in working memory, and not only to visual information. In the following we present some details of this model, and relate it to some of the functional aspects of working memory discussed by Cowan.

Retention in the cortical circuits of working memory. In the model, we assume that the cortical circuits of visual working memory are implemented in prefrontal (PF) cortex and visuo-temporal areas (e.g., the inferotemporal cortex, IT), as well as in their mutual connections. Oscillations were induced by external input and maintained after input offset through delayed feedback from PF to IT (see Fig. 1A) In Lisman and Idiart's model persistent firing is due to a slowly rising after-depolarization (ADP) state in combination with a sub-threshold oscillatory modulation in the theta range. In our view, reverberatory oscillations are more likely to occur as a result of cooperative interactions among many neurons, involving both prefrontal assemblies and assemblies located in higher-level perceptual areas.

Between-item segregation in working memory. Assuming that time-resolved neural coding plays a key role in (visual) working memory processing, the limited capacity of working memory may be explained in terms of spurious synchronizations of neuronal spike trains coding different unrelated features or items (Luck & Vogel 1997). It is unclear, however, how such spurious synchronizations of to-be-retained features could give rise to critical capacity limits, even if averaged across trials and subjects. We believe a more satisfactory explanation of the limited capacity of visual working memory is provided by a between-item segregation hypothesis. According to this hypothesis, neural assemblies in high-level visual areas, coding unrelated features or objects, exert mutual inhibitory or desynchronizing actions. The network automatically scales the phase-lag between the different reverberations in order to maintain an optimal phase segregation (see Fig. 1B). Between-item segregation depends on the inhibitory synaptic parameters. The essence of the model is that its limited capacity depends on the functional balance of a sufficiently high oscillation frequency (firing rate) and sufficient phase-segregation between disjoint assemblies. The same desynchronizing mechanism optimizing phase-segregation between assemblies coding for separate items, poses a limit to the number of oscillatory reverberations. Thus, our model suggests plausible neurocomputational reasons of the short-term memory (STM) capacity limit. As the model operates in a stochastic manner, it exhibits a psychologically plausible variability across trials of the number of objects retained, given the same number of to-be-retained objects as input.

Within-chunk integration in working memory. As is clear from Cowan's discussions, the notion of a capacity limitation is inextricably interwoven with the concept of chunking. In our view, chunking is related to neural binding processes, which may originate from bottom-up attentive operations or may be guided by long-term memory. Following the suggestion of Luck and Vogel (1997), we modeled intra-chunk integration or binding in short-term memory retention in terms of synchrony. Chunking may be based on pre-existing neural assemblies ("chunking fields"), with synchronizing connections between assemblies coding bound features, or on correlated input from earlier processing stages.

In the model, within chunk integration coexists with between-chunk segregation (see Fig. 1C). Such a property cannot be accounted for by delay activity defined in terms of firing rate attrac-


a
average activity
1
.8
.6
.4
.2
0
0
100
200
300
400
500
time [ms]
b
assembly 1
assembly 5
assembly 3
assembly 4
assembly 2
0
100
200
300
400
500
time [ms]
c
0
100
200
300
400
500
time [ms]

tors (Amit 1995). However, the network capacity for single- and multiple-feature chunks is equivalent only when there is a high "chunk stability." This functional scheme can potentially account for the different degrees of chunking effectiveness in short-term memory retention discussed by Cowan.

Central capacity versus separate capacities. In several places Cowan discusses the "central versus separate capacity" controversy. Our proposal can account for various degrees of domain specificity in processing limitations, by the assumption that competition and desynchronizing actions between neural assemblies coding for unrelated features are stronger within than between specific representational domains.

Storage versus processing in working memory. In section 4.3.6, Cowan discusses the storage versus processing capacity account, concluding that there is no reason for a separation between processing and storage in working memory. We fully agree with this view. In the cerebral cortex, a bi-directional interaction between short-term reverberations and long-term structured neural assemblies is likely to occur. In such a scheme, time-resolved synchronization (related to firing rate enhancement) and Hebbian learning may functionally interact.

The focus of attention. Cowan's proposal of a central capacity limit of the "focus of attention" leads to some neurophysiological problems. That is, how does a mechanism both collect and read out information from other sub-systems? Is it centralized or distributed in structure? How does it access and operate on the distributed representations in long-term memory? Cowan repeatedly refers to the widely shared idea of "activation" of memory representations. However, such an idea can lead to coding problems when multiple distributed neural representations are active, leading to the so-called "superposition catastrophe" (Engel et al. 1992; Von der Malsburg 1901). We suggest that a neural mechanism based on both activation (firing rate) and coherence (neural synchrony) is more plausible for the "central read-out bottleneck." We endorse Varela's notion of "resonant assemblies" in the brain (Varela 1995). In terms of our model, this would imply that only four "magical" neural assemblies can dominate the brain at any given time, saving other domain-specific active neural representations through selective synchronization.

Four-sight in hindsight: The existence of magical numbers in vision

Ronald A. Rensink

Cambridge Basic Research, Nissan Technical Center North America, Inc., Cambridge, MA 02171-1494. **rensink@cbr.com**
www.cbr.com/~rensink/

Abstract: The capacity of visual attention/STM can be determined by change-detection experiments. Detecting the presence of change leads to an estimate of 4 items, while detecting the absence of change leads to an estimate of 1 item. Thus, there are two magical numbers in vision: 4 and 1. The underlying limits, however, are not necessarily those of central STM.

In his target article, Cowan provides a wide-ranging review of data supporting the existence of a "magical number 4" – a common limit on the capacities of various perceptual and cognitive mechanisms. He suggests (sect. 3.1.1) that a similar limit may apply to *change blindness,* the finding that large changes become difficult to see when information about the location of the change is swamped by concurrent transients elsewhere in the visual field (Rensink et al. 1997; Simons 1996).

In a typical change-blindness experiment, an original and a modified image of a real-world scene are presented in succession, with a brief blank field between them; alternation continues until observer detects the change. Even though the changes are large and the observer knows they will occur, several seconds are often required before a change is seen. This has been explained by the hypothesis that focused attention is needed to see change (Rensink et al. 1997). Given that focused attention can be largely identified with visual STM (vSTM), and that vSTM has a limited capacity, only a few items can be attended at any time. Thus, the detection of change requires a time consuming attentional scan of the image.

But how many items can be attended at any one time? (Or, equivalently, how many can be held in vSTM?) This can be determined from change-detection experiments based on arrays of simple items (Rensink 2000a). The critical parameter here is *on-time* (the length of time each array is visible during a cycle). The time needed to detect a changing target item among nonchanging distractors depends linearly on the number of items in the display. For orientation change, the slope of this function (i.e., search speed) is much the same for all on-times up to 600 msec, indicating that the rate-limiting step is one of processing rather than memory. But for on-times of more than 600 msec, speed becomes proportional to alternation rate, indicating that only a limited amount of information can be held in vSTM at each alternation – more display time does not allow more items to be entered into memory. When the interstimulus interval (ISI) between displays is 120 msec, this limit is 5–6 items (Rensink 2000a). Further experiments have shown this to be a compound limit: when a short-term – presumably iconic – component is eliminated by increasing ISI to 360 msec, the estimate falls to 3–4 items (Rensink et al. 2000).

As Cowan points out, it is important to establish the absence of rehearsal or recoding processes that might cause estimates to be artificially high. For change detection, this is straightforward. First, the situation is one of information overload: not all the visible items can be placed into memory. Second, little recoding or rehearsal can occur (at least for cycle times of a second or less), since most of the available time is spent either loading items into memory or comparing them with the current input. Third, capacity is determined by a genuine discontinuity in performance, namely, a proportionality constant that appears when on-times are 600 msec or greater (Rensink 2000a). Finally, the estimate is largely unaffected by temporal decay: if ISI is greater than 360 msec, there is little further decrease, even for intervals as high as 8 sec (Schneider et al. 1999). Thus, the magical number 4 does seem to exist.

But the story does not stop here. If targets and distractors are switched so that the subject must detect a nonchanging target among changing distractors, a different limit is reached: 1.4 items (Rensink 1999; 2000b). This suggests that attended items are not independent but are instead pooled into a single collection point, or *nexus* (Rensink 2000b). Such a "magical number 1" may correspond to the limit alluded to by Cowan in his proposal that "the [4 separate] parts are associated with a common higher-level node" (sect. 2.6).

Therefore, there is considerable support for the claim of at least

---

Figure 1 (Raffone et al.). (A) Individual assembly dynamic behavior with feature input in IT and active feedback from PF. The panel shows the evolution and continuation of the average activity of one IT assembly (100 interconnected model neurons coding a single feature). Stimulus onset and offset times are marked by the vertical lines. (B) Phase segregation of IT assemblies coding for disjoint features. Four out of five reverberations remain active. Due to mutual inhibitory activity, the assemblies become spaced in the oscillatory phase, thus allowing a markedly discriminative oscillatory reverberation and retention of the coded features. Assemblies are shown in an order allowing easy inspection of phase segregation. (C) The combination of within-chunk integration and between-item segregation. Objects (chunks) consist of four interconnected assemblies. Four out of five objects are retained in terms of internally synchronized and mutually desynchronized oscillatory chunks, whereas all features coding a fifth object are suppressed.

two magical numbers in vision. However, there is less support for the claim that these are due entirely to limitations on a central working memory. To begin with, the long-term memory units (or *chunks*) accessed by STM need not be the same as the vSTM units (or *parts*) obtained from the visual input. A particular configuration might be a unit for purposes of memory retrieval, but not for visual operations such as tracking or attentional suppression. Different kinds of processes are likely to be involved, and thus, different kinds of units.

As an illustration of this, consider the detection of change in contrast sign. Whereas capacity for orientation is 3–4 items, capacity for contrast sign is at least 10 items (Rensink 2000a; 2000c). This is likely to be a compound limit, resulting from the grouping of items of similar contrast sign. But note that such groups are purely short-term visual structures – there is little likelihood that any particular arrangement had been seen before and became a chunk in long-term memory.

More generally, perception and cognition rely on systems which interact with each other to a high degree, making it difficult to determine the locus of performance limits. Indeed, there may not even be a single locus: performance on a visual task may involve both visual structures (parts) and memoric structures (chunks); a magical number might represent the number of degrees of freedom on a structure linking the two levels. Given that there are no compelling *a priori* grounds which can be appealed to, this matter will have to be settled by experiment. (The issue of individual differences would seem to be a particularly good candidate in this regard.) Until such experiments are carried out, it may be best to keep our options open as to what causes the magic in our visual world.

Which brain mechanism cannot count beyond four?

Pieter R. Roelfsema[a] and Victor A. F. Lamme[a,b]

[a]*Graduate School Neurosciences Amsterdam, Department of Visual System Analysis, Academic Medical Center (UvA), 1100AA Amsterdam, The Netherlands;* [b]*The Netherlands Ophthalmic Research Institute, Amsterdam, The Netherlands.* **p.roelfsema@ioi.knaw.nl v.lamme@amc.uva.nl www.ioi.knaw.nl/vsa.iwo**

Abstract: Cowan makes an intriguing case for a fundamental limit in the number of chunks that can be stored in short term memory (STM). Chunks are collections of concepts that have strong associations to one another and much weaker associations to other chunks. A translation of this definition for the visual domain would be that a visual chunk is a collection of features that belong to the same perceptual group (see also Mahoney & Ullman 1988). Here, we will first address the neuronal mechanisms that may demarcate visual chunks. Then we critically evaluate to what extent these mechanisms might be responsible for the limit on the number of chunks that can be held in STM. We conclude that the clarity with which the psychophysical data point to the number four is not matched by a similarly clear limit imposed by physiological mechanisms.

It is important to distinguish between two types of grouping: base grouping and incremental grouping (Roelfsema et al. 2000). Base groupings are formed rapidly and automatically after the appearance of a novel image, because they are based on the tuning of individual neurons. Many single neurons in early (e.g., Leventhal et al. 1995) as well as in higher visual areas (e.g., Kobatake & Tanaka 1994) are tuned to multiple stimulus attributes, such as color, orientation, and motion direction. The activation of a neuron tuned to, for example, red and vertical provides a base grouping between these features. Most neurons in all of these visual areas are activated within 100 msec after stimulus presentation (e.g., Oram & Perret 1992). Thus, base groupings are rapidly available, and do not depend on elaborate processing. There is no clear limit on the number of base groupings that can be computed in parallel during stimulus presentation.

The scope of base groupings has to be limited. It is unlikely that there are cells in higher visual areas that are tuned to arbitrary feature constellations (von der Malsburg 1995). This implies that an additional type of grouping, called incremental grouping, is required if base grouping fails to do the job. This may be necessary if relationships need to be established between feature domains that are not available as base groupings. In this case separate neuronal populations encode the different feature domains, and the binding problem lurks. To avoid binding problems, assemblies of neurons that respond to the various features of the same perceptual group have to be demarcated from other neurons that respond to different groups. This can be achieved by the use of *assembly labels* that are shared among neurons belonging to the same assembly. Previous theories have proposed two possible assembly labels: synchrony and firing rate modulations (reviewed by Roelfsema & Singer 1998). According to the first possibility, neurons that belong to the same assembly fire in synchrony. The second possibility is that neurons of the same assembly are labeled with an enhanced firing rate. The distribution of either of these assembly labels across neurons takes time, and incremental groupings therefore do not form automatically (Roelfsema et al. 2000). Let us now address the limits that are imposed by these assembly labels on the number of objects that can be stored in STM.

When synchrony is used as an assembly label, multiple assemblies can coexist. Neurons within each of these assemblies may fire synchronously, but they need not be synchronized to neurons of the other assemblies. Thus, multiple incremental groupings can form during stimulus presentation and they can, in principle, also be stored in STM. The target article embraces the synchrony label and suggests that the number 4 emerges as the ratio between brain rhythms, such as the 40 Hz gamma rhythm, and the 10 Hz alpha rhythm, as was first suggested by Lisman and Idiart (1995). The underlying idea is that four assemblies that are synchronized at the higher frequency can just stay out of phase in a single cycle of the low frequency oscillation. However, this line of reasoning neglects another result of studies on cortical synchronization. Neurons that respond to different objects do not have a tendency to stay out of phase. Instead, neurons that respond to different objects usually have weaker synchronization, or fire independently (Gray et al. 1989; Livingstone et al. 1995). This invalidates arguments that are based on the ratios between brain rhythms (e.g., Lisman & Idiart 1995). Because different assemblies do not stay out of phase, spurious synchronization between cells that belong to different assemblies will occur. This may also account for the interference between multiple items that are stored in STM. However, in this case the number 4 does not immediately fall out of the theory.

The other label that can be used to demarcate neurons that respond to the various features of a single object is an enhancement of their firing rate. Such response enhancements have often been documented as the correlates of visual attention that is directed to the respective object (reviewed by Desimone & Duncan 1995; Maunsell 1995). Grouping of features by attention has first been suggested in Treisman's feature integration theory (e.g., Treisman & Gelade 1980). The enhanced firing rate label can, however, only be used by one assembly at a time. If two assemblies responding to different objects are simultaneously labeled, the binding problem reappears, because neurons in both assemblies have an enhanced firing rate. This implies that rate modulations only allow a single incremental grouping at a time (Roelfsema & Singer 1998).

What happens to base- and incremental groupings when the stimulus is removed and bottom-up activation declines? The intimate relationship between STM and attention that is discussed in the target article, is also present at the physiological level. For many neurons, the enhancement of neural responses to attended features persists during intervals in which the visual stimulus is removed (Chelazzi et al. 1993; Fuster 1997; Rainer et al. 1998). Thus, the very same neurons that carry the attentional label are also involved in STM. The largely unexplored mechanisms that are responsible for persistent firing during memory episodes may

impose a limit on the number of neurons that can maintain their activity. It is unclear, at present, whether such a limit would affect the number of base groupings or rather the number of incremental groupings that can be stored. In any case, the number 4 also does not seem to fall out of the theory here.

In summary, synchrony allows the coexistence of multiple incremental groupings, whereas rate modulations only accommodate a single incremental grouping at a time. Four is not a clear-cut prediction of either mechanism. The data of Luck and Vogel (1997) show that feature conjunctions of 4 objects are held in STM. It is, however, left open by this study whether these conjunctions go beyond base grouping, since individual neurons in early visual areas can also be tuned in multiple feature domains. Moreover, the four visual objects that are held in STM cannot have arbitrary complexity. Thus an exciting arena for future psychophysical and physiological experimentation is defined, to investigate how many *incremental* groupings can be held in visual STM.

Functional neuroimaging of short-term memory: The neural mechanisms of mental storage

Bart Rypma[a] and John D.E. Gabrieli[b]

[a]*University of California, Department of Psychology, Berkeley, CA 94720-1650;* [b]*Department of Psychology, Stanford University, Stanford, CA 19001.*
rypma@socrates.berkeley.edu **gabrieli@psych.stanford.edu**

Abstract: Cowan argues that the true short-term memory (STM) capacity limit is about 4 items. Functional neuroimaging data converge with this conclusion, indicating distinct neural activity patterns depending on whether or not memory task-demands exceed this limit. STM for verbal information within that capacity invokes focal prefrontal cortical activation that increases with memory load. STM for verbal information exceeding that capacity invokes widespread prefrontal activation in regions associated with executive and attentional processes that may mediate chunking processes to accommodate STM capacity limits.

Cowan provides a thoughtful integration of broad literature supporting the view that: (1) "4 ± 1" represents the true capacity limit of STM, and (2) supracapacity maintenance (i.e., more than 4–5 items) of information occurs as a result of additional "executive" [2] processes that expand capacity by strategic data-compression (i.e., "intelligent grouping" or "chunking" e.g., Baddeley 1986; Waugh & Norman 1965) of to-be-maintained information. This fundamental idea about the structure of the human mind is, however, controversial because, as Cowan reviews, alternative interpretations of relevant experimental results are possible.

Cognitive neuroscience can offer independent and convergent constraints upon such a theory of STM functional architecture. One functional neuroimaging study that we performed (Rypma et al. 1999) offers remarkable support for Cowan's argument.

That study examined the neural correlates of STM with functional magnetic resonance imaging (fMRI), a neuroimaging technique that allows detection of hemodynamic changes associated with regional neural activity that accompanies cognitive performance. Subjects performed a verbal STM task, based on Sternberg's (1969) item-recognition task, in which STM load was varied from 1 to 3 to 6 letters. Thus, the 1-letter and 3-letter loads fell within the putative true capacity of STM, whereas the 6-letter load exceeded that capacity. Subjects saw memory-sets of 1, 3, or 6 letters per trial, kept those letters in mind for 5 seconds, and then pressed a key, if a probe-letter was part of the memory-set. As is typically found, subjects were highly accurate and reaction times increased linearly with increasing memory load.

One fMRI analysis examined the difference between 3-letter and 1-letter loads, a difference that may be conceptualized as an increase in STM demand, but within capacity limits. In prefrontal cortex (PFC), the only observable difference was increased activity for the 3-letter load in inferior (ventrolateral) parts of the left hemisphere (around Broca's area) (Fig. 1). The second fMRI analysis examined the difference between 6-letter and 1-letter loads, a difference that not only increased STM demand, but also exceeded capacity limits, thus invoking executive processes. Indeed, there was a dramatic activation increase, not only around Broca's area, but in many additional areas, including superior (dorsolateral) regions of the PFC, in both hemispheres, and in cingulate gyrus. Altogether, there was a 75-fold increase in the spatial extent of PFC activation as subjects moved from subcapacity to supracapacity SIM performance. Thus, during supracapacity memory performance, we observed a greatly disproportionate increase in PFC activation, and activation of multiple PFC regions.

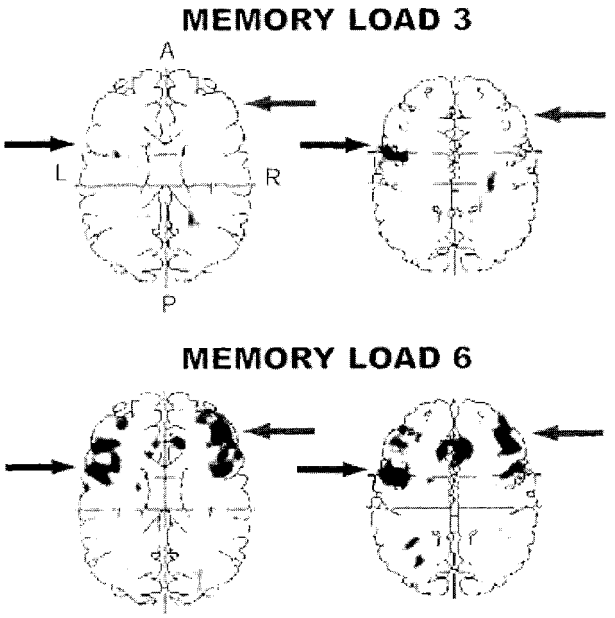


Figure 1 (Rypma & Gabrieli). Two axial (horizontal) slices showing averaged neural activity across all subjects in the 3-and 6-letter memory load conditions. A = anterior, P = posterior, L = left, R = right. Gray arrows indicate dorsolateral PFC, white arrows indicate ventrolateral PFC. FMRI results indicated ventrolateral PFC activity in the 3-letter condition and additional dorsolateral PFC activity in the 6-letter condition.

These results can be readily interpreted to support the view that distinct STM processes, mediated by separable neural systems, carry out information retention, depending on the extent of memory-demand (i.e., whether or not the memory demand exceeds a "4 ± 1" capacity limitation). In PFC, variation within that capacity (1 vs. 3 letters) resulted in focal left ventrolateral activation that increased with load. This activation may reflect a STM system that mediates subcapacity maintenance (i.e., less than 4 items) of verbal information, possibly through verbal rehearsal (e.g., Vallar & Baddeley 1984). STM loads exceeding that capacity (6 letters) invoked widespread bilateral dorsolateral PFC activation in areas not activated by subcapacity STM loads. These activations may reflect the additional "executive" involvement in memory maintenance tasks with supracapacity memory loads discussed by Cowan (e.g., Baddeley & Hitch 1974) Indeed, multiple studies suggest that dorsolateral PFC is activated whenever executive processes are required (Cohen et al. 1994; D'Esposito et al. 1995; Petrides 1996; Prabhakaran et al. 2000; Smith & Jonides 1999). Further, concomitant activation in dorsolateral PFC and cingulate gyrus have been observed in attention-demanding tasks (e.g., Corbetta et al. 1991), supporting Cowan's contention that attention is one component of the additional mechanisms that support chunking.

The behavioral results reviewed by Cowan and our fMRI results together indicate that verbal STM loads below a "4 ± 1" capacity limit invoke a specific STM system mediated in part by left inferior PFC. Supracapacity verbal STM loads invoke multiple additional executive and attentional systems mediated by bilateral dorsolateral prefrontal and cingulate cortices. STM circuitry extends beyond these areas to at least parietal and cerebellar regions, but these prefrontal findings offer independent, convergent evidence for Cowan's persuasive argument.

Characterizing chunks in visual short-term memory: Not more than one feature per dimension?

Werner X. Schneider, Heiner Deubel, and Maria-Barbara Wesenick

Institute of Psychology, Ludwig-Maximilians-University Munich, D-80802 Munich, Germany. **{wxs; deubel; wesenick}@psy.uni-muenchen.de**
www.paed.uni-muenchen.de/mip/psych/deubel/wwwdocs/index.htm

Abstract: Cowan defines a chunk as "a collection of concepts that have strong associations to one another and much weaker associations to other chunks currently in use." This definition does not impose any constraints on the nature and number of elements that can be bound into a chunk. We present an experiment to demonstrate that such limitations exist for visual short-term memory, and that their analysis may lead to important insights into properties of visual memory.

To determine the capacity limit of short-term memory (STM) can be a tricky business with a number of potential pitfalls. Cowan provides a careful and much needed analysis of these pitfalls that can confound estimates of the real memory capacity with other limitations in the system. After reviewing a large number of studies that try to avoid these problems, Cowan concludes that the limit of (STM) is about 4 chunks, rather than the classical 7 ± 2. Concerning visual working memory, this capacity estimate has indeed been put forward by several authors for quite some time. Based on experimental evidence from studies on transsaccadic memory (e.g., Irwin 1992) and on visual STM (e.g., Shibuya & Bundesen 1988), one of us reached a similar conclusion in a theoretical analysis of visual working memory (Schneider 1999).

However, knowing how many chunks can be retained is by no way sufficient if we want to know how much information can be stored in STM – to answer this question, it is absolutely necessary to know more precisely how a visual chunk can be characterized. Indeed, we think that a central and important issue of future research will be to analyse the limits of STM in terms of the mechanisms by which more basic elements are formed into chunks, and also to study how attention and specific tasks determine this chunking.

Imagine a scene consisting of four "objects," each being a square made up of a 3*3 raster of differently coloured parts. An intuitively obvious prediction would be that these four objects *cannot* be retained as well as four homogeneously coloured squares. Amazingly, the data from one of Luck and Vogel's (1997) experiments seemed to suggest otherwise. In this experiment, subjects had to retain such stimuli as shown in Figure 1 that were defined by various colour-colour-conjunctions. The results seemed to show that objects that are made of a conjunction of two colours could be retained equally well as objects with just one colour value per object. This implies that visual chunks may contain at least two feature values for the colour dimension.

For storage of visual information such an analysis is possibly somewhat easier than for verbal material, which is the focus of Cowan's article. In the visual format it is intuitively plausible to assume that chunks are direct reflections of consciously perceived visual "objects." A frequent assumption in vision is that an object can be described in terms of its basic visual dimensions, such as its colour, shape, or motion. Within each dimension, the stimulus can be specified by "features" or "feature values" (e.g., Treisman &

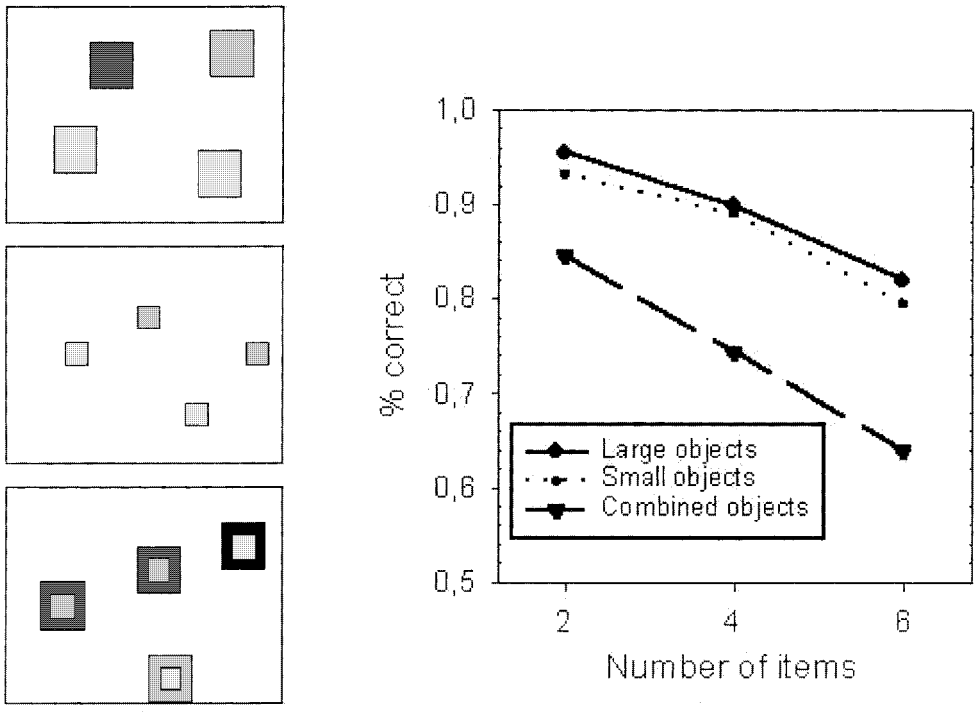


Figure 1 (Schneider et al.). Left: Stimuli used in the experiment. Subjects had to memorize configurations of 2, 4, or 6 coloured squares for 900 msec. After the retention period, either the same display or a display with one of the squares differing in colour reappeared; subjects were asked whether they had detected a change. Squares could be large (0.75*0.75 deg) or small (0.35*0.35 deg), or conjunctions of both. Right: Percent correct report as a function of the number of items presented.

Gormican 1988) For instance, the dimension "colour" of an object can be characterized by one of the features "red," "blue," "purple," and so on. After breaking down a visual object into these more basic elements, the question arises as to whether the above definitions of "dimensions" and "features" in vision constitute an appropriate basis for characterizing chunks in visual memory. In other words, one may ask (1) whether chunking in visual working indeed functions by binding features into integrated visual objects, and (2) analyse the rules and limitations of this binding.

With regard to the first question there is recent evidence from Luck and Vogel (1997, also referred to in sect. 3.1.1), suggesting that the capacity limit of visual STM indeed refers to feature bundles in the form of objects. In their experiments, subjects were required to retain simple geometrical visual objects made up of feature conjunctions such as of a certain colour, orientation, and length. The data showed that objects defined by conjunctions of two or more dimensions (e.g., a line of a certain orientation, colour, and length) can be retained as well as objects defined by only a single dimension (e.g., orientation only). For any of these combinations, the estimated memory capacity was about four objects.

The second question is strongly related to the problem of how many of these basic elements can be bound into a single chunk, and whether there exist limitations as to the possible combination of features.

In our own experiments, we attempted to replicate this surprising finding (Deubel et al., in preparation). We used the same stimuli as Luck and Vogel (1997), and identical experimental parameters such as presentation and retention times. Our experimental results (Fig. 1) clearly show that retention of objects defined by a conjunction of two colours leads to a strong drop in performance, as compared to the condition in which the objects consisted of one colour only. So, external objects with two colours seem to require two chunks for the internal coding. This finding is in obvious contrast to the result of Luck and Vogel (1997). The reason why we could not replicate their data is unclear to us, however, in an independent study, Wheeler and Treisman (submitted) recently reported a finding similar to ours.

These data are clear evidence that there exist prominent limitations to chunking in visual memory. As a possible, preliminary rule of thumb suggested by the result, one may assume that a visual chunk can consist of not more than one feature per dimension, that is, one colour, one shape primitive, and so on. A further, yet unresolved important issue in this context is the question whether there is also a limit in the number of possible dimensions that define a chunk. Luck and Vogel (1997) found no limit (i.e., no drop in memory performance) up to a conjunction of four different dimensions (colour, orientation, length, gap). However, it might be that a limitation larger than that can indeed be found.

The empirical task of the future will be to determine more precisely the limitations of chunking and how they relate to visual features and dimensions. Indeed, we think that the paradigm presented here offers a promising experimental approach to answer questions about the nature of chunks in vision. Measuring memory performance for a variety of stimuli and features could reveal the basic dimensions and features in vision in a very straightforward way: If adding the feature in question to the stimuli leaves the memory capacity (in terms of number of objects) unaffected, one may conclude that it is really a basic visual feature, forming an elementary part of a visual chunk.

Cowan defines a chunk as "a collection of concepts that have strong associations to one another and much weaker associations to other chunks currently in use." This definition does not impose any constraints on the nature and number of elements that can be bound into a chunk. Our experiment is a demonstration that such limitations exist, and that their analysis may lead to important insights into properties of visual memory.

The magical number 4 in vision

Brian J. Scholl[a] and Yaoda Xu[b]

[a]*Department of Psychology, Vision Sciences Laboratory, Harvard University, Cambridge, MA 02138;* [b]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.*

scholl@wjh.harvard.edu yaoda@psyche.mit.edu

www.wjh.harvard.edu/~scholl

Abstract: Some of the evidence for a "magical number 4" has come from the study of visual cognition, and Cowan reinterprets such evidence in terms of a single general limit on memory and attention. We evaluate this evidence, including some studies not mentioned by Cowan, and argue that limitations in visual processing are distinct from those involved in other memory phenomena.

Cowan's discussion of the "magical number 4" synthesizes evidence from domains which are rarely discussed together. In particular, Cowan draws on work from the study of visual cognition – such as studies of subitizing (sect. 3.3.2) and multiple object tracking (sect. 3.3.3) and attempts to reinterpret such evidence in terms of a general memory limitation, which he suggests is a reflection of the underlying capacity of the "attentional focus" (a thesis which is discussed in Cowan 1995, but which he does not argue for in his target article). Here we note additional evidence for a limit of approximately 4 objects in certain types of visual processing, and discuss why these limits are probably distinct from those involved in other (e.g., verbal) tasks.

Additional evidence from visual cognition. Additional evidence for a "magical number 4" in visual processing comes from studies of infants, normal adults, and neuropsychological syndromes. Recent looking-time studies with infants have suggested that they are able to keep track of arrays of objects through additions and subtractions, but only if there are less than 4 objects in these arrays (e.g., Uller et al. 1999; Wynn 1992), and this evidence has been interpreted in terms of developing mechanisms of visual attention (e.g., Carey & Xu, in press; Scholl & Leslie 1999). In normal adults, there appears to be a limit of 4 on the number of objects which can receive prioritized processing due to attentional capture (Yantis & Johnson 1990), and the number of items which can be simultaneously examined in a visual search for a change (Rensink 2000).

Finally, it has been shown that bilateral lesions of the parietal lobes in Balint's syndrome can reduce visual processing capacity. Patients with Balint's syndrome have great deficits in perceiving complex visual scenes, although their ability to recognize individual objects is usually preserved (for a review, see Rafal 1997). Dehaene and Cohen (1994) studied visual enumeration in 5 Balint's patients and found that these patients could enumerate sets of 1, 2, and sometimes 3 items correctly, but not sets comprising more than 3 items. Reaction time slopes for these patients were flat for set sizes of 1 and 2 items, but increased sharply for set sizes of 3 or more items. Treisman and colleagues (Friedman-Hill et al. 1995; Robertson et al. 1997) reported another Balint's patient who could not correctly enumerate more than one or two objects even when he was aware that more were present. In rare and extreme cases, Balint's patients report seeing only one object when presented with multiple objects (e.g., Coslett & Saffran 1991).

Specific visual limits or general memory/attention limits? Cowan views such evidence as continuous with data concerning the number of chunks which can be simultaneously active in short term memory (STM). In contrast, we think there are good reasons to resist this reinterpretation, and to view the limits on visual processing as separate from those involving verbal and other non visual material. (In this respect we take a position similar to that of Miller 1956 who suspected that STM limits and subitizing limits were independent.) Given space restrictions, we will largely restrict our discussion of this issue to the evidence which Cowan does discuss in his target article: subitizing (wherein observers can determine the cardinality of sets with less than 5 items roughly in parallel and without errors) and multiple object tracking (MOT;

wherein observers can attentionally track up to 4–5 independently and unpredictably moving identical items in a field of identical distractors).

Cowan presents only a few arguments for interpreting these phenomena in terms continuous with general STM limits. For MOT he provides no arguments, simply stating that one could use a general STM-based theory to explain performance. (Such an explanation, it seems to us, could not easily account for the strong dependency of MOT performance on subtle visual details such as the type of accretion and deletion behind occluders; Scholl & Pylyshyn 1999). For subitizing, he notes the vision-based theory of Trick and Pylyshyn (1994a), and argues against it mainly by appeal to two phenomena. First, he suggests that the "pop-out" alluded to by Trick and Pylyshyn can also occur for larger numbers of items, for example "when all of the eggs [in a carton] pop out against the surrounding carton" (sect. 3.3.2). This, however, is clearly not the type of pop-out that Trick and Pylyshyn (and others who have investigated visual search) have in mind, since the eggs in this case do not pop out as individuals, but as a group. Second, Cowan suggests that focused central attention is more important to enumeration than is suggested by Trick and Pylyshyn's theory, since other researchers (Atkinson et a1. 1976; Simon & Vaishnavi 1996) studying the enumeration of dots in afterimages have claimed that observers cannot enumerate sets greater than 4 without eye movements. This claim is false, however, and the limits these investigators found were due to the confounding effects of crowding (He et al. 1997).

Beyond Cowan's arguments, we think there are several additional reasons to view these limits as distinct from those involved in verbal STM. First, viewing them as identical seems to necessitate a prediction that one should not be able to track 4 targets in the MOT task and simultaneously acquire and hold 4 verbally-presented items in STM. However, this is trivial to do, and such tasks seem not to interfere at all. (In an informal test, two observers tracked 4 in 8 items for 10 sec with an accuracy of 87.5% averaged over 10 trials. When they also had to remember 4 random digits presented auditorily as the targets were being specified, they tracked with an accuracy of 92.5%, and made no errors on the memory task.) Cowan notes in section 4.2 of the target article that such evidence against a single capacity limit could be explained away by appeal to attentional switching back and forth between the two tasks, but in this respect MOT is an ideal foil, since one can succeed in the task only by continuous tracking (Pylyshyn & Storm 1988). Second, an explanation based on a single general limitation of memory or attention predicts that these limits should stand or fall together in neuropsychological impairments, which they do not. For example, none of the Balint's patients mentioned above exhibited deficits in short-term memory span. There are patients who, after lesions in the left hemisphere language areas, exhibited reduced STM span despite normal speech production in some cases (e.g., Baddeley 1986; Shallice & Warrington 1970). However, none of these patients showed any signs of Balint's symptoms or deficits in visual processing. Moreover, although these patients showed very poor retention of auditorily presented digits, with a span in the region of two items, they usually showed better retention of visually presented digits, with a span in the region of 4 or 5. These double dissociations in lesion sites and patient performance argue strongly against the notion that a common capacity limitation underlies capacity limited performance in both verbal and visual tasks.

Visual objects vs. chunks in memory. The view that these limitations in visual processing are distinct from those involved in other memory phenomena is further strengthened by the fact that the "units" of processing in each case are quite different. The "chunks" of memory can be almost infinitely flexible in their composition, and are thus defined by Cowan and others simply in terms of association networks (see sect. 1.3). This flexibility is in marked contrast to the units of visual attention – visual objects – which appear to be characterized by highly constrained and inflexible rules (Scholl, in press). In MOT, for instance, observers can track 4 dots in a field of 8 dots, but completely fail when trying to track 4 line endpoints in a field of 4 lines (and thus, 8 endpoints). In general, very specific rules involving connectedness and part-structure seem to determine whether a feature cluster can be tracked in MOT (Scholl et al., in press). Similarly, in visual short-term memory studies using a change detection paradigm, color and orientation features are best remembered if they belong to the same part of an object and less well remembered if they belong to different parts of an object (Xu, submitted). All of these constraints are in marked contrast to the robustness and flexibility of potential STM chunks with verbal materials.

We think the considerations discussed here provide good reasons for thinking that the limits of approximately 4 involved in various types of visual processing are distinct from other similar STM limits. We remain agnostic on the question of why there should exist similar independent limits. It could be for the teleological and computational reasons discussed by Cowan (in sect, 4.1), or it could be – as George Miller (1956) suspected of the similarity of memory capacity and subitizing limitations – "nothing more than a coincidence."

How unitary is the capacity-limited attentional focus?

Torsten Schubert and Peter A. Frensch

Department of Psychology, Humboldt-University Berlin, D-10117 Berlin, Germany. **{torsten.schubert; peter.frensch}@psychologie.hu-berlin.de**

Abstract: Cowan assumes a unitary capacity-limited attentional focus. We argue that two main problems need to be solved before this assumption can complement theoretical knowledge about human cognition. First, it needs to be clarified what exactly the nature of the elements (chunks) within the attentional focus is. Second, an elaborated process model needs to be developed and testable assumptions about the proposed capacity limitation need to be formulated.

One of the main contributions of Cowan's important target article is the assumption of a unitary limitation of the attentional focus. Cowan's arguments in favor of this assumption should reinforce the current discussion about the nature of unitary (e.g., Baddeley 1986; Norman & Shallice 1986) or distributed attentional mechanisms (Allport 1987; Meyer & Kieras 1997; Neumann 1987). Although we agree that this assumption is intriguing, we are somewhat disappointed by its theoretical elaboration and by the absence of significant support advanced in its favor by Cowan.

As supporting evidence for his assumption, Cowan lists different studies that all yield performance restrictions of about 4 items in different experimental contexts and over a wide range of stimulus materials, for example, dots, digits, screen locations, auditory signals, and so forth. Because all these studies somehow yield the number "4," Cowan's main argument is that there should be a unitary mechanism underlying this limitation. However, for this argument to be convincing, it needs to be shown, first, that the items across the different reported experimental contexts are comparable entities. Second, one would need to describe an elaborated process model with a set of mechanisms formulated that allows an integrated understanding of the findings.

Unfortunately, the present work falls somewhat short on both of these points. For example, Cowan uses the concept of chunks in stressing the equality of items in different experimental contexts. However, he does not formulate a convincing operational definition of what exactly a chunk is. How do we know that 4 dots, screen locations, or digits correspond to 4 chunks as, for example, 4 words might? The main question is: How can we measure or define chunks independently of the experimental context we are dealing with in a concrete experimental situation? If there is no sufficiently constraining definition, what prevents us from arguing

that 4 items held in visual short-term memory (STM) and 4 legs of a chair have a common causality in the limitation of subjects' attentional focus?

Furthermore, Cowan does not formulate an elaborated process model to explain the findings in different experimental contexts on the basis of a common set of mechanisms. Lacking such a model, Cowan's position is descriptive but not explanatory. Thus, one could argue that the reported results concerning STM (Sperling 1960), articulatory loop (Baddeley 1986), visual search (Fisher 1984), enumeration (Trick & Pylyshyn 1993), and so on, are best explained by different (sometimes even computational) models assuming separate attentional and memory mechanisms (Allport 1987; Neumann 1987). The alternative assumption of a unitary attentional mechanism as a common cause for the reported results must remain highly speculative until tested.

On the basis of an elaborated process model, one could formulate specific predictions for an empirical test of the common-cause hypothesis. For example, if a process or mechanism A (e.g., the attentional focus) is identified as essential for tasks 1, 2, 3, . . . , and *n*, then one should be able to find a factor X, the manipulation of which will have equal consequences on subjects' performance in tasks 1, 2, 3, . . . , and *n*. In contrast the manipulation of a process B, essential only for tasks 1 and *n*, should influence performance only in tasks 1 and *n*, and not in the remaining tasks. Converging evidence of this kind would provide most valuable to support for assumption of a unitary capacity limited attentional focus.

If these caveats can be resolved, the assumption could significantly deepen our theoretical understanding of the human cognitive system. To illustrate, consider a different research area that also deals with capacity limitations, research on dual tasks. Studies using the psychological refractory period (PRP) paradigm, in which subjects perform two choice-reaction tasks simultaneously, have yielded rather contradictory results on capacity limitations. Many studies with this paradigm have provided support for a capacity-limited central mechanism related to response selection (McCann & Johnston 1993; Pashler 1994; Schubert 1999; Welford 1980). These studies suggest that only one response selection can take place at a time, thus limiting the capacity of a central attention mechanism to 1 item (chunk) contrary to Cowan's proposal of 4 chunks. However, other contradictory findings with the PRP paradigm suggest no capacity limitation at response selection at all (Hazeltine et al. 2000; Schumacher et al. 1998). The latter results stimulated Meyer and Kieras (1997) to propose a scheduling account (EPIC) for dual-task situations in particular and sensorimotor tasks, as well as working memory tasks in general. According to EPIC, elements of the task to be performed are maintained in a working memory system without any limiting attentional capacity. Executive processes allow the scheduling of task processes according to subject's specific instructions and goals. Cowan expresses doubts that this "scheduling theory" can account for the findings on attentional and memory limitations. However, looking at the contradictory evidence and theoretical positions, the question arises whether there are different task-specific working memories with different capacity limitations, for example, one for dual tasks and one for storage tasks, or whether there exists no central capacity limitation at all.

We assume that the assumption of a capacity-limited attentional focus may account for the contradictory results in dual-task research if one localizes the capacity limitation strictly at the level of conscious information processing. A plausible explanation is that in dual-task studies a capacity limitation of the central mechanism can be observed only when subjects carry out the tasks in a relatively unlearned state. In this case, subjects probably make a conscious decision which response to map to which stimulus when performing the task. The conscious decision requires maintaining a set of different stimuli and different responses in working memory together with a set of S-R mapping rules. One could easily imagine that, in this case, the number of chunks in working memory exceeds the proposed limit of 4, thus causing a deterioration of dual-task processing.

What about studies suggesting no dual-task costs at all (e.g., Hazeltine et al. 2000; Schumacher et a1. 1998)? One important feature of such studies appears to be that the reduction of dual-task costs emerges after long, specialized training. It is thus plausible that training leads to an over learned mapping of stimuli and responses and consequently to automatic activation of the response when a stimulus is presented. In this case, no conscious decision is necessary to select the appropriate response for a special stimulus, and the proposed capacity limit on the attentional focus is not observed (see also Greenwald & Shulman 1973).

The above conjectures show that the assumption of a unitary capacity limit on conscious information processing may shed new light on different and, even, contradictory findings in other fields, that have been investigated in isolation. Given this and given an understanding of the elements of consciousness (chunks) as well as an elaborated process model, the assumption of a unitary capacity limitation could be fruitfully included in a broader theory of the human cognitive system.

Dispelling the magic: Towards memory without capacity

Niels A. Taatgen

Department of Artificial Intelligence, University of Groningen, 9712 TS Groningen, Netherlands. **niels@tcw2.ppsw.rug.nl**
www.tcw2.ppsw.rug.nl/~niels

Abstract: The limited capacity for unrelated things is a fact that needs to be explained by a general theory of memory, rather than being itself used as a means of explaining data. A pure storage capacity is therefore not the right assumption for memory research. Instead an explanation is needed of how capacity limitations arise from the interaction between the environment and the cognitive system. The ACT-R architecture, a theory without working memory but a long-term memory based on activation, may provide such an explanation.

The goal of science has always been to show that things in the world that appear to be accidental can be explained by a set of systematic and fundamental principles. Miller's (1956) magical number seven and subsequent theories based on the idea are attempts to find such principles. Cowan tells us that the magical number is not seven, but actually four. Still the word "magical" lingers around this mysterious capacity. My proposal is not to attack the "seven" aspect of the principle, but rather the "magical" part of it, since we all know that magic doesn't really exist. Whereas in Miller's original article short-term capacity was just an empirical fact, it has subsequently grown into a theory that people actually have a pure storage capacity.

Let us elaborate on this idea. If we take capacity seriously, the number of items that can be stored by an individual has to be an integer. A capacity of 3.5 only makes sense as a group average, not as a property of an individual. An individual can either retain three items or four items, not three-and-a-half. An individual capacity of 3.5 only has meaning if the individual can sometimes remember three items, and sometimes four items. But this is hard to reconcile with the idea of a fixed capacity. It becomes even harder to explain development. Even according to Cowan's own data, the capacity of adults is larger than the capacity of children. But how then does this capacity grow? Are there sudden increases in which the capacity is incremented by one?

The problem with the target article is that it already assumes there is a capacity limit, and that it can be studied separately from the rest of memory. If one wants to prove there is indeed a limit-capacity short-term store, the relation to long-term memory (LTM) has to be taken into account. When something drops out of short term memory (STM), is it really gone? Sometimes the exact information is irretrievable, but the vast literature on implicit learning and priming suggests that everything that happens in

STM has some long-term impact. So what of the alternative account, that STM is no separate entity, but just a part of LTM? This would be a much more parsimonious solution, provided it can explain the empirical facts of a limited short-term store.

I would like to argue that an explanation of short-term store based on properties of LTM is much more interesting than assuming a separate entity. Why is the capacity four, and not five? A theory that proposes a buffer of limited size does not provide any answers. Take for example the subitizing phenomenon, the fact that people seem to be able to recognize up to four dots in the visual field, but have to count if there are more. One could postulate the theory that the visual system has an built-in capacity to recognize up to four things, and be done. Peterson and Simon (2000) offer an alternative account. According to their theory, the visual system can immediately recognize a set of dots, if it has seen these dots in the same array before often enough. The number of possible configurations of dots increases exponentially by the number of dots. Therefore, the human visual system receives enough examples of four-dot configurations to recognize any of them instantly, but not of five or more. Except of course when a particular configuration occurs often enough: anyone can recognize the five-dot pattern on a die instantly. The advantage of the Peterson and Simon account is that they show how the seemingly magical number four can be explained by an interaction between environment and the cognitive system.

Short-term memory capacity is not something that can be used to explain the outcomes of experiments, but is rather something that needs to be explained itself. One possible explanation is the one offered by the ACT-R architecture (Anderson & Lebiere 1998). ACT-R has a long-term declarative memory that also serves as working memory. To keep track of the current context, a single-item focus of attention is used. All items that have to be memorized are stored in declarative memory. Since declarative memory is activation-based, interference and decay can produce the same sort of effects usually assumed to be produced by limited STM. These limitations are, however, context dependent: if there are associations between the items to be memorized or with other items in memory, it is easier to retrieve the information. Short-term memory without context is only important if one presumes its capacity is a fundamental property. Short-term memory within a context is much more useful. I have demonstrated (Taatgen 1999a; 1999b) that individual differences on simple memory task might be used to explain individual differences in skill acquisition.

Figure 1 shows some results from an ACT-R model of short-term store that I have adapted from an earlier version (Taatgen 1999a). The original model memorized a list of up to ten digits, and attempted to reproduce them. It neatly reproduced Miller's 7 ± 2 effect. Since the original model was allowed to rehearse, I removed the rehearsal, and obtained the results in Figure 1: the magical number four, but without any internal capacity limitations. The figure shows three curves, a simulated low, average and high capacity individual. The individual differences were produced by variation of an ACT-R parameter that controls the spread of activation (based on Lovett et al. 1997). The reason why the curve drops off so dramatically at around four items has nothing to do with the number four itself. It rather has do to with the fact that as the string of numbers grows, the effects of decay, interference, and the increased probability of doing something wrong if more responses are required are multiplied, and cause performance to drop suddenly at this point.

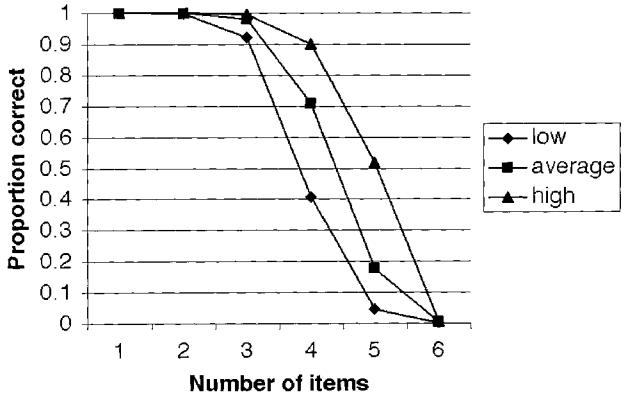


Figure 1 (Taatgen).

How to interface cognitive psychology with cognitive neuroscience?

Hannu Tiitinen

Department of Psychology, University of Helsinki, Helsinki FIH-00014, Finland. **hannu.tiitinen@helsinki.fi** **www.helsinki.fi/hum/ylpsy**

Abstract: Cowan's analysis of human short-term memory (STM) and attention in terms of processing limits in the range of 4 items (or "chunks") is discussed from the point of view of cognitive neuroscience. Although, Cowan already provides many important theoretical insights, we need to learn more about how to build further bridges between cognitive psychology and cognitive neuroscience.

Cowan's target article on the limits of mental storage capacity delineates experimental observations on short-term memory (STM) into either pure or compound limits with an average of 4 chunks and equates this limit with human attentional mechanisms. In my view, this is exactly what is needed in cognitive psychology and even more so in cognitive neuroscience today.

In cognitive psychology, the quest for an understanding of human attentional processes has its roots in the groundbreaking work of Cherry (1953), Miller (1956), and Broadbent (1958). As Cowan points out, sometimes these major findings become such a major influence, far beyond the expectations of individual researchers, that they can distort scientific progress. Over the past decades, observations have accumulated at such a rapid rate that it is very important to evaluate related findings and try to find some common denominators in the data. Cowan's target article is an important theoretical organizer and a logical continuation of his previous work (1988; 1995).

Current brain research has a tendency to focus on the study of a given response and its dynamics, obtained with a given method, but it is the theory, not the method, that should guide the researcher. This brings us to a few specific comments on the target article.

First, I would like to question the assumption of human brain oscillations, especially in the gamma (ca. 40 Hz) range, serving as a "binding mechanism" for sensory input and, in the present context, as the neuronal mechanism underlying chunking of information. The role of gamma activity is still under debate and, although the linking of STM and chunking with oscillatory brain activity is an attractive new idea, one must be very cautious in interpreting the data, such as those of Lisman and Idiart (1995), or as suggested by Cowan (sect. 4.1.2). As Cowan shows, one can calculate and show correlations in "cycles" and "subcycles," and estimate the ratio of slow and fast oscillatory rhythms in various ways, ad infinitum. Furthermore, 40-Hz oscillations are actually much more widespread in the frequency domain than is generally assumed and highly susceptible to, for example, the use of recording and filtering settings. Is this line of correlative research fruitful and promising enough? More specifically, can Cowan envision other prospective avenues of research which might help bridge the theoretical concepts of cognitive psychology and the measures of cognitive neuroscience?

In bridging the gap between these two domains in the research of STM and attention, at least two issues suggest themselves for

further inspection. (1) Experiments explicitly designed within the theoretical framework provided by Cowan and (2) the inherent limitations in brain research techniques.

Most of the neurophysiological studies cited in the target article do not seem to specifically address STM. For example, Gray et al. (1989) studied cats using invasive measures, and Tiitinen et al. (1993), using EEG, focused only on the attentional effects reflected in the human transient 40-Hz response. The latter observations were extended by Cowan (sect. 4.1.4) as consistent with the idea of a STM storage capacity of 4 chunks when taken together with Cowan et al.'s (1999; Fig. 4) very interesting simple relationship between attended and unattended speech, with attentional allocation presumably drawn from the same processes in both conditions.

This interpretation is however complicated by the fact that Tiitinen et al. (1993), using sinusoidal stimulation only, observed not only the attentional enhancement in the 40-Hz range, but a prominent response under passive (reading) conditions too. This already poses two variables that need to be taken into account in the quest for pure STM limits: the type of stimulation used (sinusoids vs. speech) and brain activity related solely to attention versus that observed in passive conditions (such as reading). Sensory-specific, task-independent brain processes (easily observable in passive conditions) might actually have a much more crucial role in "higher-order" stages of information processing than previously assumed (Tiitinen et al. 1994; Tiitinen & May, preprint). This, I feel, must be emphasized despite the fact that our understanding of sensory-level memory processes is still evolving (see, e.g., May et al. 1999).

A more straightforward way to understand the STM limitation might be the design of experiments based on the framework provided by Cowan. This impressive target article already suggests several interesting research avenues for brain research: For example, the results of Cowan et al. (1999) could be further extended in the context of EEG and MEG measurements, in which one could also take into account the problems mentioned above. An equally interesting issue is that of the flexible use of chunk sizes, which can range from small groups to "supergroups" (Ericsson et al. 1980; Ericsson 1985) or "active superconcepts" (Shastri & Ajjanagadde 1993). These not only provide humans with a powerful operating advantage in complex cognitive environments, but might even help explain the observed individual differences in STM limits. These observations can readily be made in fMRI and/or PET measurement, which might then shed light on where in the brain the STM chunks of varying size are located. These, and similar attempts, should eventually provide us with a map of the "architecture of cognition" and one of its important aspects: flexible memory storage and attentional influences.

Studies of STM properties in animals may help us better understand the nature of our own storage limitations: The case of birdsong acquisition

Dietmar Todt

Institute of Biology: Behavioural Biology, FU Berlin, D-12163 Berlin, Germany. **todt@zedat.fu-berlin.de**
www.verhaltensbiologie.fu-berlin.de

Abstract: I like Cowan's review of STM properties and especially his suggestions on the role of attention. I missed, however, a consideration of studies which provide evidence for STM properties in animals. In my commentary, I argue that such evidence can elucidate the biological basis of storage limitations, validating this view by discussing mechanisms which constrain the acquisition of serial information in songbirds.

In the introduction to his target article (sect. 1, para. 2) Cowan emphasises that "we are still uncertain as to the nature of storage capacity limits." The essence of this statement is substantiated in other parts of his paper, and I admit that I fully agree with it. Nevertheless, I would like to submit a proposal which is guided by a comparative perspective, suggesting that we should extend the research focus from the STM mechanisms of human beings to those of animals.

Several studies on animals have shown that limitations in the capacity to process acquired information are not unique to humans, but found in nonverbal organisms as well (Chen et al.; Koehler 1954; Roitblat et al. 1991; Terrace 1987; 1998; Todt et al. 2000). It may therefore be wise to select animal models that either allow inquiries into STM mechanisms which are difficult to address in humans, or that permit us to uncover memory properties which are biologically basic and thus point to precursors of accomplishments regarded as human characteristics. A paradigm for such an approach was recently published by Hultsch (1992; 1993). She studied acquisition of serially structured information in the nightingale (*Luscinia megarhynchos*), a songbird who is able to learn and memorise a vocal repertoire of more than 200 different sound patterns (= songs). In the following, I first recapitulate some background facts, then outline the methods and results of Hultsch's experiments, and finally discuss implications for special issues treated in the target article.

When singing, most songbirds produce a series of well-structured vocal patterns that typically have a duration of a few seconds and are called "songs" (Fig. 1). Although each single song encodes information about several biological details, for example, both individual and species-specific cues, many species develop and use repertoires of several different types of songs (Catchpole & Slater 1995; Kroodsma & Miller 1996). Song development is a matter of vocal learning (Hultsch 1991; Marler 1976; 1991; Nottebohm 1993). This accomplishment covaries, however, with the hierachy level of learning stimuli: Whereas birds normally copy the acoustical patterns of single songs, they appear less precise when learning at a higher level of song organisation. That is, their acquisition of information encoded in a series of different songs can be constrained by a mechanism called "package formation" (Hultsch et al. 1995).

The phenomenon of "package formation" reflects properties of STM involved in the acquisition of songs and indicates that this achievement is mediated by a process reminiscent of the chunking of items in human serial item learning. The phenomenon was discovered by analyses of groups of sequentially associated songs, or "song packages," respectively. These were developed by nightingales, who during their first weeks of life had been exposed to long strings of stimulus songs. Such packages had a mean size of 4 ± 2 songs (Hultsch & Todt 1989). After several studies on this matter, it became evident that the packages were not a result of song recall or memory retrieval, but induced by STM properties upon early auditory exposure to learning stimuli. In addition, "package formation" is explained by a model postulating a joint operation of the STM and a battery of submemories: When a young bird is exposed to a succession of several new songs, his STM processes information of about 2–6 different songs, and then transmits the information to a specific submemory where the second step, that is, a longterm storage of song material, takes place. It was hypothesised that, only after the transmission is the STM ready to process a further number of songs, which are then stored in a different submemory (Todt & Hultsch 1996; 1998). In other words, the model predicts that the STM segments a given sequence of learning stimuli and the battery of submemories consolidates this effect by processing the segments as different packages of songs.

To further test the memory model and, in particular, to clarify whether the segmentation would be a genuinely unit related (i.e., information constrained) process or whether time related effects could also play a role, Hultsch examined how young nightingales would cope with learning programs that differed in the rate of stimulus songs (Hultsch 1992). Programs were prepared by modifying the duration of silent intervals between successive stimulus

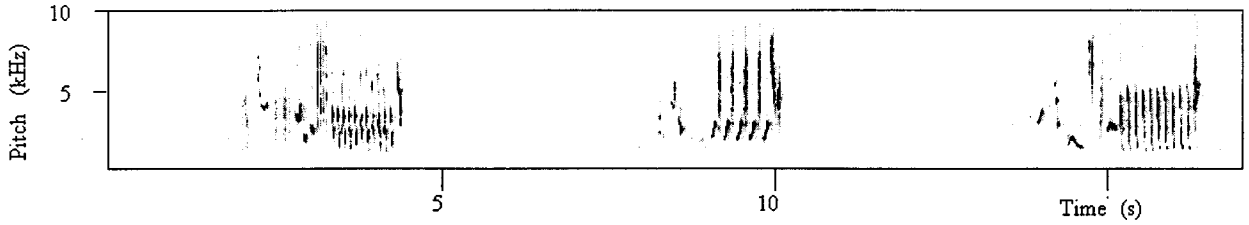


Figure 1 (Todt). Frequency spectrograms of three nightingale songs (section of a longer singing episode). These songs are developed by vocal imitation, and during their early acquisition processed analogous to "chunks" (see text).

songs. In one test series, the normal intervals (duration: 4 sec) were shortened to just 1 sec; thus 1 min included 13 learning stimuli here ("dense designs"). In a second test series, such intervals were prolonged up to 10 sec; thus 1 min included only 5 learning stimuli here ("spaced design"). For a contro1, the subjects were also exposed to temporally unmodified series ("norma1 design" that is, 8 learning stimuli per min). In order to adjust the experimental design to the methods of former studies, each of the three stimulus sequences was composed of 20 different types of songs which the birds could not experience in any other sequence or learning program. Analyses of singing behaviours performed by the trained birds yielded results that can make an interesting contribution to the issues raised by the target article (Fig. 2).

Above all, the study showed that the early segmentation song stimulus sequences is controlled by two components: a unit (or information) based capacity buffer (evidence: constraints uncovered by the "dense" program) and a time window based gating mechanism (evidence: constraints uncovered by the "spaced" program). The capacity buffer limited the sizes of stored packages at 3–5 songs. The time related component, on the other hand, limited such sizes by a time window of ca. 32 sec. Based on new evidence (Hultsch et al. 1999). This window indicates a specific span of attention as suggested in Cowan's article.

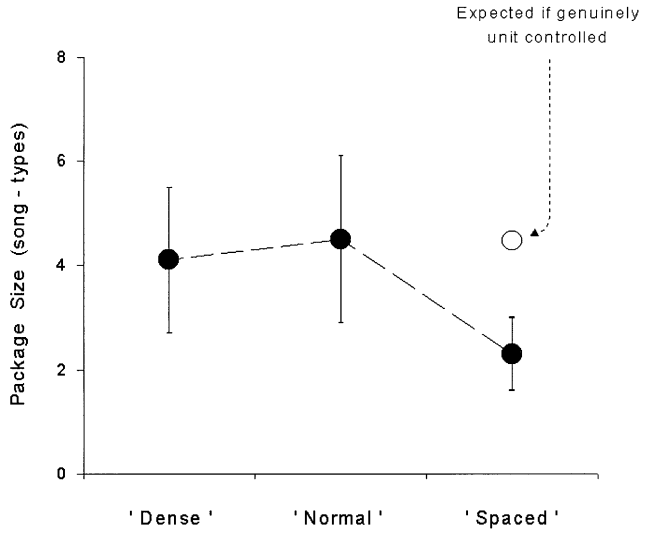


Figure 2 (Todt). Illustration of package sizes (number of associated song types) that nightingales (n=8) developed under three different learning programs. Filled circles show: Packages developed from the "dense" program (left) were not larger than those developed from the "normal" program (middle). This suggested unit-related capacity constraints on data processing to account for package formation. On the other hand, packages developed from the "spaced" program (right) were significantly smaller (see open circle) than predicted by a genuinely unit-controlled limitation process. This suggested that the postulated segmentation was concurrently controlled by time factors (see Hultsch 1992).

The good accord between properties of song acquisition in birds and properties of STM in humans, prompts three final comments. First, it seems evident that a study of STM properties in animals can help to elucidate the nature of our own storage limitations (Roitblat et al. 1991). Second, song acquisition is a useful biological model for comparative studies, for example, one can test memory mechanisms by applying learning stimuli that are completely new for a young subject and do not produce cognitive problems as do so many learning stimuli in human serial item learning (Todt et al. 2000). Third, the finding that birds use memory mechanisms which operate similarly to ours points to interesting parallels in the evolution. Since the brains of birds are not homologous to the those of mammals (Konishi 1989; Pepperberg 1993), similar operational properties suggest that they have evolved because they allow an optimal solution to a given problem; that is, here, a strategic acquisition of perceptional information.

ACKNOWLEDGMENT

The memory research on nightingales is supported by DFG grants.

Memory limits: "Give us an answer!"

John N. Towse

University of London, Egham, Surrey, TW20 OEX, United Kingdom.
j.towse@rhbnc.ac.uk **www.pc.rhbnc.ac.uk/jt/jt.html**

Abstract: Cowan has written a meticulous and thought-provoking review of the literature on short-term memory. However, reflections on one area of evidence, that of working memory span, shows the extent to which the research debate can be circumscribed by choice of experimental paradigms.

In the entertaining radio series, "The hitch-hiker's guide to the galaxy," Douglas Adams recounts how, many millions of years ago, a race of hyper-intelligent, pan-dimensional beings became so annoyed at all the constant bickering over the meaning of life, that they decided to solve the arguments once and for all (Adams 1985). They created Deep Thought, the most awesome computer imaginable and then gave him the task of finding the Ultimate Answer To Life, The Universe, and Everything. Deep Thought took his time to consider this challenge (seven and a half million years), but in the end he was ready. To the hushed anticipation of the universe, Deep Thought reluctantly announced that The Answer was in fact "42."

This produced understandable commotion. It was hardly a satisfactory pronouncement, a judgment that would allow everyone to sleep easy at night. But as Deep Thought pointed out, "I think the problem such as it was was too broadly based. You never actually stated what the question was." He had found the answer, but what exactly did the ultimate question ask? Deep Thought couldn't say, and therefore he had to design an even more powerful computer which could formulate the question to the ultimate answer (sadly, this computer, Earth, was destroyed just before it had completed the job, but that's another story).

The relevance of this whimsy lies in the illustration of how we can lose sight of the question that we're asking when searching for an answer. Cowan has amassed an impressive array of data that leads him to believe that The Ultimate Answer to Memory is "4." By constraining the circumstances for admissible evidence bearing on this question, Cowan has refined Miller's earlier conjecture that the answer is "7." In fact, whether you prefer "4" or "7" or some other number probably relates to whether you feel the various measurement conditions are appropriate or not, and Cowan's line of reasoning certainly needs to be taken seriously.

In the case of short-term memory, though, surely the question is quite transparent – "What is the limit of mental storage capacity?" So where's the fuss? Well, although we can ask this question and find an answer (even perhaps find a coherent answer) does that help us identify whether it is the best or only question to be asking? Unfortunately, this is less clear.

This problem is hardly Cowan's fault. Consider, for example, the area of working memory capacity, covered in the target article. Working memory capacity is conventionally estimated via working memory span tasks. These are supposed to require the simultaneous combination of "processing" and "retention." They exist in several forms such as reading span where individuals read a series of unrelated sentences for comprehension and afterwards attempt to recall each sentence-terminal word (Daneman & Carpenter 1980). Reading span is calculated on the basis of the number of sentences that can be presented and followed by correct recall of the relevant words. Now, one can argue about one or other theoretical account of performance (e.g., Towse & Houston-Price, in press) and doing so can certainly lead to a richer appreciation of working memory. It is important to recognise, however, that these theoretical debates all centre around what happens to span scores, that is, the maximum number of "things" that can be remembered. Under such circumstances, therefore, there are few options available to the scholar other than to do what Cowan has and derive an average value for memory capacity.

So although the target article briefly considers alternatives to the storage capacity account and finds them wanting (sect. 4.3), the evidential basis on which to evaluate the different accounts simply is not balanced. With so many studies of short-term memory, let alone working memory, based on span scores or their equivalent, we run the danger of rigging the contest. That is, predetermining the sort of answer we will look for and therefore find.

Recent, as yet unpublished collaborative work with Graham Hitch, Una Hutton, and Zoë Hamilton has begun to explore other ways of asking questions about working memory. Rather than focus on (just) the number of items that children can remember, we have asked whether, in variants of working memory span tests, children show meaningful and reliable variations in the temporal endurance of their memory traces. Potentially at least, some cognitive skills may rely more on the extent to which representations can be preserved, as opposed to the number of stimulus neighbours that can be tolerated in memory tests. Early indications of this research programme seem promising, but in the present context, it is perhaps the attempt to seek alternative dependent variables that is most relevant. Even if endurance measures turn out to be redundant, or ineffective, that itself seems an important conclusion to reach because thus far one must assume span gives all the information one needs. Strong faith in measures of memory size alone may permit us to find The Answer, but at what cost to a full understanding of immediate memory?

ACKNOWLEDGMENTS

The commentary is underpinned by support from the ESRC, grant R000222789 (awarded to John Towse, Graham Hitch, and Una Hutton). Discussions with Robin Walker are also appreciated.

Neural mechanism for the magical number 4: Competitive interactions and nonlinear oscillation

Marius Usher,[a] Jonathan D. Cohen,[b] Henk Haarmann,[c] and David Horn[d]

[a]*School of Psychology, Birkbeck College, University of London, London WCJE 7HX, United Kingdom;* [b]*Department of Psychology, Princeton University, Princeton, NJ 08544;* [c]*Department of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742;* [d]*Department of Physics, Tel Aviv University, Tel Aviv, Israel.* **m.usher@bbk.ac.uk jdc@princeton.edu hhaarmann@hesp.umd.edu horn@neuron.tau.ac.il www.psyc.bbk.ac.uk/staff/homepage/mu.html www.bsos.umd.edu/hesp/haarmann.html www.neuron.tau.ac.il/~horn**

Abstract: The aim of our commentary is to strengthen Cowan's proposal for an inherent capacity limitation in STM by suggesting a neurobiological mechanism based on competitive networks and nonlinear oscillations that avoids some of the shortcomings of the scheme discussed in the target article (Lisman & Idiart 1995).

Cowan interprets the capacity limitation of 4 ± 1 items, in relation to a bottleneck on the focus of attention. While this may be a plausible explanation, it is somehow paradoxical that recall of *unattended* material should provide a measure for the capacity of the focus of *attention;* with diverted attention, recall may be mediated by the residual decay limited components in STM. Nevertheless, we believe that the rest of the reviewed data provides convincing support for a capacity limitation of about 4 items in STM. This raises the challenge of providing a plausible neurobiological mechanism for it. Cowan adapted the model of Lisman and Idiart (1995), explaining the capacity limitation in terms of two wave frequencies. While this adaptation needs to be tested, its likelihood for confirmation is low, as the γ and θ waves used in the original schemes were based on empirical estimates of neurophysiological data. Moreover the frequency of these oscillations is broad and therefore their ratio is likely to fluctuate widely rather than provide a stable value corresponding to a specific capacity.

An alternative neurobiological account does not rely on precise frequency values but explains the capacity limitation in terms of inherent properties of competitive networks (Horn & Usher 1991; 1992; Usher & Cohen 1999). The main idea for consideration is that while LTM is mediated by structural changes in connectivity, STM (which is associated with awareness) is mediated by neuroelectric reverberations (Hebb 1949), subject to competitive interactions, whose need has often been discussed in experimental and computational neuroscience (Desimone 1998; Grossberg I976; Usher & Cohen 1999; Usher & McClelland 1995) in relation to attentional selection. Typically selection is implemented in competitive models by a mechanism of strong lateral inhibition, resulting in a winner-take-all system. We have recently presented a model which proposed that the lateral inhibition can be modulated in relation to task demand (Usher & Cohen 1999). The model is described by the following equation:

$$dx_i/dt = -x_i + \alpha F(x_i) - \beta \Sigma_{j \neq i} F(x_j) + I_i + noise$$

where x_i is the activation of the i representation, $F(x) = x/(1+x)$ is the "activation-functions" (see e.g., Tsodyks et al. 1998), I_i is the sensory input to each representation, α corresponds to recurrent self excitation and β to lateral inhibition. The self excitation allows the maintenance of activation after the input I_i is turned off. The inhibition parameter is set high when selection is required and moderate when multiple items need to be maintained together, as in immediate recall tasks. In this model we showed that moderate levels of lateral inhibition allow the coactivation of several memory representation. More interestingly, the system shows a sharp capacity limitation where only a small number of items can be simultaneously maintained. This capacity depends on the excitation and inhibition parameters and is within the

range of 3–5 items when these parameters are chosen as $\alpha = 2$ and $.15 < \beta < .20$.

Functional considerations demonstrate the system cannot increase its capacity beyond this range by diminishing the inhibition β parameter even further. Due to the recurrent excitation and to small overlaps in the input (or in the representations themselves) a minimal amount of lateral inhibition is required to prevent the activation to spread to all the memory representations. Assume, for example, that when the representation i receives input, I_i is relatively high (.33 in the simulations reported), while other representations $j \neq i$ receive a much diminished input ($I_j = .05$, due to overlaps). In this situation, we find that when the inhibition is small ($\beta < .19$) activity spreads to nonsensory-activated units, even when only a single unit, *i*, receives input. This unreliable relation to its inputs would make the system useless. Increasing the inhibition to $\beta = .19$ prevents this problem, but limits the number of co-active units to 3–4 items. Changes in the excitation parameter do not affect those considerations. If α is increased alone, the capacity increases, however, increasing the excitation requires a simultaneous increase in the minimal inhibition to prevent the unbounded spread of activation within the system. When the two parameters are changed together to preserve reliability the capacity remains in the 3–4 items range.

Does then the capacity result from a specific layout of biological parameters that characterize the system or is there a magical number 4 to be found? Two tentative answers can be suggested. In the model described so far, the capacity can be increased only by reducing the representational overlap. Such reductions may be, however, bounded due to another trade-off (Cohen 1996). Representational overlap is essential for computations that perform generalization. Thus it is possible that the system evolved so as to optimize the functions of active maintenance and generalization, providing another teleological motivation for the STM capacity (cf. Kareev et al., in the target article).

Another answer which explains a magical number 4 is suggested if one assumes that maintenance in STM requires not only the simultaneous activation of representations, but also their segmentation from each other. This can be performed by a system based on staggered oscillations as described here, but with higher inhibition, and with the addition of an adaptation/recovery variable that makes each representation oscillate (Horn & Usher 1991; 1992). Due to the lateral inhibition one can generate a situation where each representation is activated in a different phase of the temporal oscillation, as illustrated in the Figure 1 (showing segmentation with 3 units oscillating out of a background of non-active units).

Computational and mathematical analysis of this system demonstrated a fixed capacity of 3–5 items in the ability of this system to perform segmentation (Horn & Opher 1996). The system has maximal stability for 3 out-of-phase oscillations. For more than 5 oscillations the system is unable to keep each of these oscillations in a distinct phase. This characteristic does not depend on the precise values of parameters (except that they need to be within the range leading to staggered oscillations) and is a result of the fact that it is not possible to compress more than 5 narrow non-linear oscillations (one for each activated representation) within each cycle of the whole system. This mechanism explains the 4 ± 1 capacity limitation, however, unlike the model described in the target article (Lisman & Idiart 1995). It is robust to parameter changes and can function for a wide range of frequencies, resulting from competitive interactions between nonlinear oscillators.

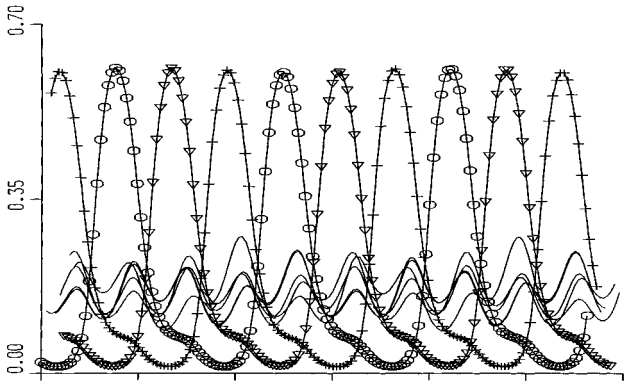


Figure 1 (Usher et al.). Segmentation of three units oscillating out of a background of non-active units.

Over the top: Are there exceptions to the basic capacity limit?

John Wilding

Department of Psychology, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, United Kingdom. **j.wilding@rhbnc.ac.uk**

Abstract: Can we identify individuals with a larger basic capacity than Cowan's proposed limit? Thompson et al. (1993) claimed that Rajan Mahadevan had a basic memory span of 13–15 items. Some of their supporting evidence is reconsidered and additional data are presented from study of another memory expert. More detailed analysis of performance in such cases may yield different conclusions.

Cowan argues for a capacity limit on short term memory of about four items (possibly rising to a maximum of six); this raises the question of whether some individuals have still larger capacity. Thompson et al. (1993) studied the memory ability of Rajan Mahadevan. His digit span before he began deliberate practice at memory tasks was recorded as 15 items and other evidence supported this estimate.

Thompson at al. (1993) argued that the lag between the end of list presentation and response initiation should remain constant until list length exceeds the basic span and rehearsal becomes necessary. In support of this, they found that individuals with normal digit span showed a constant lag until list length exceeded 7 items; lag then rose steadily as list length increased. Rajan, however, showed no rise until list length exceeded 13 items. These data exemplify one of Cowan's acceptable methods for measuring limited capacity; the discontinuity in lag is assumed to reflect the point at which basic span is exceeded. However, the data indicate a discontinuity at seven items in normal memorizers, rather than at Cowan's proposed limit of four items. Rajan's basic span of 13 items suggests that some individuals may have much greater capacity. With colleagues, I have attempted to identify other such individuals, using a computerised version of the task and recording additional measures, which clarify whether lag is a valid indicator of basic span.

Lists of digits were presented visually at 2 items per sec; an asterisk indicated the end of the list and requested recall, which was achieved by typing numbers on the keyboard. Time to initiate recall and to input each item was measured, plus the positions of errors. Initially a five-item list was presented and 1ist length increased by one item following a correct response and decreased by one item following an error, matching Thompson et al.'s (1993) ascending method. Twenty trials formed a run. Data have been collected from two memory experts, but results are discussed here only from MB, who comp1eted 50 runs on the task. MB held the world record for reciting the expansion of pi in 1977 and later achieved correct recall to 7,769 places. He can recite the first 1,000 digits of pi in 170 sec. He claims special memory ability only for numbers and uses none of the standard mnemonics, relying only on his knowledge of mathematics, which he teaches at secondary school level. If his digit span depends on chunking based on mathematical knowledge, we would expect a discontinuities in the lag measure around the normal span and between chunks in the inter-item recall latencies.

Table 1. (Wilding). *Time to initiate recall(s) for lists of 5 to 17 items in a visual digit span task*

List Length	5	6	7	8	9	10	11	12	13	14	15	16	17
MP	1.1	1.3	1.3	1.3	1.3	1.6	1.5	1.5	2.1	2.1	2.3	2.1	2.6
Rajan	—	—	0.2	0.3	0.2	0.3	0.7	0.7	1.4	4.2	3.0	8.4	11.2

The mean of the longest span achieved on each run was 14.96 (range 12–18, s.d. 1.4), less outstanding than Rajan, but about twice the norm. Table 1 gives the time to initiate recall at each list length (successful recalls only) for MP and corresponding figures for Rajan (estimated from Thompson et al., Fig. 2.1; absolute times are not comparable, due to the different methods of recording used). While Rajan certainly shows a steep rise in times for lists of more than 13 items, there is also some suggestion of an increase at 11 items; no data are given for lists of five or six items and presumably measurement of times by stopwatch was imprecise for short intervals. MP, however, shows a gradual increase, with possible stops at 10, 13, and 17 items. There is, however, no clear discontinuity, which would unambiguously indicate a basic span limitation. The other expert we tested, who overtly converted numbers into images, showed a sharp discontinuity at 9 items, followed by a continuous rise in times, like Thompson et al.'s (1993) subject GN, who also developed a mnemonic method, so the pattern shown by MP was not some artefact of the measurement method.

Inter-item latencies for successful recall, however, provided unambiguous evidence for chunking in MP's recall; however, the pattern obtained was only partly consistent with Cowan's argument. Figure 1 gives cumulative recall times for lists of 5 to 16 items. Once lists exceeded 7 items, a pause occurred after the fourth item, once lists exceeded 11 items, another pause occurred after the eighth item, and once lists exceeded 15 items, a third pause occurred after the twelfth item. MP apparently divided the lists initially into groups of four items (though lists of 14 items seem to be divided 4–4–3–3), but the final group could be as large as seven items before it was further subdivided. MP confirmed that, though he does not have a deliberately practised strategy for memorisation, he tried to find some mathematical pattern or relation, normally in each group of four.

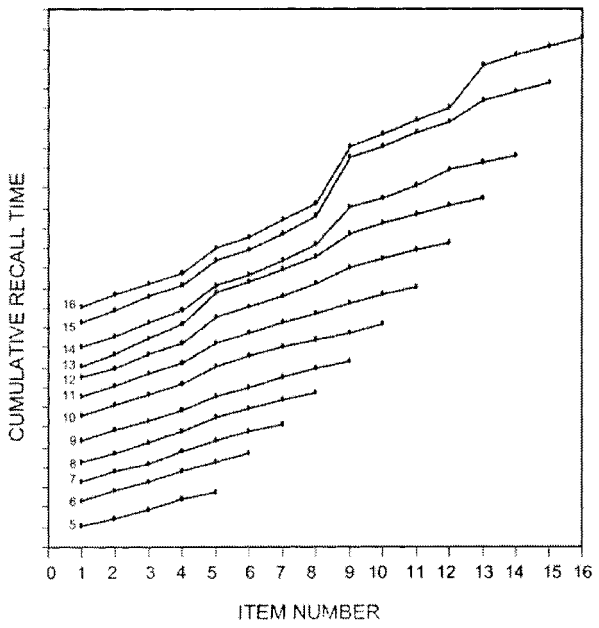


Figure 1 (Wilding). Mean cumulative recall times by MP for each item in lists of 5 to 16 items. Numbers on the left indicate list length for each function. As list length increases, each function has been displaced upward by 1 sec to avoid overlap.

While any conclusions clearly need data from other individuals before they can be firmly established, these results imply that: (1) Thompson et al.'s (1993) response lag measure is not decisive as an indication of basic span. Though no clear discontinuity in MP's data indicated a basic span limit, discontinuities in inter-item latencies suggested chunking was occurring. (2) items were in general grouped into fours, but such grouping was not mandatory and up to seven items were treated as a group in some circumstances. Such flexibility may explain the long running argument as to whether the limit is four or seven items and raises further questions about the mechanisms involved.

Attention is not unitary

Geoffrey F. Woodman, Edward K. Vogel, and Steven J. Luck

Department of Psychology, University of Iowa, Iowa City, IA 52242-1407.
{geoff-woodman; steven-luck}@uiowa.edu vogel@sdepl.ucsd.edu
www.psychology.uiowa.edu/Faculty/{woodman/woodman; luck/luck}.html

Abstract: A primary proposal of the Cowan target article is that capacity limits arise in working memory because only 4 chunks of information can be attended at one time. This implies a single, unitary attentional focus or resource; we instead propose that relatively independent attentional mechanisms operate within different cognitive subsystems depending on the demands of the current stimuli and tasks.

Cowan's model of working memory posits that the limited capacity of working memory is entirely caused by the limited capacity of attention. He further proposes that a unitary mechanism of attention operates across cognitive subsystems, with a single capacity that must be shared across subsystems. Here, we argue in favor of an alternative view in which the brain makes use of a collection of loosely interconnected attention mechanisms that operate in different cognitive subsystems and that reflect the properties of the individual subsystems. We will focus specifically on the operation of attention in three coarsely defined cognitive subsystems, namely perceptual analysis, working memory, and response selection.

We would expect that different mechanisms of attention would operate in visual perception and in visual working memory for the simple reason that these subsystems appear to operate at very different rates. For example, Potter (1976) showed that observers can identify pictures of complex real-world scenes at rates of up to 8 pictures per second, but much slower presentation rates were required for observers to store the scenes in working memory. More recent studies have used a combination of psychophysical and electrophysiological measures to demonstrate that attention shifts at different rates in visual perception and in visual working memory. Specifically, studies of the attentional blink phenomenon, which reflects the operation of attention in working memory, indicate that hundreds of milliseconds are required to shift working memory-level attention from one stimulus to another (Vogel et al. 1998). In contrast, studies of visual search, which empha-

sized the operation of attention in perception, have indicated that shifts of perceptual-level attention can occur every 50–100 msec (Woodman & Luck 1999). Horowitz and Wolfe (1998) and colleagues have also provided evidence that the operation of attention in perception is relatively independent of memory.

Given that attention operates at different speeds in perception and in working memory, it would be sensible if these mechanisms of attention operated asynchronously; otherwise, the fast perceptual system would be continually waiting for the slow working memory system to catch up. We have recently provided evidence for asynchronous operation of perception and working memory by showing that shifts of attention during visual search are not slowed when visual working memory is full (Woodman et al., in press). In this study, participants were required to maintain a set of object representations in visual working memory while performing a visual search task. We observed that subjects could perform a difficult visual search task just as efficiently when working memory was full as when working memory was empty, indicating that perceptual-level attention can be allocated and shifted very efficiently even when working memory-level attention is operating at maximal capacity.

It is also reasonable to suppose that visual perception and visual working memory might differ in their spatial properties as well as in their temporal properties. In particular, visual perception relies on topographically mapped representations, but there is no evidence that the representation of objects in working memory is topographic. Thus, perceptual-level attention might be expected to have various spatial properties that are absent from working memory-level attention. A recent study by Vogel (2000) supports this hypothesis. Specifically, Vogel found that working memory-level attention can be divided among multiple noncontiguous locations without being allocated to the regions between these locations, whereas perceptual-level visual attention must be focused on a contiguous region of the visual field. Thus, perceptual-level and working memory-level mechanisms of attention differ in both their spatial and temporal characteristics.

It also appears that the mechanisms of attention that operate during response selection and initiation are different from those that operate at earlier stages. The operation of attention during response-related processes has been studied by means of the psychological refractory period (PRP) paradigm. In this paradigm, two stimuli are presented to subjects in rapid succession and the subjects make separate speeded responses to each stimulus. Reaction time to the second stimulus is slowed when the delay between the two targets is short, and several studies indicate that this is due to capacity limitations at the stage of response selection (Pashler 1994). Pashler (1991) and Johnston et al. (1995) have provided evidence that shifts of visual-spatial attention do not operate at the same stage as the capacity limits that are observed in the PRP paradigm.

In conclusion, there are now many forms of evidence indicating that there are different mechanisms of attention that operate with different properties within different cognitive subsystems. Moreover, it is worthwhile to ask what is gained by proposing that the limited capacity of working memory arises from the limited capacity of attention. The real question is why there are limits at all, whether we call them limits of attention or limits of working memory-specific resources.

Author's Response

Metatheory of storage capacity limits

Nelson Cowan

Department of Psychology, University of Missouri, Columbia, MO 65211.
cowann@missouri.edu **www.missouri.edu/~psycowan**

Abstract: Commentators expressed a wide variety of views on whether there is a basic capacity limit of 3 to 5 chunks and, among those who believe in it, about why it occurs. In this response, I conclude that the capacity limit is real and that the concept is strengthened by additional evidence offered by a number of commentators. I consider various arguments why the limit occurs and try to organize these arguments into a conceptual framework or "metatheory" of storage capacity limits meant to be useful in future research to settle the issue. I suggest that principles of memory representation determine what parts of the representation will be most prominent but that limits of attention (or of a memory store that includes only items that have been most recently attended) determine the 3- to 5-chunk capacity limit.

R1. General reaction to commentaries

In *BBS*'s first round of refereeing of the target article, it seemed to several of the referees that the paper was not controversial enough to generate useful commentaries. I think that the present set of commentaries has proven to be very interesting, diverse, and thought-provoking after all.

By my count, about 15 of the 39 commentaries solidly accepted the hypothesis that some core working memory faculty is limited to about 4 separate chunks of information. Indeed, at least 8 commentaries (**Lane et al.; Nairne & Neath; Pothos & Juola; Rensink; Rypma & Gabrieli; Todt; Usher et al.; Wilding**) presented additional evidence for a 4-chunk limit in many circumstances. However, at least seven other commentaries seemed strongly opposed to that concept and the remaining 17 seemed less committed either way. Many of the commentators who more or less agreed with the 4-chunk hypothesis offered alternative theoretical explanations.

R1.1. Self-justification

I believe the target article has made three main contributions. First, principles were offered as guidelines for identifying experimental situations in which the number of separate chunks can be estimated. Second, a diverse field of evidence was shown to yield similar estimates of capacity in those identified situations; specifically, 3 to 5 chunks on the average, among normal adult humans. Third, various plausible theoretical explanations for the capacity limit were described.

The eligibility of evidence often depended upon the application of critical theoretical assumptions. For example, one theoretical assumption provides an answer to **Usher et al.**, who commented that "it is somehow paradoxical that recall of *unattended* material should provide a measure for the capacity of the focus of *attention*". It is not so paradoxical if one considers that, ordinarily, attention exerts effects broadly across the perceptual encoding, mnemonic storage, and retrieval processes so that any one particular effect of

attention is difficult to separate from others. By diverting attention during encoding and storage phases, the limitations of attention during retrieval, in particular, can be isolated and identified. Under these circumstances, chunking processes are limited and attention can be seen to have a limit of about 4 chunks of information.

It now seems likely that further, interesting cases of the capacity limits will emerge. For example, Jinks and Laing (1999) investigated capacity limits in odor processing, stating in their literature review (p. 311) that "the discrimination and identification of components in odor mixtures is a difficult task for humans, with 3–4 odorants being the limit."

Although I argued for particular fixed capacity limits, I did not offer a precise, detailed theory of these limits. Some commentators (especially **Halford et al.; Pascual-Leone; Milner; Schubert & French**) seemed disappointed that the rules for determining chunks or storage requirements were not worked out with more theoretical precision, and that a clear theory of the capacity limits was not proposed. I agree that such clarity would be desirable if the resulting theory were correct. However, I disagree with these commentators' prescriptions for reaching that goal. In the 45 or so years since Miller's (1956) article, this may be the first extensive discussion in which an attempt is made to consider the relevance of the burgeoning cognitive literature, taken whole, to the notion of capacity limits. It seems most important now to sort out and compile the basic messages from diverse paradigms in order to provide guidance for theory, and it would seem unwise to attempt too much theoretical precision at the same time.

Many of the commentaries offered theoretical views that were more precise than my own but those theoretical views conflict with one another (as discussed in sect. R4), and therefore cannot all be correct. If I had foreclosed on a particular theoretical explanation of the 4-chunk hypothesis, I suspect that the empirical generalization would tend to be ignored as soon as the particular theoretical explanation of it was disproved, as it most likely would be. Therefore, I staunchly continue to view the theoretical indecision of the target article as a strength, not a weakness. It took 45 years for the field to advance from the notion of chunking as an aid to remembering, given an unknown capacity limits pegged at about 7 items (Miller 1956), to an earnest quantification of those limits and the generation of some testable explanations of the limits. If the result of this article is a new surge of research and theoretical debate on the reasons for capacity limits, I believe that the field will be well-served.

If a fundamental branch of the tree of possible hypotheses (Platt 1964) could be ruled out, that would be the soundest gain that reasonably could be made at this time. Toward this end, a taxonomy of capacity theories will be suggested as a "metatheory" of capacity, and I will consider how theories might be compared and assessed.

R1.2. Organization of the responses to specific commentaries

Considering the commentaries can take us a fair distance toward understanding capacity limits: Toward that end, I will first consider evidence for and against the 4-chunk hypothesis (sect. R2). In the process, arguments for alternative formulations will be considered, including the hypothesis that there is a capacity limit of only 1 chunk, the hypothesis that there is a capacity limit of 7 chunks as Miller (1956) suggested, and the hypothesis that the number of chunks that can be brought into working memory concurrently is unlimited. The next section (sect. R3) will be directed at alternative cognitive theoretical accounts of capacity limits. A taxonomy of theoretical positions will be described; it is this taxonomy that serves as a basic cognitive "metatheory" of capacity limits alluded to in the title of this response. Biological underpinnings of the capacity limit will be discussed (sect. R4) and, finally, I will attempt to reach a tentative verdict as to the reason for capacity limits (sect. R5).

R2. Validity of the 4-chunk hypothesis

The target article has put forward considerable evidence that, in situations in which a reasonable assumption can be made about how many separate chunks are in the focus of attention, a consistent capacity estimate of 3 to 5 chunks emerges. Naturally, this evidence hinges on the ability to identify chunks correctly. This section first presents a parable that helps to illustrate the nature of controversies like the 4-chunk hypothesis (sect. R2.1). Next, the key issue of how to identify chunks is taken up (sect. R2.2). This is followed by a detailed discussion of alternative hypotheses about capacity limits (sect. R2.3). Finally, there is a discussion of a new method for the quantification of capacity limits in visual array comparison tasks (e.g., Luck & Vogel 1997), which is capable of adding considerable strength to the theoretical framework.

R2.1. Capacity limits and methods in science: A parable

Naturally, a capacity limit would be only one of several factors influencing memory performance; another, for example, would be the contribution of information that can be replaced into the focus of attention from long-term memory (i.e., the information within each chunk). Such other factors will cause apparent departures from the 3- to 5-chunk limit that certainly allow room for doubters. It might prove helpful to provide an analogy from an area of physics, illustrating how one would expect the 4-chunk hypothesis to be criticized even if it were true. Such an illustration cannot prove the hypothesis, of course, but it can at least encourage further research designed to identify and explain capacity limits.

Thus, consider the case of a now-proven concept in physics, gravity, and how it might have become established in science (though many of these historical details may be fictitious). Everyone could plainly see that most objects thrown into the air fall down again. Moreover, some saw that objects fall at an accelerating rate that could be roughly estimated even though the correct equation was unknown. Although most scientists informally talked of gravity-related ideas, they could not measure gravity exactly. Some argued that it is not a law because some objects do not come down (stars and planets) or come down at much slower, unpredictable rates (birds, bats, flying squirrels, leaves, and dust). Proponents of gravity then admitted that other considerations such as distance (in the case of planets) and air resistance have to be taken into consideration. Formulation of a complete model taking into account acceleration, size, distance, and air resistance was a long way off. Some believed in the

theoretical concept that there was a certain rate of acceleration at which objects tend to fall, even though it only provides an approximation and there are boundary conditions to its use. Given that some of the boundary conditions were unknown, many people remained agnostic and a few denied the existence of gravity as a general rule. It may only be through hindsight that the concept of gravity has come to seem obvious and universal. We have some detailed models of it now, but perhaps we must wait for a future generation to determine exactly why gravity occurs.

The known history of the discovery of gravity seems consistent with this progression of thinking. Tycho Brahe established a wealth of observations on planetary movement, an arena in which there is little wind resistance unlike earthly applications of gravity. Johannes Kepler used Brahe's observations to establish regularities or laws of planetary motion, though without really understanding the general principles behind the laws; and later, Isaac Newton "stood on the shoulders" of Brahe and Kepler to establish more general laws of gravitational force. Albert Einstein later reformulated the law of gravitation in a more penetrating manner that tied it to other forces in the universe, representing it as curvatures in space; and surely the final word is yet to come.

The history of the memory capacity concept arguably could be similar to this. On this topic we may have Brahe, and this target article reaches toward Kepler, while merely speculating about Newton and Einstein. Of course, the force of this analogy would quickly break down were I to choose a shakier concept (e.g., the concept of ether occupying empty space, or of a life force). However, it is worth contemplating that apparent exceptions to a causal principle such as capacity do not automatically rule it out if a plausible case can be made as to why the exceptions occur. The fact that we do not yet have a complete theoretical model of capacity (at least, not a confirmed one) does not rule it out, either.

It is with this parable in mind that I would address criticisms such as the one **Avons et al.** made, that "The capacity estimates cited by Cowan are less consistent than he claims." We may not have a clear enough understanding of the "wind resistance" type factors in psychological performance, but we know enough to support a "gravity" type concept (i.e., the capacity concept). So, for example, when Avons et al. state, "If four of the most recently presented items exist within the attentional focus, then recall of about four of the most recent items would be expected." I do not fully agree. The recall can sink below 4 not only because some individuals have a lower span than that (e.g., see Fig. 3), but also because some attention may be allocated earlier in the stimulus sequence or elsewhere entirely. This type of variation is bound to depend on factors such as the motivational level of subjects, the amount of diligence and vigilance that the task requires, the task complexity, and so on.

R2.2. Can we identify chunks?

Obviously, if chunks cannot really be identified, one cannot quantify the capacity limit in terms of chunks. Several commentaries (**Beaman; Schubert & Frensch**) questioned the target article's ability to determine chunks. Undoubtedly, this is a key issue. The theoretical definition of a chunk offered in section 1.3 was "a collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use" (mathematically, an equivalence class). To apply that definition without direct measurements of chunks, one needs a task analysis of every procedure. It is difficult to see how it would be possible to do better than to analyze each task in a logical manner, taking into account what is known about the task. The fact is that when I used this case-by-case method, the result was a capacity estimate that remained fairly consistent at 3 to 5 chunks across a wide variety of situations. This constancy in the results itself lends credibility to the method, except in the eyes of those who suspect an unfair use of the method. They must try it themselves.

R2.2.1. A strategy for research to identify chunks. Clearly what is needed in the future is an attempt to determine chunks more empirically in many relevant procedures. This would be a major shift of emphasis for researchers of working memory. Measures of reaction time may be of limited usefulness because chunking of two elements, A and B, is only one possible reason why these elements are recalled together quickly. Supplementary types of relevant evidence that two elements are chunked together may include (1) recall of A and B consistently near each other, and (2) the high probability of recalling B conditional on recall of A, or vice versa. Another promising strategy would be to begin with unconnected units and teach the subjects new chunks (with learning assessment), so that the chunks are known to the experimenter before they are used in a working memory task.

Kleinberg and Kaufman (1971) carried out one procedure in which the role of chunking in recall was tackled rather directly. Subjects were presented with clusters of 13 dots, each forming a complex, symmetrical pattern. The vocabulary for a particular group of subjects consisted of 2 or 4 such distinct patterns. Half of the subjects learned to group pairs of patterns together to form larger chunks that had verbal labels. Each test array consisted of an 8×8 matrix of such dot patterns. Recall depended on the exposure duration, which also was varied, but asymptotically high performance levels were obtained with an exposure duration of 6 or 8 sec. For those exposure durations, recall amounted to about 3–5 of the designated chunks on day 2 for subjects who received 4-pattern vocabularies, or for subjects who received 2-pattern vocabularies and learned to chunk them. Recall was at about 8 alleged "chunks" for subjects who received a 2-pattern vocabulary but were not able to chunk them; but one could well imagine that these subjects learned to chunk the patterns into pairs even without training in that condition (there being only 4 possible pairs, AA, AB, BA, and BB). On days 3–4, there was a slight increase in performance in all groups, which could indicate further chunking.

Illustrating how much work is left to be done, there may be subtle strategic modes that subjects can adopt to alter the ways in which chunking takes place. **Schneider et al.** failed to replicate an instance of grouping observed by Luck and Vogel (1997), who found that subjects could recall as many items in arrays of dual-color squares (composed of one color peripherally and a different color centrally) as they could single-colored squares. According to the evidence of Luck and Vogel, each dual-colored square was perceived as a chunk. Schneider et al. instead found, using a very similar procedure, that memory for dual-colored

squares was considerably poorer, suggesting that any such chunking does not take place automatically. The simplest kind of explanation is that subtle differences in instructions, subjects, or materials promoted chunking in one case but not the other. We do not yet know what they are.

Schubert & Frensch emphasized the need for an operational definition of chunks, saying, "what prevents us from arguing that 4 items held in visual short-term memory (STM) and 4 legs of a chair have a common causality in the limitation of subjects' attentional focus?" Obviously, in the absence of a detailed model (and not just a model, but a *correct* detailed model), one must use educated judgment. (**Beaman** put the situation in its proper light by calling the task of determining the number of chunks "unenviable." Certainly, it is the entire field of investigators interested in capacity limits who must share this burden.) Schubert & Frensch suggested an experimental approach in which one would identify a process or mechanism (such as the attentional focus) that is essential for various tasks showing an apparent capacity limit, susceptible to manipulation of a Factor X. The key finding would be that manipulating Factor X would affect all of the tasks similarly. This is indeed a key research strategy. However, I would emphasize that one must be careful to distinguish between the *4-chunk capacity limit* (that only 4 separate chunks can be held in the focus of attention at one time) and the *theoretical explanation* for this limit. The latter topic will be taken up in detail in section R5 but requires a brief discussion now. According to the type of theoretical account that I favored in the target article, the limit is in the capacity of the focus of attention itself. If this is the case, X could be a manipulation of the amount of supplementary material that must be held in the focus of attention during the working memory task. However, according to an alternative theoretical account, there is something about the representations of information of various kinds that prevents the focus of attention from taking in all of the information. For example, perhaps the most distinct or prominent 4 or so chunks in the field of stimuli somehow overshadow the others, making them unavailable to conscious awareness. In that case, the same research strategy applies but the successful X might be some manipulation that affects the distinctiveness of chunks in the stimulus field.

R2.2.2. Further clarification of chunking: Binding and chunking in working memory tasks. Commentator **Beaman** made the important observation that "A number of studies which Cowan takes as evidence for his 4-chunk capacity limit were serial recall studies in which items were only marked correct if they were recalled in the correct serial order." **Nairne & Neath** commented similarly, as did **Avons et al.** I completely agree that this is important in understanding the capacity limit. In the experiments of serial recall, a small number of items were used over and over on every trial (e.g., spoken digits in the study of memory for unattended speech by Cowan et al. 1999). Presumably, all of the items remained active in memory (or readily accessible in the test context). It was instead the *binding* of items to particular serial positions within lists that comprised the material that must be held in the focus of attention. It would be easy to guess that a particular item was a member of the current list, so it was the binding of items to their serial positions that was most indicative of working memory. Similarly, in tasks involving the presentation of a simultaneous array of items (e.g., the color squares of Luck & Vogel 1997), it is not difficult to guess colors that were presented so it is the binding of colors to spatial locations that is critical.

Given the importance of binding, a chunk might be described in these procedures as a direct association between adjacent items, which together become bound to only one higher-order serial position or spatial location in the stimulus array, as a group. For example, in a visually presented letter span experiment, upon seeing F-B-I-C-I-A, one might attach the chunk FBI to a first serial position and the chunk CIA to a second serial position. In a color-array memory task, if one were allowed to study the color array for some time, chunks would consist of multi-color patterns (most likely composed of three or four colors) assigned to spatial locations in the array. Perhaps this type of consideration will help to provide the narrower definition of chunks that **Schubert & Frensch** craved.

Beaman also noted the use of serial-position-specific scoring and suggested that "If capacity as measured by serial recall studies is indeed 4 chunks, then capacity is for 4 chunks plus some extra information connecting those chunks." No, because the items themselves (presumably the chunks in the relevant procedures) do not have to be held in a limited-capacity store. Stated more precisely, they are automatically activated or accessible and only their binding to serial positions within a list (or spatial locations within an array) have to be held in the limited-capacity store. We (J. Lacey, R. Brunner, J.S. Saults, and I) have preliminary evidence against a capacity-limited account of performance in a modification of the color-array task (Luck & Vogel 1997) in which it is the presence or absence of a color anywhere in the array that is tested.

In understanding the demands of another type of task, running memory span, it is important to realize that serial positions in a list are not defined on an absolute basis starting from the beginning of the list and numbered 1, 2, 3, and so on. If they were, then inter-list confusions would occur between the same numerical positions of lists differing in length. Instead, relative position seems more important. For example, the last item of a j-item list is most likely to be confused with the last item of a k-item list, not with the j^{th} item of the k-item list (Henson 1999). Until the relative position of an item is known, it may not be possible to execute an efficient grouping process. If the length of the list is known in advance then the relative position is known; but such is not the case in a running memory span task (Pollack et al. 1959). It may be for that reason that running span results in memory of about 4 items per trial, not 7. Items cannot be grouped efficiently as they can in a regular span task because they cannot be assigned to relative serial positions until the list has ended.

R2.2.3. Chunking and scene coherence. The principle of scene coherence (sect. 2.6) adds a bit of complexity to the notion of capacity limits in terms of chunks, but it may be a necessary complexity. Its most basic contribution is that it shows how the 4-chunk limit could be compatible with the subjective impression (and behavioral finding) that only one channel of information can be processed attentively at one time (Broadbent 1958). As noted above, the idea is that one can hold in mind about 4 chunks only if they can be integrated into a coherent scene; only one coherent scene can be held at one time. It is interesting because it is the oppo-

site of a similarity principle. One might propose that two chunks will tend not to reside in the focus of attention at the same time if they are too similar and therefore easily confused. However, another possibility is that two chunks will tend not to reside in the focus of attention at the same time if they are too dissimilar and therefore difficult to meld into a coherent scene. This is a completely open question that has not been empirically studied, to my knowledge. **Beaman** suggests that the work on rapidly recycled sequences of four sounds (e.g., Warren & Obusek 1972) shows this effect inasmuch as memory for the serial order is much better when the four sounds are similar to one another than when they are dissimilar. However, this difference can be attributed to perceptual stream segregation of dissimilar sounds (Bregman & Campbell 1971; Bregman & Dannenbring 1973) and it remains to be seen if there is also a "conceptual stream segregation" that comes into play even in conditions in which the perceptual factors do not operate. For example, I can think of no work indication whether an array consisting of a color, a tone, a tactile sensation, and a printed or spoken word could be held in mind for comparison with a second, identical or slightly different stimulus array.

Schneider et al. referred to Luck and Vogel's (1997) finding that four features of a visual object can be held in working memory as easily as one and suggested that there may be a limit of four; that, say, six features per object could not be held in this way. I personally would not expect such a limit, provided that the features are truly integrable. I would not expect it because I would think that the binding of features into objects occurs in parallel once the appropriate stimulus location has been attended. If there were such a limit it would suggest that the possible dimensions of difference themselves have to take up space in the limited-capacity store, in which case the storage limit might have to be revised upward. It seems difficult to think of relevant evidence in the literature to date.

R2.3. Critiques of alternative hypotheses regarding capacity

R2.3.1. The 1-chunk hypotheses. Commentators **Baars** and **McElree & Dosher** placed their stock in the opinion matching the common subjective impression, that one can hold in mind only one chunk at a time (see also **Taatgen**). These investigators do not deny that some faculty of the mind is limited to about 4 chunks, but they consider that to be a separate organizing principle for activated information outside of awareness. A variety of evidence was adduced in favor of this alternative, 1-chunk hypothesis, which deserves a detailed discussion.

These commentators do not mention the concept of *scene coherence* (sects. 2.6 and R2.2.3), which logically could have made them feel better about the 4-chunk hypothesis. Given scene coherence, the subjective impression of concentrating on only one thing at a time may be real but deceptive. The global broadcasting function of consciousness (Baars 1988) ensures that all items that are present simultaneously in awareness are linked together and, by dint of that, are experienced as a single chunk. However, this linkage is new and can be accomplished only through the momentary collection of these items all at once in awareness. The true capacity limit can be observed in that the new chunk that is formed can be constructed out of no more than about 4 previously existing chunks; not, say, 7 or more (at least, not without a reiterative use of attention over time to build up larger chunks). In the four-chunk view, the single channel held in mind comprises a small collection of chunks that are newly linked together (see sect. 2.2 in the target article). The force of this thesis will be explained below in relation to a number of ostensible counterexamples to the four-chunk hypothesis that commentators brought up.

Baars appealed to *the case of ambiguous figures,* evidence of the inability to hold in mind two alternative organizations of a stimulus at once. Examples include ambiguous figures and binocular rivalry. However, these examples should be disqualified because the alternative organizations of the stimulus field are logically inconsistent with one another. This inconsistency should be viewed as a form of specific interference. Consider, for example, the Necker Cube, which can be perceived in either of two orientations (but not both simultaneously). If the reason for the limit were that capacity is limited to one chunk then it should be possible to perceive only one Necker Cube in one orientation or, at maximum, a field of Necker Cubes all in the same orientation (presumable forming a chunk). In contrast to this suggestion, though, I was able to demonstrate to my own satisfaction that it is perfectly possible to draw two Necker Cubes side by side and to see them in different orientations at the same time. If this is correct, the inability to hold contradictory forms in mind all at once should not be viewed as a capacity limitation.

Another common objection to the 4-chunk hypothesis, brought up by **Baars**, is the case of *multiple channels of stimulation.* People can attend to only one channel of stimulation at once (e.g., not both channels in a dichotic listening task). That certainly seems true if the channels are complex messages to be understood. However, each channel then contains multiple propositions that must be processed and integrated. It is therefore important to discuss task demands in a cautious way so as to be sure not to confuse a channel with a chunk and, as well, not to confuse a perceptual limitation with a working memory limitation.

Darwin et al. (1972) presented characters in three different spoken channels at once and the task required that all three be processed. Recall that the whole-report limit in this experiment was about 4, suggesting that more than one chunk of information could be held in mind at once. It could be argued that there are other factors to consider such as output interference, which will be discussed below. Nevertheless, this experimental result with relatively simple channels illustrates that the results of studies with more complex channels are difficult to analyze theoretically for capacity limits inasmuch as we do not know how many chunks are present in each channel.

Commentators **McElree & Dosher** brought up *the case of retrieval dynamics.* They presented evidence from very interesting probed recall tasks in which the speed criterion was varied so that they could plot the increase in accuracy of recall as a function of the speed of recall. In an unstructured list, this retrieval dynamic shows the fastest retrieval for the most recent item and no difference between the remaining list positions. In a list composed of strings of three consecutive items from a single semantic category, with a category shift between strings of three, the retrieval dynamic was fastest for the entire last category (which was assumed to be represented as a single chunk in memory). A

category cue was capable of conferring this advantage on a non-final category. On the basis of these findings they proposed that one chunk is in the focus of attention in these tasks, resulting in the fast retrieval dynamic.

There are at least three problems with the interpretation of this evidence. First, assuming that the most recently presented or cued chunk is in awareness, the advantage seen for this chunk may not indicate that it is the only chunk in awareness. For example, in a list of six unrelated items (presumable separate chunks), perhaps the last item is consistently in awareness and the rest of the available capacity is used up with a rehearsal set that includes a subset of items from anywhere else in the list. Sometimes it would include early items, sometimes it would include medial items, and sometimes it would include penultimate items. If this were the case, one could well obtain exactly the pattern that was observed: a retrieval speed advantage for the final serial position compared to all other serial positions, which need not differ from one another.

Second, it may be that subjects do not always fill their capacity, sometimes choosing instead to include only one chunk in the focus of attention even though they are capable of including more (though perhaps capacity can be filled only at a cost to the strategic processing that can be done).

Third, and finally, there is a potential problem in the designation of a chunk. Above, chunks were described as sets of items with strong intra-chunk associations and weak inter-chunk associations. Presenting three items from the same category does not automatically make them into a chunk, although it does produce some degree of associations between them. For example, if the related items occupy serial positions j, $j + 1$, and $j + 2$, it is possible that within the retrieval context, item $j - 1$ (not in the categorically related set) is more closely related to item j than item $j + 2$ is, as a result of the adjacency of $j - 1$ and j in the list. If one rejects the notion that the categorically related items form a chunk, the alternative account is that the associations between items are enough to induce subjects to hold all three of the items from a category in mind independently. One critical prediction of this alternative account would be that one could not get the enhanced retrieval dynamic with, say, seven items in a row from one category followed by seven from another category because the items from a category then could not be kept in the focus of attention concurrently.

It is not altogether clear how performance should be affected if the three items from the most recent category were held in that focus of attention as a single chunk. One would think that this would provide strong enough information to distinguish members of a category that were presented, from other similar members that were not presented and therefore were not part of the chunk. Yet, for the most recently-presented category, false alarms to non-presented members of the most recent category (e.g., receiving *pig, dog,* and *mouse* and false-alarming to *cat*) occur relatively frequently (McElree 1998, Fig. 4). Another possible prediction is that, if the three recent members of a category really do form a single chunk in the focus of attention, there should not only be an *advantage* arising from their presence within that focus; there also should be a *disadvantage* arising from the need to unpack the chunk in order to access its individual members. This could be tested in a procedure in which the number of consecutive items from a category is either 1 or 3. According to the single-chunk hypothesis, a single ungrouped item in list-final position should be retrieved more quickly than any of the items within a three-item list-final category because the isolated item avoids an unpacking stage. Failure to find this would seem to indicate that items are held in memory as separate (though associated) chunks, not as one chunk.

Last, there is an alternative interpretation (mentioned by McElree 1998) in which the advantaged items are not necessarily in the focus of attention but are simply more in keeping with the contextual cues of retrieval. If this is the case, or if the true capacity limit is greater than one chunk, it should be possible to achieve a faster retrieval dynamic for more than one category at once. If such a result were apparently obtained, though, one would have to look at the distribution of responses to ensure that subjects did not simply choose one category to keep in the focus of attention on each trial; this presents a thorny problem to be addressed in future research.

McElree & Dosher also mentioned the *simultaneous versus sequential array distinction.* They suggested that the four-chunk hypothesis may be valid only when the chunks are presented simultaneously (e.g., Luck & Vogel 1997). Apparently, they dismissed the sequential evidence, such as the memory-for-unattended-speech procedure of Cowan et al. (1999). It is not clear why they would consider it invalid but one could argue that there is output interference from each item being recalled in that procedure, reducing the ability to recall subsequent items. Indeed, that is a complaint lodged by **Milner** and by **Nairne & Neath**. An unpublished experiment (by N. Cowan, J.S. Saults, E.M. Elliott, and L.D. Nugent) allays that particular concern. Lists of nine digits were presented in an unattended channel and only occasionally were tested, in order to minimize the role of attention. In this particular experiment, recall was to begin at any of the nine serial positions and was to continue to the end of the list and then starting back at the beginning, in a circular manner, until all nine items were recalled. One can look at the proportion correct for any of the nine serial positions for the first-recalled items, for which there was no output interference. The results examined in this way were bow-shaped with a relatively weak primacy effect and a stronger recency effect. Summing the proportion correct for first-recalled items across serial positions produced an estimate of the number correct in the absence of output interference. This sum came to 3.2 items, consistent with the 4-chunk hypothesis as elaborated above. There at least does not appear to be any strong evidence against the hypothesis that attention focused on auditory sensory memory can result in the apprehension of roughly 4 chunks of information. Thus, this study suggests that one can extract 4 chunks from a sequential presentation. In principle, moreover, it should be possible to use one's faculties of mental imagery to recast sequential arrays as simultaneous arrays.

In partial defense of the suggestion by **McElree & Dosher** that the 4-chunk limit could apply only to simultaneous presentations, though, one possibility is that *attended* sequential presentations do not typically result in the loading of the focus of attention with four consecutive chunks of information. Instead, perhaps rehearsal processes (of an elaborative nature?) use up some of the capacity. One thing that favors this possibility is that, as pointed out by **Nairne & Neath**, there is a study that contradicts the finding of Halford et al. (1988), discussed above, indicating that

proactive interference does not occur for a memory test in which the target list is only four words long. Tehan and Humphreys (1996) did find such proactive interference. There are numerous differences between the studies that need to be explored but one potentially critical difference is that Halford et al. presented the words within each list as a simultaneous array, whereas Tehan and Humphreys presented each list sequentially. It may be that, because of this difference, 4 chunks entered the focus of attention at once in the Halford et al. study but not in the Tehan and Humphreys study. (However, the situation may be different in free recall, where proactive interference is not found for the recency portion of the serial position curve; see Craik & Birtwhistle 1971.)

Another comparison of hypotheses has to do with the *capacity and function of the focus of attention* and with how comparisons of chunks are carried out. For example, in a probed recognition task like the one that **McElree & Dosher** often use, it is necessary to compare the various targets to the probe until a match is found or none can be found. If only one chunk can be held in the focus of attention at a time, that implies that the probe and target cannot both be held in the focus of attention at the same time. Therefore, the comparison process must take place outside of that focus. It is difficult to understand why a process should be strategically controlled and yet take place outside of attention, and why the probe itself should not be considered part of the focus of attention at the same time as at least one target item. Similarly, the global workspace idea mentioned by **Baars** seems useful largely because it provides a mechanism whereby independent chunks can come together within the focus of attention to form a new chunk that reflects the present episode (a theory that also provides a teleological argument for the between-object links hypothesized by **Davis** and consistent with the episodic buffer of **Baddeley)**. If multiple chunks cannot be present in the focus of attention long enough for them to be linked together, it is not clear how or where in processing this chunking process can be carried out or why newly formed chunks are limited to a combination of about 4 pre-existing chunks (see sect. 2.2 in the target article).

McKone reviewed results that appear to be at odds with those of McElree and Dosher. Recognition judgments were to be made for items presented at various points within a long list; the reaction times were faster for the most recent 4 items, in keeping with a prediction made in the target article. However, this experiment also showed that the reaction times were shorter for stimuli presented more recently within the last 4 lags. This suggests that the more recent items were fresher in mind, not equally accessible within a focus of attention. The implications of this finding will be discussed further in section R2.3.5. However, for now it is worth noting that more work must be done to reconcile these findings. McKone's experiment did not examine retrieval dynamics (changing accuracy as a function of response time criterion), so it cannot be directly compared to the test situation described by McElree and Dosher; but the two experimental procedures may well differ in how much rehearsal they allow. In McElree and Dosher's situation, rehearsal may displace some specific items (e.g., the three penultimate items) from the focus of attention.

Data basically similar to **McKone**'s can be seen in Atkinson and Shiffrin (1968, Fig. 24). Lag judgments were to be made for stimuli presented in a long, running list, with test lags varying from 1 item to 16 items. For this sort of list, it is impossible to assign items to specific serial positions, so grouping and rehearsal become difficult or impossible. The results show facilitated (fairly accurate) lag judgments for lags of 1 through 4, with an asymptotic average judgment of about 5–7 items applied indiscriminately to items with an actual lag of 5–16. It is as if the capacity-limited processing system is capable of noting only 1, 2, 3, 4, and many.

It is going to take additional work before we will understand exactly what testing procedures best index information in the focus of attention. Oberauer (in press) has devised a modified probe reaction time testing method that seems very promising for distinguishing between information that is in the focus of attention versus information that is activated, yet outside of the attentional focus (Cowan 1988; 1995; 1999). Oberauer presented two lists of items (digits or words in different experiments) concurrently on each trial, in different colors. The set sizes of both lists varied. A postlist cue indicated which list would be tested. After a variable interval, a recognition test probe followed the cue. The measure of the irrelevant list remaining *activated* in memory was an intrusion effect: specifically, the slowing effect that occurred when a negative probe was an item from the irrelevant list rather than a nonpresented item. However, the measure of the irrelevant list remaining in the *focus of attention* was modification of the intrusion effect by the set size of the irrelevant set. The intrusion effect lasted a long time and was larger for older adults, whereas the irrelevant set size effect occurred only for several hundred milliseconds, presumably until the irrelevant list could be removed from the focus of attention, and did not differ with age. In another experiment, in which subjects (young adults) knew that both lists had to be recalled, the results reassuringly suggested that both lists remained in the focus of attention. This procedure could be used to estimate the contents of the focus of attention in a more principled manner than has been done previously.

R2.3.2. The 4-chunk perceptual/1-chunk memory hypothesis. Commentator **Jou** posed in interesting criticism. If the memory search procedure of Sternberg (1966) and the enumeration procedure (e.g., Mandler & Shebo 1982) both reflect the focus of attention or window of simultaneous consciousness (Ebbinghaus, cited in Slamecka 1985), then why do they show such different patterns of results? Specifically, memory search tasks yield a linear increase in reaction times as a function of the set size up through about 6 items and then a shallow slope, whereas enumeration tasks yield a shallow slope for 1–4 items and a steep slope after that. Jou's suggested explanation appears to be that, in perception, the simultaneous presence of several stimuli results in parallel processing of the items in the focus of attention; whereas, in memory search, items enter the focus of attention only one at a time in a serial search process.

There are several points that must be made in response. First, it seems unlikely that memory search occurs in a straightforward, serial fashion. It has often been found that search times are faster for items at the end of the list than for items earlier in the list (Clifton & Cruse 1977; Diener 1990; Dosher 1981; Monsell 1978; Ratcliff 1978), which appears to rule out a serial search of items in the presented serial order. Although it is possible that memory search occurs in some random or backward order, it seems more

plausible, and there is better evidence, that memory search occurs for list items in parallel, in a capacity-limited manner (Ashby et al. 1993; Ratcliff 1978; Townsend 1976), with recent items often having an advantage within the competitive process because of their greater salience in memory.

Second, the difference between reaction time patterns is not the result of a contrast between perceptual tasks on one hand and memory tasks on the other. In visual search tasks with a varied target set from trial to trial, which are perceptual tasks, one sees a reaction time that increases steadily with set size, not the two-part pattern seen in subitizing (Dosher 1998; Schneider & Shiffrin 1977).

Third, one must take into consideration differences in task demands. In enumeration, the entire pattern of objects is relevant to the count. In memory search, in contrast, what is needed is not a conglomeration but a discrimination among items. In one way or another, the probe must be compared to each item in memory at least until a match is found. For any memory set size greater than one, the total collection in the focus of attention will not match the probe. There is no reason to think that the process of determining whether one of the items within the focus of attention matches the probe should be automatic, so it will depend on the set size.

Fourth, it is easily understandable from a fixed-capacity view that the supra-capacity slope differs markedly from one task to another. Specifically, whether the reaction time function at large set sizes becomes steeper (as in enumeration tasks) or flattens (as in memory search) clearly must depend on what process takes over at that point. In the case of enumeration, it is a serial counting process, whereas in the case of memory search, it is a parallel direct access process that is not capacity limited, yet is slower than the capacity-limited process that one finds at smaller set sizes. So this difference between findings is not a problem for the 4-chunk hypothesis.

Despite these points, it must be acknowledged that a central riddle remains unanswered. Why should the reaction time in memory search increase linearly as the set size increases to 6 items if only 4 can be held in the focus of attention? That is a good question. Nevertheless, it is worth pointing out that subjects in memory search tasks are allowed to rehearse the stimuli. The process of rehearsal may keep up to 6 or so items in such an active state that the response times remain faster than is found through capacity-unlimited memory access. According to this suggestion, it would be predicted that the inability to rehearse during presentation of the memory list should shift the point of discontinuity leftward, so that the linear increasing portion of the function should extend through set sizes of 4 but no more.

This still does not really address the question of why reaction times increase in a *linear* fashion through 6 items in memory search. That point has been a problem, though, for all theories of memory search. We know that various types of serial or parallel capacity-limited processes can produce linear functions but we do not know why the functions happen to be linear despite nonlinear serial position effects in reaction time. I am confident that it would be possible to put forth a theory of the linear function that assumes a core attentional capacity of 4 chunks but, given how little is understood here, it would be post hoc and unhelpful to do so. Instead, I would point to this as an exciting area for further empirical research in the near future.

R2.3.3. The 7-chunk hypothesis. Two commentators (**Bachelder** and **Pascual-Leone**) found value in the 7-chunk hypothesis, although presumably for very different reasons.

Bachelder showed that points from several types of experiments (memory span, subitizing or span of apprehension, and absolute judgment) could be plotted on a similar, inverse-ogive function (as in his figure). He suggested that the "magical number 4" emerges from a perfect-performance criterion on this function, whereas Miller's "magical number 7" emerges from a criterion at which 50% of the trials result in correct responses. (At least, I assume that is what the curve shown in Bachelder's figure represents.) Although this function is empirically beautiful, I have some doubts about its interpretation, for several reasons. First, correct performance on a memory span trial requires success at all serial positions. In contrast, the span of apprehension and absolute judgment tasks required just one response per trial.

More important, there is no way that the curve shown in **Bachelder's** figure can represent all of the evidence that I take to reflect capacity limits. Consider, for example, the 35 adults who carried out the memory-for-unattended-speech task of Cowan et al. (1999). These subjects completed attended-speech as well as unattended-speech tasks. The results corresponding to Bachelder's plot (proportion of trials correct at each list length) are shown in Figure R1. The bold solid line is a portion of Bachelder's normal ogive. The plain solid lines with solid points represent data from the attended-speech procedure. Each subject received trials at only 4 list lengths, equal to the longest list that was recalled correctly in a pretest (defined here as "span") and three shorter list lengths (span-1, span-2, and span-3). To obtain a wider range of list lengths, data are shown separately for subjects with span = 6 (N = 5), span = 7 (N = 14), span = 8 (N = 11), and span = 9 (N = 5). The ogive approximates these functions fairly well; it seems nearly perfect for subjects with a span of 6. However, look now at the dashed lines, representing data from the unattended-speech procedure. In this procedure, scores are considerably lower. In fact, for these data, it is approximately the case that the 50%

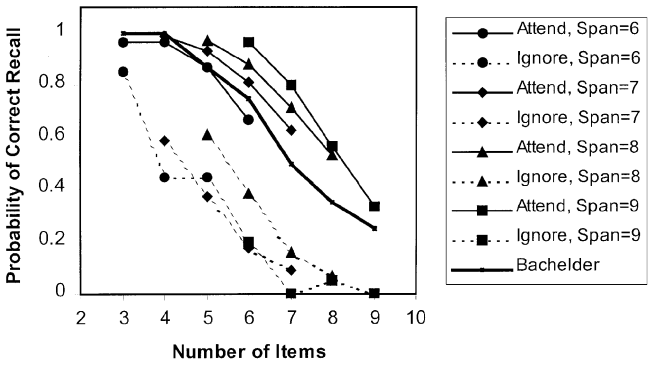


Figure R1. Proportion of lists correctly recalled by the 35 adults who carried out the task of Cowan et al. (1999); these are the same subjects shown, for example, in Figure 3 of the target article. Here they are reported separately for subjects with each memory span (defined as the longest list repeated correctly in a pretest). *Thin solid lines*, attended-speech condition; *dashed lines*, ignored-speech condition. Only the attended-speech condition matches the normal ogive (*bold solid line*). Based on data by **Bachelder.**

correct criterion occurs for 4 items, not 7. This may reflect the fact that the estimated capacity is 4 chunks or higher only in about half of the adults, as shown in Figure 3.

A devil's advocate could point out that one measure taken to indicate a pure STM capacity limit, the subitizing (span of apprehension) limit of Mandler and Shebo (1982), fits **Bachelder's** function, whereas another measure, the unattended speech procedure of Cowan et al. (1999), does not, suggesting that they cannot represent the same concept. I would argue, though, that this discrepancy can be explained within the fixed-capacity view through a theoretically-oriented task analysis. Memory span and absolute judgment, which I consider to be compound STM estimates, could both require rehearsal of a stimulus series (the list items for span and the response choices for absolute judgment). In subitizing, however, all items are considered at the same time and only one response is to be emitted. Loss of some precision of information may or may not throw off the single response, whereas loss of precision in a span situation is likely to result in one or more incorrect item response, making the entire list response count as incorrect for the type of data shown in Bachelder's figure. Thus, the empirical functions shown by Bachelder do not consistently indicate the pure capacity limit.

There is, nevertheless, a problem in understanding subitizing according to the 4-chunk hypothesis. If only 4 items are apprehended as separate chunks and the other items in an array are not apprehended, shouldn't performance be approximately a step function of items, with the step occurring at the capacity limit? At least two alternative explanations of subitizing have been introduced. As mentioned by **Taatgen**, Peterson and Simon (2000) suggested that rapid enumeration of small numbers of objects comes from the recognition of specific patterns that the objects happen to form, and showed that the number of patterns to be learned begins to skyrocket for numbers above 4. **Avons et al.** Suggested that numbers of objects are judged in a continuous manner and that the just-noticeable difference (jnd) between two amounts depends on the Weber fraction. (**Milner** made a similar suggestion.) The idea would be that for a 4-item array, the Weber fraction is small enough that the true number plus or minus the jnd is still nearer to 4 than to the neighboring numbers; whereas for arrays of 5 items or more, the jnd is large enough for the number to be mistaken for a neighboring number. This is an interesting possibility, although it is unclear if the magnitude of differences could account for the discontinuity observed between 4 and 5 items. For example, according to an application of Weber's law, suppose that the just-noticeable difference for 4 objects was .44, so that the perceived number could not reach any lower than 3.56 or any higher than 4.44. In that case, the predicted jnd for 5 objects would have to be (5/4) × (.44), or .55. With some random error thrown in, it is unclear whether such a subtle change in the jnd could truly account for the discontinuous performance function. One prediction of the model, however, is that the point of discontinuity should increase markedly if the task requirement is relaxed (e.g., not asking for an exact count but asking whether there were <X objects or not). At present, I am far from certain that the limit in subitizing occurs for the same reason as the other capacity limits reviewed in the target article. This does not, however, provide any support for the 7-chunk hypothesis or weaken the 4-chunk hypothesis much, generally. (It also is possible that there are logical constraints that apply similarly to both STM and LTM. For example, the explosion in the number of ways in which one can arrange n objects as n exceeds 4 could add a difficulty to both STM and LTM mechanisms even if they operate separately.)

Pascual-Leone was more explicit in arguing for a 7-chunk hypothesis. The basis of the argument was that the analysis in the target article was incomplete and did not take into account the memory capacity needed for the procedures involved in carrying out the task. However, this analysis would be invalid if, in fact, the procedures have become sufficiently automatized so that they do not compete with limited-capacity storage. A case in point is Pascual-Leone's analysis of the unattended-speech task of Cowan et al. (1999). According to that analysis, 3 or 4 slots in working memory are used up to carry out the rhyme-matching primary task, and a remaining 3 or 4 are used to retain in memory digits from the unattended channel, which are recognized following an orienting response. This analysis contrasts with the analysis of Cowan et al., who assumed that the digits were held automatically in a sensory memory and that attention (and, therefore, capacity-limited storage) was focused on the speech only after the rhyming task was replaced by the auditory-memory recall cue. One problem with Pascual-Leone's analysis is that Cowan et al. found no effect of the concurrent memory task on performance in the rhyme-matching task. The same was true of Cowan et al. (2000), who used the same task but manipulated the memory retention interval rather than the list length. Also, if working memory slots were used up retaining some of the supposedly unattended sounds, one might expect the capacity to remain fairly constant across a silent retention interval. This would not be expected if performance depended on a sensory memory that was lost during the retention interval. A reanalysis of adult data from Cowan et al. (2000) shows that the capacity (and its SD) was 3.34 (1.53) after a 1-sec retention interval, 2.27 (0.85) after a 5-sec retention interval, and 1.90 (0.73) after a 10-sec retention interval. This is in contrast to performance in the attended-speech condition, which showed no memory loss across the retention intervals. Thus, the evidence is more consistent with the interpretation of Cowan et al. (1999), involving a 3- to 4-chunk limit drawing upon a sensory memory representation of the last spoken list at the time of the recall cue.

It also would be difficult for **Pascual-Leone**'s model to interpret one finding of Luck and Vogel (1997) in their color-array memory task: that performance was unaffected by a concurrent, 2-item, verbal memory load.

How, then, would we account for the evidence upon which **Pascual-Leone**'s M-space theory is based? One possibility to consider is that the tasks he used allowed a contribution of rehearsal and chunking strategies to improve performance, providing compound STM estimates rather than pure STM capacity estimates.

Kareev (2000) offered another argument in favor of a magical number 7 ± 2. The argument was based on considerations about correlations between binary variables. Small samples tend to overestimate the correlation found in the entire population. Given that a smaller working memory provides an individual with smaller samples, a limit on working memory could be useful for detecting weak correlations (see sect. 4.1.1). Kareev (2000) argued that the usefulness index was greatest for samples of around 7 pairs of values. Although this might be viewed as embarrassing,

given that the target article used the same phenomenon to argue for a capacity limit of 4 rather than 7, I stand by my estimate, on the basis of all of the evidence taken together.

The arguments based on correlations are lax enough to allow either estimate. An argument demonstrating the plausibility of 4 emerges from a rudimentary simulation I have carried out with slightly different assumptions than Kareev used. I created a random list of 1,000 numbers ($x1$, $x2$, . . . $x1000$) with the value of 1 or 2. I then created another list of 1,000 numbers ($y1$, $y2$, . . . $y1000$) with the value of 1 or 2 so that the correlation between the two lists equaled .4. A third list was constructed, so that its correlation with the first list was .6. For each pair of correlated lists (1–2 and 1–3), a moving window of 2, 4, 6, or 8 number pairs was used to calculate sample correlations. If there was no variation in the sample obtained from at least one of the lists, no correlation was calculated. A particular value of the sample correlation was taken as the minimum criterion (C) indicating that the subject would "notice" the correlation, and the results were examined with various criteria. The question then was, which number-pair sample set size (2, 4, 6, or 8) produced the largest number of "noticed" correlations in the same direction as the population correlation? The number of noticeable correlations was corrected by subtracting from the number of correlations with $r \geq C$ the number that were noticeable, but in the wrong direction ($r \leq -C$), to produce an adjusted success value. The outcome was similar for a population correlation of either .4 or .6. With C set to .9, set size 4 was clearly the winner. (With a population correlation of .4, the set sizes of 2, 4, 6, and 8 produced adjusted success values of 198, 212, 114, and 57, respectively. With a population correlation of .6, the same set size produced adjusted success values of 290, 352, 263, and 182, respectively.) Set size 4 also consistently came out ahead with C set to .6, whereas set size 6 won with C set to .7 or .8, and set size 8 won with a C set to .5. Given that we do not know the appropriate detectable sample correlation level C for human subjects, this exercise yielded no clear-cut winner. It suggests that in our present ignorance, one should not take Kareev (2000) as strong teleological evidence either for or against a basic capacity limit of 4 chunks.

Wilding presented protocols of mnemonists who could recall long strings of memorized digits. Reaction times showed breaks after sets of about 4 digits but, in an apparent departure from the 4-chunk hypothesis, the last series could be as long as about 7 digits. One might suspect that a decay-based mechanism held the last string of 7 or so digits until they were available for entry into the focus of attention. However, according to the theoretical view suggested in the target article, it is mysterious how the focus of attention could collect and recall the last 7 in a single cluster. Close inspection of Wilding's Figure 1 suggests that there actually may have been slight pauses in the middle of the last group of 7. They could be difficult to detect if the retrieval process for the last chunk could take place largely during the recall of the next-to-last chunk. This could occur at the end of the list because the overall structure of the list no longer had to be held in mind. Another possibility is that the last two chunks of 3 or 4 numbers were melded into a single super-chunk in the period after the list was presented and before recall.

R2.3.4. The unlimited-capacity hypothesis. A completely different view from the present target article was taken by **Ericsson & Kirk**. They stated that "Even in task environments where the functional independence of chunks is convincingly demonstrated, individuals can increase the storage of independent chunks with deliberate practice – well above the magical number four." One may well question, however, what the effects of expertise are and what was meant by the functional independence of chunks. For example, in the memory performance of a trained individual who can repeat lists of many digits in a row (Ericsson et al. 1980), there may be a hierarchical organization in which chunks are composed of other chunks. It is the higher-level chunk that should govern recall ability (e.g., features are chunked into letters but capacity is not measured at the letter level when letters are in turn grouped into words).

Ericsson & Kirk noted two encoding mechanisms by which information can be stored beyond any specific capacity limit. They were said to be able to "generate associations between different presented chunks of information and build new integrated structures in LTM." It was noted that "If experts can encode associative relations between virtually all chunks within their domain of expertise, then the concept of chunk independence would not apply. Second, and more important, skilled individuals acquire skills to associate chunks of presented information with internal cues organized in retrieval structures." I do not deny any of this. However, if experts are able to do this they may be able to represent an entire stimulus field as a single chunk that can be unpacked (with the help of long-term memory) into smaller chunks as needed. Therefore, the ability of experts to overcome the conventional capacity limits can be viewed as an extraordinary instance of the capacity limit but not an exception.

An important complication for future research is that some associations may be intermediate in strength. So, for example, if one were asked to recite the alphabet and then the American pledge of allegiance to the flag, one would pause briefly within the alphabet because some groups of letters form a subgroup; and one might pause during the pledge because a certain phrase is not immediately accessible. Nevertheless, if we can learn to understand effects of intermediate-level associations and of grouping hierarchies used in multiple reiterations, it is reasonable to apply a capacity model. Basically, all one has to do in the aforementioned situation is to hold in mind a node for the alphabet and a node for the pledge of allegiance, while unpacking and recalling subgroups of elements one by one. The more expertise is involved, the more theoretical work investigators must do to understand the associative structure before a capacity principle can be applied. This, however, does not rule out the principle. Ruling out the existence of capacity limits on the basis of the phenomenon of expertise would be much like ruling out the existence of controlled processing (Shiffrin & Schneider 1977) on the basis of the phenomenon of automatic processing, or more broadly, like ruling out effects of gravity on the basis of effects of wind resistance.

Schubert & Frensch made an interesting suggestion along this line. They suggested that "in dual-task studies a capacity limitation of the central mechanism can be observed only when subjects carry out the tasks in a relatively unlearned state." They go on to suggest that "training leads to an over-learned mapping of stimuli and responses, and, consequently, to an automatic activation of the response when a stimulus is presented." In other words, expertise

uses a type of automatic processing that can obscure the effects of capacity limitation within controlled processing. The task for a capacity account would be to determine what little controlled processing is left to be accomplished by the expert in a capacity-limited manner. For example, a chess expert may be able to go around a room playing multiple chess games at once, but it seems doubtful that chess expertise would allow an individual to play one game with the left hand and a different game with the right hand, without some loss of speed or playing quality. I, thus, adhere to the principle that expertise is likely to alter the way in which a limited capacity can be filled, but without doubting the capacity itself in terms of chunks (or a more sophisticated, graded, hierarchical associative structure once it is understood and studied).

One commentary (**Lane et al.**) reviewed work on complex tasks such as chess, in which, following Chase and Simon (1973), they have been able to obtain empirical evidence of chunks. They were able to implement a specific computational model of the process and found that a 3- or 4-chunk capacity best matched performance of players at different levels of expertise. Although this theoretical success is in only one research domain, it is a reassuring demonstration that such work can be done.

R2.3.5. Decay-and-interference-only hypothesis. Several commentators (**Davis; Grondin; McKone; Milner; Muter; Taatgen; Towse**) suggested that the possible influences and results of limiting factors other than chunk capacity (in particular, time-based forgetting and interference) were neglected or dismissed. For example, **Muter** suggested that "a comprehensive account of STM should surely include treatment of the nature of forgetting after attention has been diverted." I could not agree more and addressed those questions at length in several previous works (Cowan 1988; 1995). I have studied time-based forgetting intensively in the past (see for example Cowan 1984; 1988; Cowan et al. 1990; 1994; 1997; 2000; in press; Gomes et al. 1999; Keller & Cowan 1994; Keller et al. 1995; Saults & Cowan 1996). Time-based forgetting has been difficult to understand inasmuch as, after all these years, there still may be no clear evidence for forgetting as a function of the passing of a particular, absolute amount of time as opposed to the passage of time relative to the times of presentation of potentially interfering stimuli (e.g., see Cowan et al. 1997 vs. Cowan et al., in press). Nevertheless, time-based forgetting and interference from other items are both very important factors to consider in recall. The present target article is simply intended to cover only one theoretical mechanism of short-term memory, the capacity limit.

Towse noted that "Strong faith in measures of memory size alone may permit us to find The Answer, but at what cost to a full understanding of immediate memory?" The implication is that he believes that chunk-based capacity limits may eventually turn out to result indirectly from decay and interference factors. Although I believe that decay and interference are critical for remembering, their usefulness seems reduced when one is talking about the contribution of the focus of attention, the topic of the target article. What is crucial here is that the 4-chunk capacity seems to apply even in situations in which conditions have been selected so as to minimize time-based forgetting of information before it reaches the focus of attention and output interference that could conceivably displace it before it is recalled (e.g., the single-decision procedure of Luck and Vogel 1997; the unpublished paper on unattended speech discussed in sect. R2.3.1, in which just the first-recalled item was examined). It also is worth recalling evidence that information in the focus of attention may be impervious to some types of interference that can occur outside of that focus (Craik & Birtwhistle 1971; Halford et al. 1988, discussed in sect. 3.3.4). Therefore, I believe that the concepts of time-based forgetting and material-specific interference are more appropriate for describing passive storage buffers (Baddeley 1986) or activation outside of the focus of attention (Cowan 1988; 1995). To understand short-term memory, one must know about the capacity of attention (in chunks) as well as the passively held information available to it (affected by time and interference). The intent of the target article is to improve our understanding of short-term memory by analyzing a key component of it separately but we will then have to put the whole working memory system back together. (For attempts to do so see Cowan 1999; also see the commentary by **Baddeley** and see Baddeley, in press.) Additional discussion will be in response to the specific points of each commentator, who used the notions of decay ind interference in different ways.

Davis described an unpublished experiment in visual perception by Davis et al. that he thought ruled out a chunk-based capacity limit in favor of "decay and interference." There was not much actual emphasis on decay. The "interference" that he mentioned apparently referred to a limit in how many binding links could remain activated, between objects and between features within an object. That type of interference, if correct, could serve as a "subatomic" type of analysis of capacity limits that is essentially friendly to the notion of a 4-chunk limit, which he predicted on the basis of the number of links for ordinary objects, in which the total number of inter-object links increases rapidly with the number of objects in the display.

Although this reasoning is exciting, the conclusion that **Davis** described on the basis of an experiment seems open to question. In this experiment, subjects had to compare the shapes of two notches among 3 or 6 irregularly-shaped objects, as shown in Parts C and D of his figure. The number of objects (3 or 6) did not matter overall, but one case was an exception. In that case, horizontally displaced notches appeared on the same object in either situation, whereas vertically displaced notches appeared on the same object only in the case of 3 objects, not in the case of 6 objects. In this case, there was an advantage of fewer objects. This exception showed that the number of perceived objects in the displays really did differ as intended. Thus, the absence of a number-of-objects effect overall was seen as evidence against a capacity limit of 4 objects. The ordinary effect of number of objects was said to be an indirect effect of the number of inter- and intra-object links combined, which here were held constant. My concern with this logic is that the demands of the task are not clear. Not all visual searches are capacity-limited; simple feature searches are independent of the number of objects. If the two target notches in the display can be located in a capacity-unlimited search, then there is no reason to expect that the comparisons would be made more slowly with 6 objects present than with 3. Also, the more specific effect that was obtained could have occurred not because of a difference in the number of objects perceived in the two types of display but because of subjects' use of local symmetry as a cue. It may be especially

easy to compare two notches when they flank a large, simple object because, in that circumstance, the notches are the same if and only if there is local symmetry in the design. For the horizontally displaced notches, symmetry is always a useful cue because a rectangular edge is flanked by the notched (though perhaps this cue is not as useful in the case of 3 large objects because the horizontal symmetry of the semi-elliptical shapes could make it difficult to perceive local asymmetry). For vertically displaced notches, in the case of 3 large objects, the notches flank a white semi-elliptical shape; that design feature might be used to judge symmetry. For 6 small objects, in contrast, vertical displacement leaves no large shape for the notches to flank, which may make symmetry less useful. Davis's point is potentially very interesting but it seems that these possibly confounding factors first need to be examined further.

Grondin brought up an interesting temporal limit in cognition. The comparative judgment of time intervals benefits from rehearsal only, according to his findings, for intervals longer than about 1.2 sec. Presumably, this is an estimate of the "psychological present," which could be described as the period for which stimulus input seems to cohere into the same perceived event. (It may be worth noting one important nuance, that the "psychological moment," the time during which sequentially-presented stimuli are perceived as simultaneous, is an order of magnitude smaller; see, for example, Allport 1968.) I would not in any way dispute such an interval or its importance in limiting the psychological present. The stance taken in the target article was just that these factors could not explain storage limitations in the full range of situations in which they have been observed; for example, it is hard to see how they could explain the limitations observed by Luck and Vogel (1997). Nevertheless, I agree that there are some likely links between any such psychological present and the chunk-based capacity limit. For example, a rapid sequence of 4 stimuli presented within a 1.2-sec window would allow 300 msec per stimulus presentation, just long enough for brief stimuli to be perceived (Massaro 1972; Turvey 1973). Perhaps there is both a temporal and a chunk limit to the psychological present, leading to a rule such as this: *The psychological present can include up to 3–5 chunks of information at once, provided that these chunks all can be perceived, conceived from long-term memory, or reactivated through rehearsal within the period of one psychological present.* Such a rule might make many investigators happy (e.g., Baddeley 1986). However, it should be noted that the reason for the rule might not be decay of the first-presented chunks. Rather, chunks presented farther apart might be difficult to integrate into the same perceptual present (or coherent scene, in present terms). Given that this rule is stated in such abstract terms, considerable empirical work would be needed to assess its validity. It would have to be shown to apply with rehearsal processes blocked. Rehearsal may serve the purpose of *allowing the subject to re-present all of the items within one psychological present in order to encode them into a common mental context.*

Finally, **Grondin** found the postulation of a chunk limit (rather that a time limit) as central to short-*term* memory to be an oxymoron; a "term" is a time period. Perhaps so, but the continual need for the focus of attention in processing practically ensures that any particular set of chunks must leave to make room for other chunks within a rather short term. This explains the great difficulty of vigilance during a secondary task, for example, remembering to check the oven in a timely fashion while busy composing a journal article. Despite the importance of not burning down the house, the stove presumably cannot remain in working memory and must be reloaded periodically, most likely during breaks in the mental workload of the ongoing composition.

McKone presented results that were discussed briefly in section R2.3.1. and are relevant to the decay issue. Recognition times were found to be shortest to the most-recently-presented item and to increase steadily until a lag of 4, after which the response times stay much more constant. Given this pattern, she noted that it conflicts with the idea that the most recent 4 items are equally well represented in the focus of attention. There, she suggested that "older items fade from the limited capacity mechanism (at least without rehearsal)." What makes me doubt that interpretation is the pattern of results from our experiments on memory for unattended lists of spoken digits (e.g., Cowan et al. 1999; 2000b). These experiments show a distinct bow-shaped serial position function. One interpretation is that, at the time of the recall cue, the focus of attention casts its net upon the available activated memory (in this case, presumably a stream of auditory sensory memory) and picks up the 3 to 5 most salient chunks within the representation. The pre-list silent period makes the sensory information from the beginning of the list salient and the end-of-list silent period, as well as perhaps the highly activated state of recent sensory memory, makes that part of the list salient. In McKone's procedure, also, there is the possibility that the focus of attention is refilled with the appropriate items, from activated memory outside of the focus, only after the recall probe is presented. Thus, time-based forgetting is only one part of what differentiated the salience of different items. McKone's reaction time function could be accounted for if there were a probabilistic tendency to fill the available 4 or so slots with the most recent items, but if on some trials some of the slots were filled with items from other serial positions or not filled with items. The monotonically increasing advantage across the last 4 serial positions would occur because, on many trials, the slots were indeed filled with the last 4 items.

McKone also objected that the capacity was much lower than 4, specifically about 2, in the implicit memory of pseudowords. I would suggest here that under implicit memory situations, it may not be the case that each pseudoword was learned as a single chunk. Instead, many pseudowords may be recognized only as several separate chunks, as was suggested in the target article (sect. 3.2.1).

Taatgen presented an ACT-R model that combines a 1-chunk focus of attention with activation, decay, and interference of information from long-term memory. Except for the postulation of only a 1-chunk focus, this model is in keeping with that presented by Cowan (1999). On one hand, the output of the model (shown in Taatgen's Fig. 1) seems very promising. The low-capacity curve appears to match the data for the unattended speech condition shown in Figure R1, whereas the high-capacity curve appears to match the attended-speech condition and the curve presented by **Bachelder**. However, it is difficult to understand how the function based on "the effects of decay, interference, and the increased probability of doing something wrong if more responses are required" (**Taatgen**) can account for evidence with simultaneous arrays followed by only one deci-

sion (Luck & Vogel 1997) unless a metric is proposed for mutual interference between elements within a simultaneous display. Again, it also is difficult to understand how a 1-chunk focus of attention can accomplish comparisons between items. The success of the modeling is encouraging but we probably do not know how many different forms of model can accomplish the same thing. I agree with the implication that capacity, an integer quantity, ultimately is likely to result from developmental changes and individual differences in some underlying continuous quantity or quantities. However, I see no reason to exclude attention as the basis for that change or the locus of that capacity.

In sum, I agree with many of the commentators that decay (or any type of time-based forgetting) and interference are important factors to consider in short-term recall. However, the present thesis is that a 4-chunk limit of the information in the focus of attention at one time can be identified *apart from these other factors.* According to the model of Cowan (1995; 1999), decay and interference determine what information will be available and accessible to the focus of attention, from which about 4 chunks will be selected. While in the focus, items may be protected from further decay or interference. Evidence presented by **Rensink** from change-detection experiments is especially helpful in illustrating this point of the separability of capacity from other processing limitations. For example, when the interval between displays increases to 360 msec, eliminating a visual sensory memory contribution, a change can be identified reliably among at most 3–4 items, and this limit does not change with further increases in the interval.

R2.4. An improved procedure to estimate capacity in visual arrays

The majority of the criticisms of the 4-chunk hypothesis have been directed at the interpretation of verbal, serial recall measures. Therefore, it is important to communicate that, since the target article was written, I have developed (with co-investigators J.S. Saults and N.M Fristoe) an improved means to assess capacity in visual array experiments. Specifically, we have developed a formula to index memory capacity in the procedure of Luck and Vogel (1997) in which a single item within the second array is encircled and the task is to determine whether that item has changed color from its counterpart in the first array (the arrays being otherwise identical). The formula was a modification of one presented earlier by Pashler (1988). The logic of the measure is as follows.

Upon examining a briefly presented array of N items, the subject is able to apprehend a certain fixed number of items, k. The apprehension of these items would allow a change to be detected if one of these k items should happen to be the changed item. Thus, with probability k/N, the change is detected. If the change is not detected, the subject guesses "yes, there was a change" with probability g. Thus, the formula for the hit rate H is: $H = k/N + [(N - k)/N]g$. If there is no change between the two arrays, and if the cued item happens to be an item that is included within the set k that the subject apprehended, then that knowledge will allow the subject to answer correctly that no change has occurred (and this is where our formula differs from Pashler's). If there is no such knowledge (for $N - k$ items), then the subject still will answer correctly with a probability $1 - g$, where g is again the probability of guessing "yes." Given that memory is used to respond in the no-change situation, it is useful to define performance in terms of the rate of correct rejections, CR. The assumptions just stated then lead to the following expression: $CR = k/N + [(N - k)/N](1 - g)$. Combining equations, $H + CR = 2k/N + (N - k)/N = (k + N)/N$. Rearranging terms, the capacity can be estimates as $k=(H + CR - 1)N$. The resulting estimates appear to be compromised slightly by a performance ceiling in set size 4 but they are rather constant across larger set sizes, which we have tested up to 10. The resulting capacity estimate in several data sets is about 3.5 items.

To illustrate the working of the model, estimates can be obtained on the basis of the data figure provided by **Schneider et al.** for a set size of 6. Inasmuch as H+CR equals twice the proportion correct P, for large objects or small objects, assuming P = .80 (from the data graph), $k = (2(.80) - 1)\ 6 = 3.6$. For combined objects, if we assume that 6 small objects plus 6 large objects equal 12 objects total and P = .63 (from the data graph), then $k = (2(.63) - 1)12 = 3.12$. The lower number in the second case suggests that not only were large and small objects perceived separately, they also may have interfered with one another slightly. In sum, simultaneous arrays lead to quantifiable capacity-limit estimates, much as sequential presentations do.

R3. Cognitive theoretical accounts of the 4-chunk limit

R3.1. Metatheoretical considerations

As mentioned above, my strong belief in the 4-chunk limit is not accompanied by convictions that are as strong regarding the theoretical reason for the limit. Many of the reviewers offered their own theoretical approaches to understanding capacity limits. The theoretical problem we have as a field is certainly not the absence of a plausible explanation, but rather many different, and in some cases conflicting, explanations. Therefore, a first step toward the understanding of the capacity limit may be to set out a simple taxonomy of capacity-limit theories. In doing so I will assume that the 4-chunk limit applies. Inasmuch as I already discussed alternative theoretical approaches that do not lead to the prediction of a 4-chunk limit in STM, the following discussion will be limited to theories that are, at least to my understanding, compatible with a 4-chunk limit. The goal is to try to discover what types of experimentation will be most helpful in distinguishing between these possibilities.

After making certain core assumptions (from Cowan 1995; 1999), I find that the most general question that arises is about the mechanism that constrains capacity. A basic suggestion in the target article was that the limit is in the capacity of the focus of attention. Upon reflection, however, one might interpret that suggestion in two different ways. First, the *focus of attention might be able to hold only a limited number of items* from the activated representation concurrently. Second, the *activated representation might be limited* in some way that constrains how many items can be recovered by the attentional mechanism.

These two possibilities are illustrated in Figure R2 for processing of a single row of 7 letters in Sperling's (1960) whole-report task. Before explaining the figure, the underlying assumptions must be stated as they apply to this case. All 7 items are presumably contained briefly (for several hundred milliseconds) in a visual sensory store. This visual

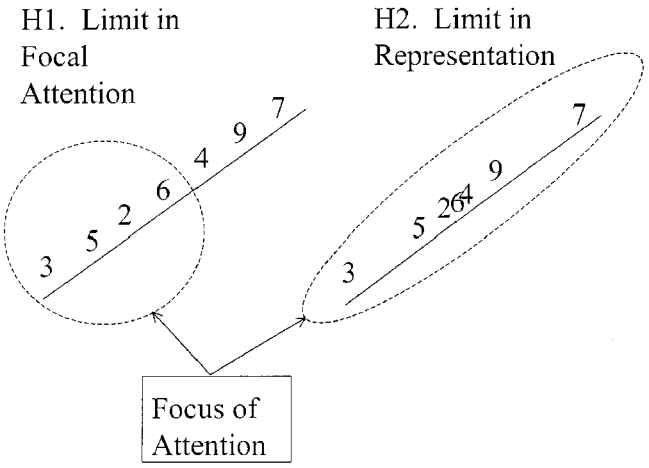


Figure R2. An illustration of two alternative, cognitive theoretical hypotheses as to why a capacity limit of about 4 chunks is obtained. The illustration is for a 7-item linear character array in a whole-report condition like that of Sperling (1960) but the depiction should be similar for any estimate of pure STM capacity limits.

sensory story might consist of a feature map that deteriorates unless it is enhanced by attention quickly. (The feature-map concept was included because most errors in the Sperling-type procedure are location errors rather than omissions or substitutions, as shown for example, by Irwin & Yeomans 1986; and by Mewhort et al. 1981.) Each item is assumed to be encoded as a separate object (or chunk) because the rapid presentation did not allow enough time for items to be grouped. Not all of this information is held in the focus of attention at one time. Items presumably must come into the focus of attention before they can be deliberately recalled. In the example, the activated representation comes from sensory memory but, presumably, long-term memory retrieval also produces an activated representation. The main difference is that the latter tends to have fewer sensory features and more semantic features than activated memory originating in sensory memory.

Given these assumptions, two loci for the limit in capacity then can be articulated. In the first hypothesis (H1, shown on the left of Fig. R2), the focus of attention, depicted by the dashed circle, is limited to four items within the encoded set. In the second (H2, on the right), the focus of attention (dashed ellipse) can process all of the items but their representations interfere with one another. This particular example shows one plausible deficiency, a bow-shaped recall function; items surrounding the center of the display are too indistinct from one another within the memory representation for them all to be recalled. Variants of both of these alternative views may be viable, given the available evidence.

Let us refer to these hypotheses shown in Figure R2 as (H1) the limited-focus view, and (H2) the limited-representation view. The figure does not capture all theoretical views but it can serve as an anchor point from which other views also can be described: for example (H3). It is possible that there are limits in *both* the focus of attention and the representation upon which attention is focused. For example, one could propose that the focus of attention can include only *one coherent scene* at a time and also that the scene can include no more than 4 chunks because of limitations in the representation. One could also propose (H4) that attention is *not a unitary concept* and that, therefore, limits apply separately in different domains. This type of proposal appears to be a variation on H1 in which different attentional limits operate at the same time in different domains. There are other possible accounts, but these are adequate for the present exposition.

Some commentators' alternative views (sect. R3.2) will be classified with respect to these four categories along with further specifications as appropriate. The theoretical views include separate-stores accounts (R3.2.1), unitary memory accounts (R3.2.2), multiple-attention accounts (R3.2.3), and storage-plus-processing (neo-Piagetian) accounts (R3.2.4). Biologically oriented explanations will be presented and some will be classified according to these categories later on in section R4. Last, a tentative appraisal of the models will be offered in section R5.

R3.2. Alternative theoretical views

R3.2.1. A separate-stores account. Whereas I suggested that the focus of attention works with an activated portions of long-term memory, **Baddeley** proposed that the focus of attention works with a number of passive stores. He suggested that my characterization of his model contains some misinterpretations and pointed out both similarities and differences between our views. A comparison of our approaches could be helpful.

At the time that I wrote the section on his view of working memory, I had not yet heard of his addition of an episodic buffer (and did not realize until now that he considered it to be of limited capacity in terms of chunks). With that addition, I agree that his view and mine are rather similar despite different uses of terminology.

In the target article I used the word "arbitrary" to describe the stores in Baddeley's model, as he noted; but that term was misleading. I meant to imply only that the stores are not exhaustive; some types of coding do not appear to fit in well. The contribution of the episodic buffer may change this. Baddeley (in press) described this buffer briefly. The main reason for proposing it appears to be that activated elements of long-term memory do not include the novel combinations of these elements that occur in new experiences, which clearly must be preserved if new episodic information is to be remembered (in both the short and the long term, although the episodic buffer is for the short term). With this addition, the view is either a limited-representation view (H2) or, with another limit in attention thrown in, a version of H3.

Although Baddeley (in press) contrasted this point to my own view, I actually agree with him. Thus, Cowan (1999, p. 89) stated,

> Finally, there is one important qualification of the statement that working memory contains activated elements of long-term memory. Most stimulus situations in life include novel combinations of familiar features. In memory the elements are activated independently, but the particular links between those elements are often novel. The current combination of elements may, however, be stored as new long-term memory trace.

Similarly, Cowan (1995, p. 101) previously stated,

> The information that is in an activated state must include new information as well as the activated set of previously learned information. Links between items and/or between each item and its serial position context must be generated and preserved as additional activated information. The new links comprise an episodic record that will become part of long-term memory.

A remaining, subtle difference between views is that I have envisaged a mechanism different from **Baddeley's** to accommodate this information. I have imagined that the newly combined information immediately forms new paths in LTM, while at first being in an activated state (and often, or always, at first in the focus of attention) and then subsiding out of activation. The newly combined information includes, for example, the binding of elements to serial positions in the current list. In contrast, Baddeley sees the need for a separate episodic buffer to hold the newly formed combinations temporarily.

Several questions about the episodic buffer seem to remain unanswered at this point. What passive storage system would hold, say, imagery for musical timbre (Crowder 1989), which is neither phonological nor spatial in nature? If the episodic buffer holds such a code, does the code then have a status comparable to phonological and spatial codes or is the episodic buffer different (e.g., less passive in nature)? Presumably, one can deposit a musical timbre in memory either passively, through a sensory stimulus, or actively, through imagination. A potential concern is that the system, even as amended, seems to suggest that some codes (phonological and spatial) are represented in a more fundamental or enfranchised manner than others (such as acoustic quality or, say, tactile quality) and there may not be sufficient empirical support for that particular assumption.

My statement about the focus of attention that "no other mental faculties are capacity limited" was followed immediately by a qualifier: "although some are limited by time and susceptibility to interference." Thus, I use the term "capacity" in a narrow sense. I agree that this set of processing limits is probably similar to Baddeley's (1986) model, except that chunk limits were not emphasized within that earlier description of his model.

Baddeley pointed out three apparent differences between our models, to which I will respond.

1. I was said to regard short-term memory as "simply the currently activated areas of long-term memory," which was said to be inconsistent with neuropsychological data inasmuch as "deficits in long-term memory are found which have no apparent impact on immediate memory, or on working memory more generally." I cannot claim to be an expert on the neuropsychological data. I would caution, though, that what appear to be separate short-term stores could be, instead, separate routes to the activation or use of information in long-term memory, so that it would be specific retrieval processes, rather than stores, that are impaired or left intact in some patients. Processes versus stores can be notoriously difficult to distinguish.

2. My limited capacity system was said to be unitary, in contrast to the "fractionable" central executive of Baddeley's model. Actually, I distinguish between the control aspects of attention and the focus of attention (e.g., Cowan 1995), in a manner that seems similar to Baddeley's conception. His central executive control processes appear similar to my central executive or control of attention (heavily involving the frontal lobes), whereas his episodic buffer probably is more closely aligned to what I call the capacity-limited focus of attention (heavily involving the parietal lobes and adjoining temporal areas, according to suggestions made by Cowan 1995). For the present target article I saw no reason to emphasize control processes. Further fractionation of the central executive may be possible, as Baddeley believes, but I simply have not dealt with any such further fractionation.

3. As mentioned, my focus-of-attention mechanism was said to have storage capacity, whereas in Baddeley's recent conception that capacity resides in the episodic buffer. An important difference between the predictions of these views appears to be that momentary distraction would cause items to be lost from a focus of attention (though the information would stay in an activated form for a while and therefore could be returned to the focus of attention without much difficulty). In contrast, distraction would not be expected to have as severe an effect on an episodic buffer. Clearly, the effect of distraction on a limited-capacity system is a fruitful question for future research and may help to resolve this issue.

R3.2.2. Unitary-memory accounts. At least four commentaries (**Nairne & Neath; Avons et al.; Rensink; Taatgen**) questioned, in one way or another, a fundamental assumption underlying the present thesis that there is a 4-chunk capacity: the assumption that there exist separate short- and long-term memory faculties. This assumption is necessary if one wishes to hold the belief that long-term memory is unlimited in capacity, whereas short-term memory is limited. Instead, the theoretical account would presumably draw on general principles of memory, such as contextual cues in retrieval and the distinctiveness of memory representations within a retrieval context. As such, it appears to be a version of H2, a limited-representation view.

Although this type of account seems reasonable, it seems theoretically implausible that there is not *some* sort of STM/LTM distinction. I define STM rather broadly (as the focus of attention along with other activated elements of LTM; see Cowan 1988; 1995; 1999). It seems to me that something of this nature must exist. First, I think that no one truly believes humans can attend to an unlimited amount of information at once, and that the concept of an attentional limit is practically self-evident. Second, the concept of memory activation probably maps onto the distinction between information held through synaptic connections (long-term storage) and the subset of that information that is currently in an electrochemically active state (short-term storage). My view of short-term storage does not require that a separate copy of the information be formed in a different faculty within the brain, in contrast to a view offered of the basis of neuropsychological considerations by **Baddeley**; though this view was apparently attributed to me by **Taatgen**, probably because of my use of the term STM to refer to this activation, against the wise counsel of **Morra**. Even though most cognitive researchers may accept the idea of a limited attention and memory activation of some sort, I suppose that many of them do not see these as scientifically helpful concepts. I would argue, however, that one cannot apply the general principles without at least an implicit use of the attention concept to describe capacity limitations during encoding and retrieval, even if the 4-chunk limit turns out to be a result of limitations in the focused-upon memory representation.

The unitary-memory theorists presumably want to use general mechanisms of memory to account for the 4-chunk limit. It is worth noting that they have various options open to them. I can identify basically three: *overwriting, perturbation,* and *contextual distinctiveness.* I will try to describe the potential relevance and boundaries of each of these.

In an overwriting model (e.g., Nairne 1990; Neath & Nairne 1995) each stimulus is assumed to be represented as a bundle of features, and features of each stimulus are said to overwrite the same features that happened to occur within previous stimuli in the sequence. (By the way, I would think that long-term memory is preserved intact, whereas the overwriting takes place in activated memory, in which case this type of model would seem to be misclassified and actually is a dual-store model; but its users typically see it as a unitary model.) **Murray** presented an account of this type. Overwriting models of this sort seem possible for the description of memory for sequential stimuli but not simultaneous arrays. It is not clear to me, in this type of model, what overwriting is supposed to take place among simultaneously presented stimuli as in the Luck and Vogel (1997) task in a manner that depends on the number of stimuli in the array. Thus, someone with this view might well be forced to suggest that capacity limits in simultaneous arrays arise from a different mechanism than memory limits in sequential presentations. Then it is not clear if the limit of about 4 chunks in each realm would be judged the result of a coincidence or some deeper design principle. In any case, if one accepts that attention must be focused on the representation in order for recall to take place, it is a version of H2 or H3. (The same probably could be said for perturbation and contextual distinctiveness; all of them focus on limitations in the representation rather than in attention per se.)

In a perturbation model (e.g., Lee & Estes 1981; Nairne 1991), the items do not obliterate one another in memory as in the overwriting model. Instead, representations of items are said to switch serial positions within the memory representation of a list, a process termed perturbation. Usually, it appears to be nearby positions that switch. One could also generalize this model to a spatial array with the assumption that adjacent spatial positions switch within the representation. This type of model widens the types of representation that can be considered. Whereas an overwriting model may be inherently committed to some sort of temporary representation because an overwritten trace is destroyed, one can talk of perturbations taking place in a representation without it being destroyed. Nevertheless, if one holds that long-term memory representations are not destroyed, it is a bit mystifying how they can be so altered. Perhaps the perturbations take place at the time of retrieval. If so, multiple retrievals from the same list might tend to show a perturbation during one retrieval that is unperturbed in a later retrieval. It is not clear to me how good the evidence for a perturbation mechanism is because guessing factors in serial recall can produce patterns of results that look like perturbation patterns (Anderson & Matessa 1997). On the other hand, the target article brought up the possibility that items that had been encoded in the same attention cohort tend to be confused with one another (sect. 3.4.3) and that could be the mechanism of perturbation (which would seem to place perturbation in the camp of H1 rather than H2 or H3, where it otherwise would fall).

Finally, in a contextual distinctiveness model, the sets of features established during encoding are not necessarily degraded but there is a failure to distinguish some items (or chunks) from others within the representation at the time of retrieval, and therefore an inability to recall all of the chunks appropriately. This confusion among items can depend on the similarity between the context of encoding and the context at the time of retrieval. One type of contextual cue is a timing signal from the time of encoding (Brown et al. 2000). Contextual distinctiveness is the principle underlying the ratio rule that has been used to explain the recency effect in recall: the principle that the ability to distinguish two nearby items correctly depends on the ratio of the time between the items and the time between the last item and the recall cue. It could account for the 4-chunk limit in sequential recall on the basis of diminishing distinctiveness of items earlier in the list. An analogy could be drawn so that spatial distinctiveness could govern the ability to tell apart objects within the visual field. That presumably is the gist of the finding of He et al. (1997) mentioned by **Scholl & Xu**, showing that with great enough spatial distinctiveness one can enumerate numbers above 4 in a visual afterimage; although I have no access to the research report. The theory of subitizing mentioned by **Avons et al.**, based on the application of the Weber fraction for a continuous perception of numbers, may be an example of the distinctiveness principle but based on a perceptual representation (and sometimes its afterimage) rather than a longer-lasting memory representation. It involves the comparison of the perceived quantity with a continuum of quantities encoded in long term memory. The alternative account of Peterson and Simon, mentioned by **Taatgen**, can be described similarly except that it is a pattern rather than an amount that is compared with the contents of long-term memory.

There are other principles of memory but they may not apply directly to the question of the reason for the chunk limit in memory. For example, memory can be improved with long-term recoding in order to form associations between items and between these items and existing structures in memory. However, by definition, this type of process alters the size of chunks rather limiting how many chunks can be recalled, and so is not relevant here. Now, let us turn to some of the evidence provided by commentators taking a unitary-memory view.

Nairne & Neath presented a provocative result. Words presented in lists for pleasantness ratings were presented again in alphabetical order, the task being reconstruction of order in the long term. Lists of an average of 3.75 items could be ordered correctly 50% of the time. Their conclusion was, reasonably enough, that a limited capacity does not indicate a separate STM mechanism.

This experiment tells us a great deal about the mechanisms of working memory, which will be addressed in later sections. However, it does not pose a problem for the present theoretical view. As Shiffrin (1973) noted: in a sense, every deliberate task is a short-term memory task. The memory representation must be brought to consciousness before it can be overtly recalled. Thus, it is perfectly possible for the representation of a list to be retrieved from long-term memory into an activated state so that the most distinctive aspects of the representation (limited to about 4 at a time) can be drawn into the focus of attention and overtly recalled.

If there are separate STM and LTM faculties, the STM faculty could affect LTM recall not only at the time of retrieval as Shiffrin noted, but also at the time of encoding. With the latter case apparently in mind, Broadbent (1971) addressed the STM/LTM relationship as follows (pp. 342–43):

> There remain to be considered two points urged by interference theory: the existence of effects on short-term memory from previous long-term experiences, and the continuity which

> seems to exist between memory at long and short periods of time. The first of these must be admitted straight away, and is perfectly consistent with a view of short-term memory as due to recirculation into and out of a decaying buffer storage . . . In general one must beware of concluding that the appearance in short-term memory of an effect known from longer-term studies is evidence for identity of the two situations . . . Only the success or failure of attempts to show differences between the two situations is of interest in distinguishing the theories.

Pothos & Juola reinforced the logical case introduced by Broadbent in the previous quote, that STM tasks and LTM tasks would not be expected to be process-pure. He noted that "If language is at least partly learned, linguistic dependency structure should reflect properties of the cognitive components mediating learning; one such component is STM." He showed higher contingencies between words within a range of 4 words in natural language, with remarkable convergence among samples from different languages.

Like **Nairne & Neath**, another commentary, **Avons et al.,** similarly argued for a continuity of STM and LTM. It is precisely this sort of thrust that led me to comment in the target article that the existence of memory decay (i.e., time limitations in certain kinds of mental representations) still has not been established after all these years. However, do Avons et al. actually doubt that the focus of attention can encompass only a small amount of information, in contrast to the vast amount of information in memory? I believe that this dual process, which was central to the target article, should be counted as a discontinuity between STM and LTM.

Taatgen suggested that "short-term memory capacity is not something that can be used to explain the outcomes of experiments, but is rather something that needs to be explained itself." Why shouldn't the arrow of causality go in both directions? Just as gravity helps to explain many phenomena (e.g., our ability to stay on earth) and a theory of space/time attempts to explain gravity, STM capacity may help to explain many phenomena (e.g., problem-solving proficiency) and yet itself needs to be explained at a finer-grained level of analysis.

To sum up, unitary memory accounts argue against a separate short-term storage facility. However, it is unclear where they stand on the existence of a limited-capacity attentional mechanism or on some definitions of memory activation. (Clearly, even in these unitary views, there is a special accessible status for information in memory that has not been overwritten and that matches currently available retrieval cues.) In a dual-storage view, a continuity between short- and long-term memory procedures is nevertheless to be expected, on the basis of the involvement of these attentional and activation factors both at the time of long-term learning and at the time of retrieval (Broadbent 1971; Shiffrin 1993).

R3.2.3. Multiple-attention accounts. Several commentators (**McKone; Rensink; Scholl & Xu; Woodman et al.**) argued that the attention-based account is not suitable because attention is not a unitary phenomenon. These are apparently versions of H4 above.

Woodman et al. discussed how visual search occurs on one time scale and working-memory search on another time scale, and how visual search can take place without impairment when working memory is loaded. They also mentioned research on the separability of response limits. I believe that their arguments overlook some important theoretical treatments of the phenomena mentioned. There was no discussion of the distinction between controlled and automatic processing in visual search and memory search (Schneider & Shiffrin 1977; Shiffrin & Schneider 1977). The experimental procedure in which search was investigated during a memory load (Woodman et al., under review) was one in which the search targets were the same from trial to trial and therefore was presumably an automatic search. It would be expected that, in contrast, a controlled visual search *would* be influenced by working memory load. There also was no mention of the alternative views on the reason why a response bottleneck often is obtained (Cowan 1995; Meyer & Kieras 1997).

Woodman et al. are on to something but not, I believe, something relevant to the 4-chunk hypothesis. It seems reasonable to believe that mechanisms contributing to automatic visual search and to response bottlenecks occur, respectively, before and after a central attentional process that is linked to subjective awareness. These input- and output-related processes still may be termed attention if one prefers, but their limitations may not be appropriately defined in terms of a capacity limit. They may be limited instead by other factors, such as the time of processing and the presence of domain-specific or feature-specific interference. Therefore, I do not think that they alter the arguments of the target article, although they clearly contribute a great deal to our overall understanding of information processing.

Scholl & Xu argued more directly that the phenomena discussed in the target article dealing with visual information processing come from a different source than the phenomena dealing with verbal information processing. For example, they said that a general STM-based theory capacity limits "could not easily account for the strong dependency of MOT (multi-object tracking) performance on subtle visual details such as the type of accretion and deletion behind occluders." I do not understand why this is the case. The findings appear to indicate that tracking performance is preserved with a logical disappearance and reappearance of objects (disappearance behind occluding objects or even virtual objects) but is impaired when objects inexplicably disappear and reappear. These studies point to a fascinating perceptual apparatus that influences the input to a central capacity-limited process. If one object disappears and reappears somewhere else inexplicably, it does not pick up the same object identity. However, the capacity limit still could reside in a central attentional faculty fed by this perceptual process, not in a separate visual attention mechanism per se. The ability to "track 4 targets in the MOT task and simultaneously acquire and hold 4 verbally-presented items in STM" was used as an argument against a central capacity limit inasmuch as attention-switching is impossible in this procedure. However, in adults, an articulatory loop (Baddeley 1986) would be able to hold 4 items almost effortlessly, leaving the attentional mechanism free for tracking. If verbal memory were loaded to capacity or the procedure were used in children old enough to rehearse, but not effortlessly (Guttentag 1984), I would predict more interference between tasks.

The other arguments also seem questionable. Arguments based on the subitizing procedure are open to question given the uncertainly as to the cause of that phenomenon, acknowledged above (sects. R2.3.3 and R3.2.2), although it may well be that common principles underlie subitizing and

memory limits (much as Newton's inverse-square law applies not only to gravity, but also to other phenomena such as light intensity). If subitizing limits come from limits in long-term recognition, it stands to reason that uncrowded sets of objects could be more likely to be recognized than crowded sets. Arguments based on neurological dissociations are open to question because a deficit in one area (e.g., verbal STM) with a normal performance in another area (e.g., subitizing) could be explained on the basis of a defect in a perceptual or representational process feeding into the central attentional faculty, rather than a fractionation of the attentional faculty per se. The points raised by Scholl and Xu suggest some important possibilities for future research to resolve these issues.

Some points make by **McKone** and **Rensink** can be thought of similarly. **McKone** noted that "Clear theoretical reasons exist for limits on the number of items that can be simultaneously active within a single domain, since leftover activation from previous items will interfere with identification of the currently-presented item." Here one must distinguish between similarity-dependent interference, on one hand, and a true capacity limit, on the other hand. Both may well occur. Theoretically at least, the difference would be that a capacity limit depends on the number of chunks in memory at the same time, rather than specific similarities in the features of those chunks; whereas interference depends on similarity rather than number per se. **Rensink** pointed to evidence for "at least 10 items" probably "resulting from the grouping of items of similar contrast sign." No doubt such a visual mechanism (one that initially breaks the world into parts) occurs, but it is presumed to be a preattentive mechanism that feeds into the capacity-limited process. This type of distinction is reminiscent of the distinction (Norman & Bobrow 1975) between data-limited processes, such as specific processes in visual perception, and resource-limited processing, such as central attention, a distinction that seems very important. Again, I maintain that data-limited processes are affected by decay and interference factors but are not truly capacity-limited, unlike resource-limited processes. Nevertheless, I am sympathetic to Rensink's conclusion that "Given that there are no compelling a priori grounds which can be appealed to, this matter will have to be settled by experiment."

R3.2.4. Storage-plus-processing accounts. Three of the commentaries are from a neo-Piagetian perspective (**Halford et al.; Morra; Pascual-Leone**). What all of these accounts have in common is that they suggest that processes take up capacity. The processing accounts all incorporate the need for storage in one way or another but that is a basis of controversy. The accounts nevertheless all appear to be varieties of H1, suggesting that a central attentional limit explains the processing and storage limitations. **Halford et al.** focus on processes, suggesting that "humans are limited to relating four entities," (see also **Davis**), whereas **Morra** and **Pascual-Leone** consider storage and processing separately and suggest that the total capacity is therefore 7 chunks (some taken up by processing schemes), not 4. Ironically, the latter two theorists (like **Raffone et al.**) appear to fall into my camp in considering storage and processing to share a common resource, as opposed to the suggestion by Halford et al. that processing limits cannot be "subsumed under storage limits." Yet, it is Halford et al. who come up with a capacity estimate in keeping with my result.

The reason that I still do not agree with the limit of 7 is that the views leading to that estimate by **Morra** and **Pascual-Leone** do not appear to allow that processes can become automatized and thereafter cease to take up capacity. The research combining processing with storage loads (reviewed, for example, by Baddeley 1986) does not seem to suggest that there is a strong tradeoff between the two. In many studies, a moderate storage load proves to have minimal effects on concurrent processing. A full storage load has notable effects on processing but, in that situation, one can argue that the central executive has become involved in special processes (various types of complex, attention-demanding rehearsal) in order to manage the full storage load. Given the disagreement even among the neo-Piagetians, it is clear that more empirical work is needed to determine the relation between processing and storage. I believe that the procedures used in the target article to indicate a storage limit of 4 chunks are procedures in which everything except storage tends to become automatized and takes up little capacity.

R4. Biologically oriented accounts of the 4-chunk limit

The present section complements the previous one by examining commentaries describing theory and evidence relevant to the neurobiological underpinnings of the capacity limit. These include psychophysiological evidence (sect. R4.1), defined broadly to include electrophysiology and neuroimaging; evidence from individual and group differences in humans (sect. R4.2); and evidence from animal behavior (sect. R4.3), which is relevant to evolutionary considerations. Following these topics I will discuss several neurobiological models of capacity that were offered by commentators (sect. R4.4) and will seek clarification of the most useful relations between theoretical models and empirical evidence, and between neurobiology and behavior, in understanding capacity limit (sect. R4.5).

R4.1. Psychophysiological evidence

The evidence that was presented offers some important opportunities for further research in the near future. **Rypma & Gabrieli** showed that the brain response to one or three items is far less than the brain response to six items. This finding may offer the opportunity to use the neuroimaging to help solve the difficult problem of how to define chunks, discussed in section R2.2. For example, if one presents a sequence such as "F-B-I-C-I-A," will the brain metabolism look like two simple items (assuming that the subjects form two higher-order chunks) or like six? If it looks like six, could it be made to look like two by presenting preexposures of the acronyms? If the method does not match behavioral results, such a discrepancy could suggest that there are hidden processes that need to be added to the cognitive model. This type of research could help to answer **Tiitinen's** plea for research to address the chunking issue.

Gratton et al. considered a distributed capacity hypothesis in which a limit of about 4 chunks would result from a limit of about 2 in each hemisphere. They showed interesting evidence that reaction times were about the same when five items were distributed to the two hemispheres 2–3 or 1–4, but not when all five items were presented to one hemi-

sphere. It would seem, though, that a distributed capacity model should predict a smoother transition between the two conditions. The result that was obtained may be more consistent with a model in which groupings play a role. Groups of 1 to 4 items can be chunked together. Perhaps two such chunks can be accessed and searched in parallel; quickly, given that they are small sets. When all 5 items occur in one location, it may be more difficult to separate them into subgroups, in which case a slower search of set size 5 has to be used. This hypothesis could be investigated by dividing the stimuli into two groups above and below fixation, rather than to the left and right of fixation. The distributed memory model should predict no difference between distribution conditions, whereas a grouping account predicts a difference comparable to what was obtained in Gratton et al.'s Figure 1. This, too, is a promising area for research.

Jensen & Lisman brought up several new physiological findings related to the hypothesis of Lisman and Idiart (1995) that nested cycles of oscillation in the brain can account for the capacity limit. These, too, show promise although, sooner or later, a tighter linkage between physiological and behavioral results will be needed (as discussed in sect. R4.5). For example, how can the physiological evidence for changes in oscillation rates in the brain be reconciled, if necessary, with behavioral evidence favoring a parallel rather than serial memory-search process, as discussed in section R2.3.3? It is helpful that several investigators have begun to find behavioral evidence of changes in stimulus presentation rates that can be closely linked to the physiological data on the special role of the 40-Hz rhythm in humans; not only Burle and Bonnet (2000), as brought up by Jensen and Lisman, but also Elliott and Müller (1998; 2000). Clearly, though, it will be a while before we know what is happening in these innovative new procedures.

R4.2. Evidence from individual and group differences in humans

As discussed in the target article (e.g., sect. 3.1.3), studies of individual and group differences can contribute to our understanding of the mechanisms underlying capacity limits. Insofar as the individual and group differences in capacity can be attributed to biological sources, they have the potential to yield information about relevant biological processes. **Oades & Jemel** considered the possibility that the capacity limit may be different in individuals with schizophrenia. It is a good point that pathological populations are a likely source of information about the reason for capacity limits. Oades and Jemel brought up an interesting question. Given that schizophrenics are renowned for loose associations, why do they not chunk items more than normal individuals and therefore show a supernormal STM? I think that the answer must be that the loose associations are nonselective and are based on characteristics of the stimuli that are typically irrelevant for performance. Perhaps if those typically irrelevant associations were make relevant, schizophrenics would excel. Oades and Jemel stated that schizophrenics excel on global processing. They might excel also if tested on, say, incidental occurrences of rhymes among stimuli to be memorized on a semantic, rather than a phonological, basis. In this case, their loose associations could result in the recoding of relevant stimuli into fewer chunks. Pure capacity limits are an important area in which to look for subnormality in schizophrenics given that they do not differ in some of the temporal aspects of STM behavior. In particular, they are abnormal in the precision of auditory encoding, but not in auditory recognition speed or auditory memory duration when the overall performance levels are equated with normal individuals (Javitt et al. 1997; March 1999).

It is not only subnormal individuals who are potentially instructive. As discussed in section R2.3.3, **Wilding** illustrated that capacity limits of about 4 items emerge in the form of the number of items clustered together, within special subjects whose memory for digits is supernormal. The pattern is similar to individuals who have trained themselves to have a supernormal digit span, discussed by **Ericsson & Kirk**. In both cases, there are anomalies that the authors would want to attribute to some mechanism other than a capacity limited to 4 chunks of information, though in both cases the predominant finding is consistent with the 4-chunk hypothesis. If capacity limits are eventually to be understood well, these anomalies in the results must be understood; but it still seems to me that the vast preponderance of evidence is favorable to the 4-chunk hypothesis. We all would agree that these individuals' unusual abilities cannot be attributed to a larger-than-normal memory capacity, and therefore that supernormal performance depends more on training than on biology. Yet, it remains to be determined if any normal individual could learn to accomplish these mnemonic feats. Given that pure STM capacity measures have not been applied, we cannot rule out the possibility that there is a minimum level of capacity that an individual must have before being capable of becoming a mnemonist. It must be difficult to distinguish between raw ability and motivation, with some individuals falling by the way instead of becoming experts because of insufficient motivation or inadequate ability.

Hecht & Shackelford remarked upon the possible role of pure STM capacity limits for applied areas of cognitive development, including mathematics. Of course, this general approach is consistent with the neo-Piagetian approaches of **Halford et al.**, **Morra**, and **Pascual-Leone**. The assumption is that biological changes mediate a growth of capacity developmentally. Hecht and Shackelford appealed to Geary's (1993; 1995) distinction between biologically primary skills, which come relatively easily in development (e.g., natural language learning; math in the subitizing range), and secondary skills which are much newer in the evolution of the human species and prove difficult for many children (e.g., learning to read; math in the counting range). According to Geary's theory, the secondary skills can be learned through a "co-opting" process in which primary skills are used to bootstrap the secondary ones. Hecht and Shackelford wrote that "A limited pure short-term memory capacity, shaped by evolutionary forces, may currently be 'co-opted' for many contemporary tasks such as biologically secondary mathematical problem solving skills." If this is the case, measures of pure STM capacity could be important in understanding individual differences in applied areas. One reason why these measures could serve a different purpose from more complex measures of working memory (e.g., Daneman & Carpenter 1980; Daneman & Merikle 1996) is that the latter can incorporate many processes (item storage, sentence or task information processing, and probably rehearsal) and therefore do not make it clear which basic abilities underlie the correlation of

working memory with practical skills. Measures of pure STM capacity might be questioned on similar grounds but are at least more narrowly focused than complex measures of working memory capacity.

R4.3. Evidence from animal behavior

Several interesting biological questions can be addressed through comparisons of animals to humans. First, in a pure STM capacity measure, if capacity is linked to uniquely human abilities, there should be a substantial advantage for humans. The absence of any advantage for humans would indicate that capacity is biologically more fundamental. Second, in a compound STM measure, if performance depends on skills unique to humans, there should again be a substantial advantage for humans. (It is reasonable to assume, for example, that verbal rehearsal strategies are unique to humans and assist their performance.) The absence of any advantage for humans would suggest that any strategies unique to humans are not assisting performance.

Kawai & Matsuzawa described the interesting research from their recent paper in *Nature*, implementing a type of memory span task in a numerically trained chimpanzee, in which digits appeared in a random array about the computer screen and were to be selected by the chimp in numerical order, the digits' identities being covered by masks on the screen after the first response was made. A reasonable argument was made that the chimps remembered 5 digits 65% of the time. (The first digit in the sequence of 5 could have been identified visually, whereas the remaining digits were masked; but the response lag was relatively long before the first digit and was shorter between digits, suggesting that the first digit was part of the memorized set.) Five adult human comparison subjects were run. Their proportions correct were not reported in the article but their mean accuracy in the same condition as the chimpanzee was 97% with 4 numbers, 92% with 5 numbers, and 83% with 6 numbers (Kawai, personal communication, August, 2000). This suggests that adult human spans are about 2 items higher than the chimpanzee.

Arguments could be found for this task being either a pure STM capacity measure or a compound STM measure. I suspect the latter. The stimuli appeared in a simultaneous spatial array on the computer screen, as in many pure STM capacity measures. In this case, however, the presentation time was not short and was determined by the subject. The stimuli lent themselves to coding as a coherent series in a particular spatial configurations, and therefore might not have been too taxing to bring from sensory memory into a categorically coded form. When humans participated in a second condition in which the numbers were masked after a fixed 750 msec period from the beginning of each trial, their scores were much more similar to those of the chimpanzee (Kawai, personal communication, August, 2000). Attention to the stimuli at the time of their presentation apparently can assist performance, not only in humans through verbal rehearsal but also through a mechanism available to chimpanzees. (This should be clear also from the benefit of attention during presentation in first-grade children who should be a bit young for strategic rehearsal [Cowan et al. 1999].) Thus, the study does not provide an estimate of the chimpanzee's pure-capacity limit.

One might speculate that if the pure limit in chimpanzees is about half of the compound estimate, as is typical in humans, chimpanzees might have a pure capacity of about 2.5 chunks on the average. That would make them similar to the first-grade children studied by Cowan et al. (1999) in an unattended-speech procedure, consistent with the claim of Kawai and Matsuzawa (2000) that their chimpanzee's performance is "the same as (or even more than) preschool children." Alternatively, it is possible that adult chimpanzees have a pure capacity limit of about 4 chunks, similar to humans, and that the absence of sophisticated strategies means that there is less of a difference between pure and compound limits than there is in humans. Consequently, this study raises important questions that can only be answered in future research.

Todt presented evidence that songbirds repeat songs in packages of about 4 chirps, except when the songs are spaced out over time, in which case performance is reduced below that chunk size. It is unclear if the distinction between pure and compound capacity estimates is a meaningful one in songbirds or not. If one can assume that this is a pure capacity estimate in the songbirds, it suggests that the basic capacity limit is a fundamental property of efficient brains and not a signature of higher human intellect. It would then be a faculty that is, in the terminology of Geary (1993; 1995), "co-opted" for human intelligence; useful for many things including holding information in mind while processing it in a sophisticated manner. The similarity between evidence on the burst size in songbirds and the cluster size in human recall (e.g., sects. 3.4.1 and 3.4.2) is striking. The mystery is only how it could come about that estimates of capacity in human children (e.g., Cowan et al. 1999) fall below estimates in songbirds. That could be the result of an inappropriate comparison between different methods. Alternatively, it could be an authentic phenomenon, analogous to the observation that motor development is in many ways more advanced in adult animals than in immature humans.

R4.4. Neurophysiological models of capacity

Three commentaries briefly described explicit neurophysiological models that can predict capacity limits of about 4 chunks (**Jensen & Lisman; Raffone et al.; Usher et al.**). Probably the most important lesson here is that there are multiple ways to accomplish the same thing. I must admit straight off that I do not profess expertise in the workings of these models, severely limiting how much I can say about them.

Jensen & Lisman elaborated upon the type of oscillatory model that was described by Lisman and Idiart (1995) and discussed in the target article (sect. 4.1.2). Rather than adhering strictly to Miller's magical number, they state on the basis of recent empirical evidence constraining their modeling parameters, that the upper limit of STM is about 5 to 6 items. My main question regarding this model is how absolutely it depends upon the memory search process being serial in nature. The model is of obvious heuristic value but ultimately it will be important to see how it could, for example, produce recency advantages within search, and how plausible the additional necessary assumptions would be.

Raffone et al. presented a model that appears to be less wedded to the notion of a serial search process. In their model, "neural assemblies in high-level visual areas, coding unrelated features or objects, exert mutual inhibitory or desynchronizing actions." By stating that "competition and

desynchronizing actions between neural assemblies coding for unrelated features are stronger within than between specific representational domains," they appeared to be admitting both representation-specific and general types of limitations (Hypothesis H3 above). One important questions about this model is what the role of attention may be (for the general limitations). The conclusion stated in the commentary was that "only four 'magical' neural assemblies can dominate the brain at any given time, saving other domain-specific active neural representations through selective synchronization." Within this scheme, what happens to unattended information that is active in memory? My own view (e.g., Cowan 1999) has been that the limit of four neural assemblies applies only to attended information and that, outside of attention, numerous other features can remain active in memory (although not fully "assembled" like the attended chunks). It is not clear if this model fully agrees with my view or not and, if not, whose conception is correct.

Last, **Usher et al.** similarly presented a mode designed to overcome shortcomings of the Lisman and Idiart (1995) model. It seems similar to the model of Raffone et al., at least in its orientation and intent (with a reliance on the concept of competition between neural networks). An impressive aspect of this model is that it presents another possibility that did not emerge from my analysis of the behavioral literature. It was stated that an inhibition parameter "is set high when selection is required and moderate when multiple items need to be maintained together, as in immediate recall tasks." Yet, the capacity of the system is limited to 3–5 items and "cannot increase its capacity beyond this range by diminishing the inhibition parameter even further." The limit was rationalized by the observation that one cannot allow less representational overlap (in order to increase the limit) because "representational overlap is essential for computations that perform generalization." This model appears to be a version of H1 above (a single attentional limit), but with a flexible attentional focus ranging from 1 to 5 chunks depending on the intensity of focus. What I characterized as the singular coherent scene could be the result of a singular focus of attention when it is in the intensely focused mode. As evidence continues to accrue, it should have implications for which model is most promising. The model of Usher et al. appears to be a version of H3 because the capacity depends partly on a concept intrinsic to the memory representations (the amount of overlap) and partly on a concept related to attention (level of inhibition). However, it is an interactive concept because neither the representation alone nor the level of inhibition alone state a capacity; that is stated by the interaction of these parameters.

Such a model might be usefully tested in relation to patients with frontal lobe damage. According to a reasonable extension of the model, the ability to control the focus of attention might be altered (e.g., making a narrow focus impossible) while the maximal capacity limit (3–5 chunks) would not necessarily be affected, depending as it does on the degree of representational overlap.

Several other commentators (**Milner; Roelfsema & Lamme; Tiitinen**) did not fully present models of capacity limits, but used modeling concepts to challenge or query aspects of the present approach. **Milner** asked, "how much does it further our knowledge to be told that the span is limited because the focus of attention can hold only about 4 items at a time?" I have tried to show that although this theoretical statement may not be right, it does have empirical consequences that differ from other theoretical conceptions (as shown in Fig. R2 and accompanying text). Milner also objected that "the neural substrate of attention is not a unitary system that can be pointed like a spotlight or a camera, much less a static process into which peripatetic images can be directed. Apparently it must be a highly organized system of centrifugal paths, every bit as specific as the centripetal sensory paths that it modulates." I tend to agree but would urge caution when mixing levels of analysis. A central circle for the focus of attention in my cognitive model (Cowan 1988; 1995) was not intended to imply geographic unity of the focus of attention in the brain. It instead was meant, at most, to refer to a more abstract kind of unity, where all elements in the focus of attention adhere to a common theme (i.e., scene coherence). Moreover, the unitary nature of the representation of activated information (surrounding the focus of attention) is no more than an oversimplification in the drawing, inasmuch as the theoretical view places or limits on requirement of coherence on what can be activated at any moment.

Roelfsema & Lamme concluded that, "the clarity with which the psychophysical data point to the number four is not matched by a similarly clear limit imposed by physiological mechanisms." That sort of consideration may perhaps provide a motivation to increase the amount of complexity in neural modeling of the capacity limit, as **Raffone et al.** and **Usher et al.** did. It also may be worth noting the obvious point that, in attempting to find the correspondence between physiology and behavior, both of them have to be examined very carefully. Roelfsema & Lamme distinguished between "base groupings" (combinations of simple features) and "incremental groupings" (combinations of more abstract features) and proposed that only incremental groupings are capacity-limited, whereas base groupings "are rapidly available" with "no clear limit on the number of base groupings that can be computed in parallel during stimulus presentation." For example, it was stated that "the activation of a neuron tuned to, for example, red and vertical provides a base grouping between these features." This distinction seems to leave open a question about how an individual would find a red vertical line among red horizontal lines and blue vertical lines. We know that this requires a slow, capacity-limited search process (Treisman & Gelade 1980) but it is unclear if Roelfsema & Lamme's conception classifies this as a situation handled by base or incremental groupings.

Tiitinen asked what kind of research could best link neurophysiology to behavior. He presented some objections to the idea that 40-Hz oscillations per se must mediate a limited-capacity store, given that these oscillations occur in an attenuated form even in control conditions. I think that part of the answer to this challenge is to proceed with research examining behavioral consequences of neurophysiological concepts such as 40-Hz oscillations (already begun by Burle & Bonnet 2000; Elliott & Müller 1998; 2000). Another part is to conduct neurobiological research to capture patterns that have been observed only behaviorally. An example is the finding of **Rypma & Gabrieli** that the brain reaction is much different for 1 or 3 items versus 6 items, and the suggestion of using that finding to explore the way in which items are chunked together. All of this brings up another issue, which is what neurobiology and behavior, and what evidence and models, have to offer one another.

R4.5. Relation between evidence and models in cognition and neurobiology

We have seen that there is considerable ongoing work related to capacity limits. It is of both an empirical and a theoretical nature, and is occurring in both cognitive and neurobiological fields. (Cognitive research on populations other than adult humans can be considered in this context to encroach upon neurobiological topics, inasmuch as the differences in brain structure between types of intelligent organism should count as neurobiological quasi-manipulations.) In this section, I would like to offer a few musings about where all of this research may lead, rather than leaving everything hanging. It is important to consider the logical relations between four types of research endeavor: (1) empirical research on cognition; (2) empirical research on neurobiology; (3) theoretical modeling of cognition; and (4) theoretical modeling of neurobiology. If the aim is to account for a cognitive phenomenon, the pure STM capacity limit, the relation between these endeavors is in some cases asymmetrical.

Some cognitive theory is necessary even to observe a cognitive generalization. Thus, the observation that only 4 chunks can be held in working memory could be made only with some assumptions about what counts as a chunk. However, the simplicity of the answer that resulted lends some support to those theoretical assumptions by allowing the detection of a simple pattern in the data (consistent with Occam's razor). To go further, a theory of capacity must be stated in cognitive terms so that the elements in the theory lead to predictions in cognitive terms. Figure R2 summarizes some attempts along those lines and necessarily introduces abstract concepts (active memory representations; an attentional focus) that can be directly related to behavior.

Neurobiological research is of two types. In some circumstances, manipulations of neurobiological factors cause cognitive behavioral differences, and a causal link can be drawn. We have not reviewed any evidence of exactly that type but it is possible in the future. For example, it is reasonable to believe that alcohol may affect working memory by altering the capacity of attention (see Steele & Josephs 1990). However, a neurobiological manipulation has one limitation that is the same as for a cognitive manipulation: in either case, the manipulation may not be very specific in its effect. Any manipulation, biological or behavioral in nature, that affects capacity only by affecting every single cognitive function probably reveals little about the mechanism; more specific effects are more instructive. On the other hand, a large neurobiological change that results in little change in capacity limits is potentially very instructive. For example, if it turns out that lower animals have a capacity limit that is the same as in adult humans (see **Todt**), that would show a dissociation between the capacity limit and higher cognitive functions. It would suggest, as various commentators have proposed, that the capacity limit may reflect the optimization of certain quantities, such as the necessary degree of overlap between the neural representations of different concepts (**Usher et al.**).

The other basic kind of neurobiological research, including electrophysiological recording and neuroimaging, is essentially correlative in nature rather than manipulative (if carried out only on normal adult humans to ascertain group means). Thus, one tries to observe correlations between certain patterns of behaviors and patterns of brain response. This type of research can help in several ways. If it produces brain response patterns that are in accord with cognitive theory (e.g., the different responses for arrays of sub- versus supra-capacity numbers of items; see **Rypma & Gabrieli**), that then strengthens the cognitive conclusions by showing an orderly and potentially interpretable pattern across levels of analysis.

Any discrepancy between cognitive and neurobiological results, indicates that something unexpected may be going on behind the scenes. In that case, however, the discrepancy may be difficult to interpret without the help of neurobiological theory. For example, the complaint of **Roelfsema & Lamme** that neural populations do not tend to form the oscillatory patterns needed for a model such as that of Lisman and Idiart (1995) may provide a motivation to look for more complex theories, such as those of **Raffone et al.** and **Usher et al.**

Finally, neurobiological theory, spurred on by neurobiological as well as cognitive evidence, can contribute to cognitive theory. It can do this by serving as the implementation of the cognitive theory at a finer-grained lever of analysis (Marr 1982). It also can help by leading to possibilities that otherwise would not have occurred to the cognitive investigators. For example, the use of representational overlap and inhibition concepts (**Usher et al.**) led to the understanding that the same focus of attention could be limited to as few as one chunk or as many at three to five, depending on the setting of an inhibition parameter.

R5. Appraisal of models

In this final section I will briefly try to extract, from the immense amount of information provided by the commentators, an improved understanding of the basis of capacity limits. I will do this with respect to the models illustrated in Figure R2 and accompanying text.

The main question I will address is whether the 3- to 5-chunk capacity limit in adult humans is likely to be caused by the nature of focus of attention (H1), the nature of memory representations (H2), or both (H3). It seems obvious from the literature that was reviewed in the target article that attention exists and that performance can be limited similarly either by removing attention or by limiting its effects (e.g., through articulatory suppression or presentation of a complex array of items); so I will not entertain H4. The question for me is just where the pure STM capacity limit comes from.

Several types of study that do not yet exist seem rather crucial (and, indeed, several laboratories are working on them). First, we need studies using two concurrent pure-capacity tasks, in which memory loads in very different domains are presented concurrently to determine if the 4-chunk limit is shared across domains or is reproduced intact within each domain even in this dual-task situation. We have plenty of dual-task experiments but most have not attempted to use tasks that estimate pure STM capacity. Second, we need studies of the correlations of pure STM capacity estimates across domains.

Lacking this evidence, I will still offer an opinion because anyone who has followed this extensive controversy has earned some tentative bottom line. I lean toward H1, in which the focus of attention is limited in capacity, essentially because we have seen that *very different underlying*

forms of memory representation yield similar capacity estimates. For example, consider two situations that may be best suited to estimate capacity: those in which only one decision has had to be made per trial and there has been no interference from previous outputs (e.g., Cowan et al. unpublished, discussed in sect. R2.3.1; Luck & Vogel 1997). The unattended-speech research shows a bow-shaped serial position function, whereas running memory span shows a monotonic increasing function across the tested serial positions; and the task of Luck and Vogel is based on spatial arrays of colors represented simultaneously. Despite such representational differences, a common capacity limit seems to apply, and it requires explanation.

The best suggestion I can offer is that *aspects of the memory representation determine what chunks will be most prominent* (relative to the available retrieval context), whereas *limits in the focus of attention determine how many of the most prominent chunks in the representation can be attended at once.* Thus, for example, in memory for unattended speech, about four prominent items typically would be picked off from the two ends of the list together; whereas, in running span, about four prominent items typically would be picked off from the recency portion alone. We do not know which portions of a spatial array are most prominent in memory but there probably are especially prominent areas, perhaps those nearest the fovea. Despite the representational diversity, a common capacity limit seems to apply.

Some concepts suggested by commentaries may lead to minor modifications in this H1 account. Suppose **Usher et al.'s** concept of representational overlap does not apply for activated representations in general, but only for a special type of representation that includes just recently-attended items. It may be within *that type of representation only* that representational overlap limits the capacity to 3 to 5 chunks. If so, this could be counted either as a limit in the focus of attention per se (H1), or as a special representational limit (H2). Similarly, suppose the chunk limit occurs only in a representation corresponding to **Baddeley's** episodic buffer, and not in the simpler passive stores, and suppose this episodic buffer receives input only from the focus of attention. If so, it is difficult to know exactly how to apply the theoretical distinction shown in Figure R2. As just mentioned, I prefer to think that the situation best matches H1. While the present theoretical approach cannot yet indicate exactly the right model, it can provide a more detailed vision of the limited-capacity process than before, with the benefit of an extraordinarily thoughtful set of commentaries.

ACKNOWLEDGMENTS

This project was supported by NICHD Grant R01 21338. I thank Jeffrey Rouder for helpful comments. Address correspondence to Nelson Cowan, Department of Psychology, University of Missouri, 210 McAlester Hall, Columbia, MO 65211. E-mail: cowanN@Missouri.edu.

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Adams, D. (1985) *The hitch-hiker's guide to the galaxy: The original radio scripts.* Pan Books. [JNT]

Allport, D. A. (1968) Phenomenal simultaneity and the perceptual moment hypothesis. *British Journal of Psychology* 59:395–406. [rNC]

(1987) Selection-for-action: Some behavioural and neurophysiological considerations of attention and action. In: *Perspectives on perception and action,* ed. H. Heuer & A. F. Sanders. Erlbaum. [TS]

Amit, D. J. (1995) The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences* 18:615–57. [AR]

Anderson, J. R. (1983) *The architecture of cognition.* Harvard University Press. [SM]

Anderson, J. R. & Lebiere, C. (1998) *The atomic components of thought.* Erlbaum. [NAT]

Anderson, J. R. & Matessa, M. (1997) A production system theory of serial memory. *Psychological Review* 104:728–48. [aNC, SM, rNC]

Anderson, R. B., Tweney, R. D., Rivardo, M. & Duncan, S. (1997) Need probability affects retention: A direct demonstration. *Memory and Cognition* 25:867–72. [PM]

Ashby, F. G., Tien, J.-Y. & Balikrishnan, J. D. (1993) Response time distributions in memory scanning. *Journal of Mathematical Psychology* 37:526–55. [rNC]

Atkinson, J., Campbell, F. W. & Francis, M. R. (1976a) The magic number 4 ± 0: A new look at visual numerosity. *Perception* 5:327–34. [aNC, BJS]

Atkinson, J., Francis, M. R. & Campbell, F. W. (1976b) The dependence of the visual numerosity limit on orientation, colour, and grouping of the stimulus. *Perception* 5:335–42. [aNC]

Atkinson, R. C. & Shiffrin, R. M. (1968) Human memory: A proposed system and its control processes. In: *The psychology of learning and motivation: Advances in research and theory, vol. 2,* ed. K. W. Spence & J. T. Spence. Academic Press. [aNC, rNC]

Avons, S. E., Wright, K. L. & Pammer, K. (1994) The word-length effect in probed and serial recall. *Quarterly Journal of Experimental Psychology* 47A:207–31. [aNC]

Baars, B. J. (1988) *A cognitive theory of consciousness.* Cambridge University Press. [aNC, BJB, rNC]

(1993) How does a serial, integrated and very limited stream of consciousness emerge from a nervous system that is mostly unconscious, distributed, parallel and of enormous capacity? *Theoretical and Experimental Studies of Consciousness, Ciba Foundation Symposium* 174:282–303. [BJB]

(1997) *In the theater of consciousness: The workspace of the mind.* Oxford University Press. [BJB]

(1998) Metaphors of consciousness and attention in the brain. *Trends in Neurosciences* 21:58–62. [BJB]

Baars, B. J. & Newman, J., ed. (in press) *Essential sources in the scientific study of consciousness.* MIT Press/Bradford Books. [BJB]

Bachelder, B. L. (1978) Span Theory: Laboratory and classroom applications. Paper presented at the Eleventh Annual Gatlinburg Conference on Research in Mental Retardation, Gatlinburg, TN. [BLB]

(2000) The magical number seven: What did Miller really prove? (submitted). [BLB]

Bachelder, B. L. & Denny, M. R. (1977a) A theory of intelligence: I. Span and the complexity of stimulus control. *Intelligence* 1:127–50. [BLB]

(1977b) A theory of intelligence: II The role of span in a variety of intellectual tasks. *Intelligence* 1:237–56. [BLB]

Baddeley, A. D. (1986) *Working memory.* Clarendon Press. [aNC, ADB, SG, SM, BR, BJS, TS, rNC]

(1993) Visual and verbal subsystems of working memory. *Current Biology* 3:563–65. [BJB]

(1996) Exploring the central executive. *Quarterly Journal of Experimental Psychology* 49A:5–28. [ADB]

(2000) The episodic buffer: A new component for working memory? *Trends in Cognitive Sciences* 4:417–23. [ADB]

(in press) Levels of working memory. In: *Perspectives on human memory and cognitive aging: Essays in honor of Fergus Craik,* ed. M. Naveh-Benjamin, M. Moscovitch & H. L. Roediger. Psychology Press. [rNC]

Baddeley, A. D. & Ecob, J. R. (1973) Reaction time and short-term memory: Implications of repetition effects for the high-speed exhaustive scan hypothesis. *Quarterly Journal of Experimental Psychology* 25:229–40. [JJ]

Baddeley, A. D., Grant, S., Wight, E. & Thompson, N. (1975a) Imagery and visual working memory. In: *Attention and performance V,* ed. P. Rabbit and S. Dornic. Academic Press. [ADB]

Baddeley, A. & Hitch, G. J. (1974) Working memory. In: *Recent advances in learning and motivation, vol. 8,* ed. G. Bower. Academic Press. [ADB, BR]

Baddeley, A. D., Lewis, V., Eldridge, M. & Thomson, N. (1984) Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General* 113:518–40. [ADB]

Baddeley, A., Lewis, V. & Vallar, G. (1984) Exploring the articulatory loop. *The Quarterly Journal of Experimental Psychology* 36A:233–52. [aNC]

Baddeley, A. D. & Logie, R. (1999) Working memory: The multiple component model. In: *Models of working memory: Mechanisms of active maintenance and executive control,* ed. A. Miyake & P. Shah. Cambridge University Press. [ADB]

Baddeley, A. D., Thomson, N. & Buchanan, M. (1975b) Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* 14:575–89. [aNC, SM]

Basso, A., Spinnler, H., Vallar, G. & Zanobio, E. (1982) Left hemisphere damage and selective impairment of auditory verbal STM. *Neuropsychologia* 20:263–74. [ADB]

Bender, S., Schall, U., Wolstein, J., Grzella, I., Zerbin, D. & Oades, R. D. (1999) A topographic event-related potential follow-up study of "prepulse inhibition" in first and second episode patients with schizophrenia. *Psychiatry Research Neuroimaging* 90:41–53. [RDO]

Besner, D. (1987) Phonology, lexical access in reading, and articulatory suppression: A critical review. *Quarterly Journal of Experimental Psychology* 39A:467–78. [aNC]

Biro, D. & Matsuzawa, T. (1999) Numerical ordering in a chimpanzee (*Pan troglodytes*): Planning, executing, and monitoring. *Journal of Comparative Psychology* 113:178–85. [NK]

Bjork, R. A. & Whitten, W. B. (1974) Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology* 6:173–89. [aNC]

Boudewijnse, G.-J. A., Murray, D. J. & Bandomir, C. A. (1999) Herbart's mathematical psychology. *History of Psychology* 2:163–93. [DJM]

Bower, G. H. & Winzenz, D. (1969) Group structure, coding and memory for digit series. *Journal of Experimental Psychology Monographs* 80:1–17. [aNC]

Brandimonte, M. A., Hitch, G. J. & Bishop, D. V. M. (1992) Influences of short-term memory codes on visual image processing: Evidence from image transformation tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18:157–65. [SM]

Braun, H. A., Wissing, H., Schäfer, K. & Hirsch, M. C. (1994) Oscillation and noise determine signal transduction in shark multimodal sensory cells. *Nature* 367:270–73. [aNC]

Brebion, G., Smith, M. J., Gorman, J. M., Malaspina, D., Sharif, Z. & Amador, X. (2000) Memory and schizophrenia: Differential link of processing speed and selective attention with two levels of encoding. *Journal of Psychiatric Research* 34:121–27. [RDO]

Bregman, A. S. & Campbell, J. (1971) Primary auditory stream segregation and perception of order in rapid sequences. *Journal of Experimental Psychology* 89:244–49. [rNC]

Bregman, A. S. & Dannenbring, G. L. (1973) The effect of continuity on auditory stream segregation. *Perception and Psychophysics* 13:308–312. [rNC]

Broadbent, D. E. (1958) *Perception and communication.* Pergamon Press. [aNC, HT, rNC]

Broadbent, D. E. (1971) *Decision and stress.* Academic Press. [rNC]

(1975) The magic number seven after fifteen years. In: *Studies in long-term memory,* ed. A. Kennedy & A. Wilkes. Wiley. [aNC]

Brown, G. D. A. & Hulme, C. (1995) Modeling item length effects in memory span: No rehearsal needed? *Journal of Memory and Language* 34:594–621. [aNC, SM]

Brown, G. D. A., Preece, T. & Hulme, C. (2000) Oscillator-based memory for serial order. *Psychological Review* 107:127–81. [aNC, rNC]

Burgess, N. & Hitch, G. J. (1997) Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review* 106:551–81. [SM]

Burle, B. & Bonnet, M. (2000) High-speed memory scanning: A behavioral argument for a serial oscillatory model. *Cognitive Brain Research* 9:327–37. [OJ, rNC]

Burrows, D. & Okada, R. (1975) Memory retrieval from long and short lists. *Science* 188:1031–33. [JJ]

Burtis, P. J. (1982) Capacity increase and chunking in the development of short-term memory. *Journal of Experimental Child Psychology* 34:387–413. [JPL]

Caplan, D., Rochon, E. & Waters, G. S. (1992) Articulatory and phonological determinants of word length in span tasks. *Quarterly Journal of Experimental Psychology* 45A:177–92. [SM]

Cardozo, B. L. & Leopold, F. F. (1963) Human code transmission. *Ergonomics* 6:133–41. [aNC]

Carey, S. & Xu, F. (in press) Infant knowledge of objects: Beyond object files and object tracking. *Cognition: Special Issue on Objects and Attention.* [BJS]

Case, R. (1974) Structures and strictures: Some functional limitations on the course of cognitive development. *Cognitive Psychology* 6:544–73. [SM]

(1985) *Intellectual development: Birth to adulthood.* Academic Press. [SM]

(1995) Capacity-based explanations of working memory growth: A brief history and reevaluation. In: *Memory performance and competencies: Issues in growth and development,* ed. F. E. Weinert & W. Schneider. Erlbaum. [SM]

(1998) The development of conceptual structures. In: *Handbook of child psychology,* vol. 2, *Cognition, perception, and language,* ed. W. Damon, D. Kuhn & R. S. Siegler. John Wiley and Sons. [JPL]

Case, R. & Okamoto, Y. (1996) The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development,* vol. 61, no. 245. [JPL]

Catchpole, C. K. & Slater, P. J. B. (1995) *Bird song: Biological themes and variations.* Cambridge University Press. [DT]

Cavanagh, J. P. (1972) Relation between the immediate memory span and the memory search rate. *Psychological Review* 79:525–30. [DJM]

Charniak, E. (1993) *Statistical language learning.* MIT Press. [EMP]

Chater, N. & Christiansen, M. (1999) Connectionist and natural language processing. In: *Language Processing,* ed. S. Garrod & M. Pickering. Psychology Press. [EMP]

Chase, W. G. & Ericsson, K. A. (1981) Skilled memory. In: *Cognitive skills and their acquisition,* ed. J. R. Anderson. Erlbaum. [KAE]

(1982) Skill and working memory. In: *The psychology of learning and motivation, vol. 16,* ed. G. H. Bower. Academic Press. [KAE]

Chase, W & Simon, H. A. (1973) The mind eye's in chess. In: *Visual information processing,* ed. by W. G. Chase. Academic Press. [aNC, PCL, rNC]

Chelazzi, L., Miller, E. K., Duncan, J. & Desimone, R. (1993) A neural basis for visual search in inferior temporal cortex. *Nature* 363:345–47. [PRR]

Chen, S., Swartz, K. B. & Terrace, H. S. (1998) Knowledge of the ordinal position of list items in rhesus monkeys. *Psychological Science* 8:80–86. [DT]

Cherry, E. C. (1953) Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America* 25:975–79. [aNC, HT]

Cheung, H. & Kemper, S. (1993) Recall and articulation of English and Chinese words by Chinese-English bilinguals. *Memory and Cognition* 21:666–70. [SM]

Chi, M. T. H. & Klahr, D. (1975) Span and rate of apprehension in children and adults. *Journal of Experimental Child Psychology* 19:434–39. [aNC]

Chomsky, N. (1975) *Reflections on language.* Pantheon. [EMP]

Cleeremans, A. & McClelland, J. L. (1991) Learning the structure of event sequences. *Journal of Experimental Psychology: General* 120:235–53. [aNC, EMP]

Clifton, C & Cruse, D. (1977) Time to recognize tones: Memory scanning or memory strength? *Quarterly Journal of Experimental Psychology* 29:709–26. [rNC]

Cohen, J. D., Braver, T. S. & O'Reilly, R. C. (1996) A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society of London B* 351:1515–27. [MU]

Cohen, J. D., Forman, S. D., Braver, T. S., Casey, B. J., Servan-Schreiber, D. & Noll, D. C. (1994) Activation of the prefrontal cortex in a nonspatial working memory task with function MRI. *Human Brain Mapping* 293–304. [BR]

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J. & Smith, E. E. (1977) Temporal dynamics of brain activation during a working memory task. *Nature* 386:604–608. [aNC]

Conrad, R. (1965) Order error in immediate serial recall of sequences. *Journal of Verbal Learning and Verbal Behavior* 4:161–69. [SEA]

Corballis, M. C. (1967) Serial order in recognition and recall. *Journal of Experimental Psychology* 74:99–105. [aNC]

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L. & Petersen, S. E. (1991) Selective and divided attention during visual discrimination of shape, color, and speed: Functional anatomy by positron emission tomography. *The Journal of Neuroscience* 11:2383–2402. [BR]

Coslett, H. B. & Saffran, E. (1991) Simultanagnosia: To see but not two see. *Brain* 114:1523–45. [BJS]

Cowan, N. (1984) On short and long auditory stores. *Psychological Bulletin* 96:341–70. [rNC]

(1988) Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin* 104:163–91. [aNC, SM, HT, rNC]

(1995) *Attention and memory: An integrated framework.* Oxford Psychology Series, No. 26. Oxford University Press. [aNC, BJS, HT, rNC]

(1999) An embedded-processes model of working memory. In: *Models of working memory: Mechanisms of active maintenance and executive control,* eds. A. Miyake & P. Shah. Cambridge University Press. [rNC]

Cowan, N., Cartwright, C., Winterowd, C. & Sherk, M. (1987) An adult model of preschool children's speech memory. *Memory and Cognition* 15:511–17. [aNC]

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T. & Flores, L. (1992) The role of verbal output time in the effect of word length on immediate memory. *Journal of Memory and Language* 31:1–17. [SM]

Cowan, N., Lichty, W. & Grove, T. R. (1990) Properties of memory for unattended spoken syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16:258–69. [rNC]

Cowan, N., Nugent, L. D., Elliot, E. M. & Geer, T. (2000a) Is there a temporal basis of the word length effect? A response to Service (1998). *Quarterly Journal of Experimental Psychology* 53A:647–60. [rNC]

Cowan, N., Nugent, L. D., Elliot, E. M., Ponomarev, I. & Saults, J. S. (1999) The role of attention in the development of short-term memory: Age differences in the verbal span of apprehension. *Child Development* 70:1082–97. [aNC, SAH, JPL, HT, rNC]

Cowan, N., Nugent, L. D., Elliot, E. M. & Saults, J. S. (2000b) Persistence of

memory for ignored lists of digits: Areas of development constancy and change. *Journal of Experimental Child Psychology* 76:151–72. [rNC]

Cowan, N., Saults, J. S. & Nugent, L. D. (1997a) The role of absolute and relative amounts of time in forgetting within immediate memory: The case of tone pitch comparisons. *Psychonomic Bulletin and Review* 4:393–97. [aNC, rNC]

(in press) The ravages of absolute and relative amounts time on memory. In: *The nature of remembering: Essays in honor of Robert G. Crowder,* ed. H. Roediger. American Psychological Association. [rNC]

Cowan, N., Winkler, I., Teder, W. & Näätänen, R. (1993) Memory prerequisites of the mismatch negativity in the auditory event-related potential (ERP). *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19:909–21. [aNC]

Cowan, N., Wood, N. L. & Borne, D. N. (1994) Reconfirmation of the short-term storage concept. *Psychological Science* 5:103–106. [rNC]

Cowan, N., Wood, N. L., Nugent, L. D. & Treisman, M. (1997b) There are two word length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science* 8:290–95. [aNC, rNC]

Cowan, N., Wood, N. L., Wood, P. K., Keller, T. A., Nugent, L. D. & Keller, C. V. (1998) Two separate verbal processing rates contributing to short-term memory span. *Journal of Experimental Psychology: General* 127:141–60. [aNC]

Craik, F. I. M. & Birtwistle, J. (1971) Proactive inhibition in free recall. *Journal of Experimental Psychology* 91:120–23. [rNC]

Craik, F., Gardiner, J. M. & Watkins, M. J. (1970) Further evidence for a negative recency effect in free recall. *Journal of Verbal Learning and Verbal Behavior* 9:554–60. [aNC]

Craik, F. I. M., Govoni, R., Naveh Benjamin, M. & Anderson, N. D. (1996) The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General* 125:159–80. [ADB]

Craik, F. I. M. & Lockhart, R. S. (1972) Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11:671–84. [PM]

Crick, F. (1984) Function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences, USA* 81:4586–90. [BJB]

(1994) *The astonishing hypothesis.* Touchtone Books. [JPL]

Crowder, R. G. (1989) Imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance* 15:472–78. [rNC],

(1993) Short-term memory: Where do we stand? *Memory and Cognition* 21:142–45. [aNC]

Cunningham, T. F., Healy, A. F., Till, R. E., Fendrich, D. W. & Dimitry, C. Z. (1993) Is there really very rapid forgetting from primary memory? The role of expectancy and item importance in short-term recall. *Memory and Cognition* 21:671–88. [PM]

Damasio, A. R. (1989) Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33:25–62. [BJB]

Daneman, M. & Carpenter, P. A. (1980) Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior* 19:450–66. [aNC, JNT, rNC]

Daneman, M. & Merikle, P. M. (1996) Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review* 3:422–33. [aNC, rNC]

Darwin, C. J., Turvey, M. T. & Crowder, R. G. (1972) An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology* 3:255–67. [aNC, rNC]

Davis, G., Driver, J., Pavani, F. & Shepherd A. J. (2000) Reappraising the costs of attending to two separate visual objects. *Vision Research* 40:1323–32. [GD]

Davis, G., Welch, V. L., Holmes, A. & Shepherd, A. J. (submitted) Can attention select only a fixed number of objects at a time?

D'Esposito, M., Detre, J., Alsop, D., Shin, R., Atlas, S. & Grossman, M. (1995) The neural basis of the central executive system of working memory. *Nature* 378:279–81. [BR]

De Groot, A. D. (1978) *Thought and choice in chess*, 2nd ed. Morton. [PCL]

Dehaene, S. & Cohen, L. (1994) Dissociable mechanisms of subitizing and counting: Neuropsychological evidence from simultanagnosic patients. *Journal of Experimental Psychology: Human Perception and Performance* 20:958–75. [BJS]

Dempster, F. N. (1981) Memory span: Sources of individual and developmental differences. *Psychological Bulletin* 89:63–100. [aNC]

Desimone, R. (1998) Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London B* 353:1245–55. [MU]

Desimone, R. & Duncan, J. (1995) Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18:193–222. [PRR]

Destexhe, A., Contreras, D. & Steriade, M. (1999) Spatiotemporal analysis of local field potentials and unit discharges in cat cerebral cortex during natural wake and sleep states. *Journal of Neuroscience* 19:4595–4608. [BJB]

Diener, D. (1990) Role of the preprobe delay in memory-scanning tasks. *Memory and Cognition* 18:451–58. [rNC]

Dirlam, D. K. (1972) Most efficient chunk sizes. *Cognitive Psychology* 3:355–59. [aNC]

Dosher, B. A. (1981) The effect of delay and interference: A speed-accuracy study. *Cognitive Psychology* 13:551–82. [BM, rNC]

(1984) Degree of learning and retrieval speed: Study time and multiple exposures. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10:541–74. [BM]

(1998) Models of visual search: Finding a face in the crowd. In: *Methods, models, and conceptual issues: An invitation to cognitive science,* ed. D. Scarborough and S. Sternberg. MIT Press. [rNC]

Dosher, B. A. & Ma, J. J. (1998) Output loss or rehearsal loop? Output-time versus pronunciation-time limits in immediate recall for forgetting-matched materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24:316–35. [BM]

Duncan, J. (1984) Selective attention and the organization of visual information. *Journal of Experimental Psychology: General* 113:501–17. [GD]

Elliot, M. A. & Müller, H. J. (1998) Synchronous information presented in 40 Hz flicker enhances visual feature binding. *Psychological Science* 9:277–83. [rNC]

(2000) Evidence for a 40-Hz oscillatory short-term visual memory revealed by human reaction-time measurements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:703–18. [rNC]

Elman, J. L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition* 48:71–99. [aNC, EMP]

(1996) *Rethinking innateness: A connectionist perspective on development.* MIT Press. [EMP]

Engel, A. K., Fries, P., Roelfsema, P. R., König, P., Brecht, M. & Singer, W. (1999) Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition* 8: 128–51. [BJB]

Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B. & Singer, W. (1992) Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neuroscience* 15:218–26. [AR]

Engle, R. W., Kane, M. J. & Tuholski, S. W. (1999) Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In: *Models of working memory: Mechanisms of active maintenance and executive control,* ed. A. Miyake and P. Shah. Cambridge University Press. [aNC]

Ericsson, K. A. (1985) Memory skill. *Canadian Journal of Psychology* 39:188–231. [aNC, HT]

(1996) *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games.* Erlbaum. [KAE]

Ericsson, K. A., Chase, W. G. & Faloon, S. (1980) Acquisition of a memory skill. *Science* 208:1181–82. [aNC, HT, rNC]

Ericsson, K. A. & Kintsch, W. (1995) Long-term working memory. *Psychological Review* 102:211–45. [aNC, KAE]

Ericsson, K. A., Patel, V. L. & Kintsch, W. (2000) How experts' adaptation to representative task demands account for expertise effect in memory recall: Comment on Vincente & Wang (1998). *Psychological Review* 107:578–92. [KAE]

Ericsson, K. A. & Polson, P. G. (1988) Memory for restaurant orders. In: *The nature of expertise,* ed. M. Chi, R. Glaser & M. Farr. Erlbaum. [KAE]

Fabiani, M., Stadler, M. A. & Wessels, P. M. (in press) True memories but not false ones produce a sensory signature in human lateralized brain potentials. *Journal of Cognitive Neuroscience.* [GG]

Feigenbaum, E. A. & Simon, H. A. (1984) EPAM-like models of recognition and learning. *Cognitive Science.* 8:305–36. [PCL]

Fisher, D. L. (1984) Central capacity limits in consistent mapping, visual search tasks: Four channels or more? *Cognitive Psychology* 16:449–84. [aNC, TS]

Fraise, P. (1978) Time and rhythm perception. In: *Handbook of Perception,* vol. 8, eds. E. Carterette & M. Friedman. Academic Press. [SG]

Frensch, P. A. & Miner, C. S. (1994) Effects of presentation rate and individual differences in short-term memory capacity on an indirect measure of serial learning. *Memory and Cognition* 22:95–110. [aNC]

Frick, R. W. (1984) Using both an auditory and a visual short-term store to increase digit span. *Memory and Cognition* 12:507–14. [aNC]

Friedman-Hill, S. R., Robertson, L. C. & Treisman, A. (1995) Parietal contributions to visual feature binding: Evidence from a patient with bilateral lesions. *Science* 269:853–55. [BJS]

Frith, C. D. & Dolan, R. J. (1997) Brain mechanisms associated with top-down processes in perception. *Philosophical Transactions of the Royal Society of London B,* 352:1221–30. [RDO]

Fuster, J. M. (1997) Network memory. *Trends in Neuroscience* 20:451–59. [PRR]

Fuster, J. M., Bauer, R. H. & Jervey, J. P. (1985) Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. *Brain Research* 330:299–307. [PMM]

Gathercole, S. E. & Baddeley, A. D. (1993) *Working memory and language.* Erlbaum. [ADB]
Geary, D. C. (1993) Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin* 114:345–62. [SAH, rNC]
(1995) Reflections of evolution and culture in children's cognition. *American Psychologist* 50:24–36. [SAH, rNC]
George, N., Jemel, B., Fiori, N. & Renault, B. (2000) Holistic and part-based face representations: Evidence from the memory span of the "face superiority effect." *Current Psychology Letters* 1:89–106. [RDO]
Glanzer, M. & Cunitz, A. R. (1966) Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior* 5:351–60. [aNC]
Glanzer, M. & Razel, M. (1974) The size of the unit in short-term storage. *Journal of Verbal Learning and Verbal Behavior* 13:114–31. [aNC]
Glenberg, A. M. & Swanson, N. C. (1986) A temporal distinctiveness theory of recency and modality effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12:3–15. [aNC]
Gobet, F. (1998) Memory for the meaningless: How chunks help. *Proceedings of the Twentieth Meeting of the Cognitive Science Society.* Erlbaum. [PCL]
Gobet, F. & Simon, H. A. (1996) Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology* 31:1–40. [aNC, KAE]
(1998) Expert chess memory: Revisiting the chunking hypothesis. *Memory* 6:225–55. [aNC, PCL]
(in press) Five seconds or sixty? Presentation time in expert memory. *Cognitive Science.* [PCL]
Gold, E. M. (1967) Language identification in the limit. *Information and Control* 16:447–74. [EMP]
Gomes, H., Sussman, E., Ritter, W., Kurtzberg, D., Cowan, N. & Vaughan Jr., H. G. (1999) Electrophysiological evidence of developmental changes in the duration of auditory sensory memory. *Developmental Psychology* 35:294–302. [rNC]
Goodman, N. (1951) *The structure of appearance.* Harvard University Press. [JPL]
Graesser II, A. & Mandler, G. (1978) Limited processing capacity constrains the storage of unrelated sets of words and retrieval from natural categories. *Journal of Experimental Psychology: Human Learning and Memory* 4:86–100. [aNC]
Granholm, E., Perry, W., Filoteo, J. V. & Braff, D. (1999) Hemispheric and attentional contributions to perceptual organization deficits on the global-local task in schizophrenia. *Neuropsychology* 13:271–81. [RDO]
Gratton, G., Corballis, P. M. & Jain, S. (1997) Hemispheric organization of visual memories. *Journal of Cognitive Neuroscience* 9:92–104. [GG]
Gratton, G., Fabiani, M., Goodman-Wood, M. R. & DeSoto, M. C. (1998) Memory-driven processing in human medial occipital cortex: An event-related optical signal (EROS) study. *Psychophysiology* 38:348–51. [GG]
Gray, C. M., König, P., Engel, A. K. & Singer, W. (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization, which reflects global stimulus properties. *Nature* 338:334–37. [aNC, PRR, HT]
Greene, R. L. (1989) Immediate serial recall of mixed-modality lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15:266–74. [aNC]
Greenwald, A. & Shulman, H. (1973) On doing two at once: II. Elimination of the psychological refractory period. *Journal of Experimental Psychology* 101:70–76. [TS]
Grondin, S. (in press) From physical time to the first and second moments of psychological time. *Psychological Bulletin.* [SG]
Grondin, S., Meilleur-Wells, G. & Lachance, R. (1999) When to start explicit counting in time-intervals discrimination tasks: A critical point in the timing process of humans. *Journal of Experimental Psychology: Human Perception and Performance* 25:993–1004. [SG]
Gross, C. G., Bender, D. B. & Rocha-Miranda, C. E. (1974) Inferotemporal cortex: A single-unit analysis. In: *The neurosciences: Third study program,* ed. F. O. Schmitt & F. G. Worden. MIT Press. [PMM]
Grossberg, S. (1976) Adaptive pattern classification and universal recoding. *Biological Cybernetics* 23:121–34. [MU]
Gruenewald, P. J. & Lockhead, G. R. (1980) The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory* 6:225–40. [aNC]
Guilford, J. P. & Dallenback, K. M. (1925) The determination of memory span by the method of constant stimuli. *American Journal of Psychology* 36:621–28. [BLB]
Guttentag, R. E. (1984) The metal effort requirement of cumulative rehearsal: A developmental study. *Journal of Experimental Child Psychology* 37:92–106. [aNC, rNC]
Halford, G. S. (1993) *Children's understanding: The development of mental models.* Erlbaum. [SM, JPL]
Halford, G. S., Mayberry, M. T. & Bain, J. D. (1988) Set-size effects in primary memory: An age-related capacity limitation? *Memory and Cognition* 16:480–87. [aNC]
Halford, G. S., Wilson, W. H. & Phillips, S. (1998) Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences* 21:723–802. [aNC, GSH, JPL, rNC]
Hamilton, W. (1859) *Lectures on metaphysics and logic, vol. 1.* Blackwood. [aNC]
Hatano, G. & Osawa, K. (1983) Digit memory of grand experts in abacus-derived mental calculation. *Cognition* 15:95–110. [KAE]
Hausman, A. & Wilson, F. (1967). Goodman's ontology. In: *Carnap and Goodman: Two formalists,* ed. A. Hausman & F. Wilson, Iowa Publications in Philosophy, vol. 3. Martinus Nijhoff. [JPL]
Hazeltine, E., Ivry, R. & Teague, D. (2000) Dual-task performance with minimal cost: Evidence for concurrent response selection processes. *Journal of Cognitive Neuroscience* supplement:110. [TS]
He, S., Cavanaugh, P. & Intriligator, J. (1997) Many objects can be enumerated in afterimages when they are not crowded [Abstract]. *Abstracts of the Psychonomic Society* 2:3. [BJS, rNC]
Healy, A. F. & Cunningham, T. F. (1995) Very rapid forgetting: Reply to Muter. *Memory and Cognition* 23:387–92. [PM]
Hebb, D. O. (1949) *The organization of behavior.* Wiley. [MU]
Hecht, S. A. (1998) Toward an information processing account of individual differences in fraction skills. *Journal of Educational Psychology* 90:545–559. [SAH]
Hecht, S. A., Torgesen, J. K., Wagner, R. K. & Rashotte, C. A. (in press) The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second- to fifth-grade. *Journal of Experimental Child Psychology.* [SAH]
Hellige, J. B. & Cox, P. J. (1976) Effects of concurrent verbal memory on recognition of stimuli from the left and right visual fields. *Journal of Experimental Psychology: Human Perception and Performance* 2:210–221. [SM]
Henderson, L. (1972) Spatial and verbal codes and the capacity of STM. *Quarterly Journal of Experimental Psychology* 24:485–95. [aNC]
Henry, L. (1991) The effects of word length and phonemic similarity in young children's short-term memory. *Quarterly Journal of Experimental Psychology* 43A:35–52. [SM]
Henry, L. (1994) The relationship between speech rate and memory span in children. *International Journal of Behavioral Development* 17:37–56. [SM]
Henson, R. N. A. (1999) Positional information in short-term memory: Relative or absolute? *Memory and Cognition* 27:915–27. [rNC]
Herbart, J. F. (1890) Psychologie als Wissenschaft [Psychology as science]. In: *Jon. Fr. Herbart's saemtliche Werke in chronologischer Reihenfolge* [Jon. Fr. Herbart's collected works in their chronological order of publication], *part 1, vol. 5,* ed. K. Kehrbach & O. Flügel. Herman Beyer und Soehne. [DJM]
Hitch, G. J., Burgess, N., Towse, J. N. & Culpin, V. (1996) Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology* 49A:116–39. [aNC]
Hobson, J. A. (1997) Consciousness as a state-dependent phenomenon. In: *Scientific approaches to consciousness: The Twenty-eighth Carnegie Symposium on Cognition,* ed. J. D. Cohen & J. W. Schooler. Erlbaum. [BJB]
Holtzman, J. D. & Gazzaniga, M. S. (1982) Dual task interactions due exclusively to limits in processing resources. *Science* 218:1325–27. [aNC]
Holyoak, K. J. & Thagard, P. (1995) *Mental leaps: Analogy in creative thought.* Bradford. [GSH]
Horn D. & Opher I. (1996) Temporal segmentation in a neural dynamic system. *Neural Computation* 8:373–88. [MU]
Horn, D. & Usher, M. (1991) Parallel activation of memories in an oscillatory neural network. *Neural Computation* 3:31–43. [MU]
(1992) An oscillatory model for short term memory. In: *Advances in neural information processes systems 4,* ed. J. E. Moody, S. J. Hanson & P. R. Lippmann. Morgan and Kaufmann. [MU]
Horowitz, T. S. & Wolfe, J. M. (1998) Visual search has no memory. *Nature* 394:575–77. [GFW]
Houghton, G. & Tipper, S. P. (1994) A model of inhibitory mechanisms in selective attention. In: *Inhibitory mechanisms in attention, memory, and language,* ed. D. Dagenbach & T. Carr. Academic Press. [SM]
Hulme, C., Maugham, S. & Brown, G. D. (1991) Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short term memory span. *Journal of Memory and Language* 30:685–701. [SM]
Hulme, C & Tordoff, V. (1989) Working memory development: The effects of speech rate, word length and acoustic similarity on serial recall. *Journal of Experimental Child Psychology* 47:72–87. [SM]
Hultsch, H. (1991) Early experience can modify singing styles: Evidence from experiments with nightingales, *Luscinia megarhynchos. Animal Behavior* 42:883–89. [DT]
(1992) Time window and unit capacity: Dual constraints on the acquisition of serial information in songbirds. *Journal of Comparative Physiology A* 170:275–80. [DT]

(1993) Tracing the memory mechanisms in the song acquisition of birds. *Netherlands Journal of Zoology* 43:155–71. [DT]

Hultsch, H., Mundry, R. & Todt, D. (1998) Learning, representation and retrieval of rule-related knowledge in the song system of birds. In: *Learning: Rule extraction and representation,* ed. A. Friederici & R. Menzel. Walter de Gruyter. [DT]

Hultsch, H. Schleuss, F. & Todt, D. (1999) Auditory-visual stimulus pairing enhances perceptual learning in a songbird. *Animal Behavior* 58:1143–49. [DT]

Hultsch, H. & Todt, D. (1989) Memorization and reproduction of songs in nightingales (*Luscinia megarhynchos*): Evidence for package formation. *Journal of Comparative Physiology A* 165:197–203. [DT]

Hummel, J. E. & Holyoak, K. J. (1997) Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review* 104:427–66. [aNC]

Humphreys, G. W. (1998) Neural representation of objects in space: A dual coding account. *Philosophical Transactions of the Royal Society of London B* 353:1341–51. [GD]

Irwin, D. E. (1992) Memory for positions and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18:307–17. [WS]

Irwin, D. E. & Yeomans, J. M. (1986) Sensory registration and informational persistence. *Journal of Experimental Psychology: Human Perception and Performance* 12:343–60. [rNC]

Jacoby, L. L. & Bartz, W. H. (1972) Encoding processes and the negative recency effect. *Journal of Verbal Learning and Verbal Behavior* 11:561–65. [PM]

Jacoby, L. L., Woloshyn, V. & Kelley, C. (1989) Becoming famous without being recognized: Unconscious influences of memory produced by divided attention. *Journal of Experimental Psychology: General* 118:115–25. [aNC]

Jahnke, J. C., Davis, S. T. & Bower, R. E. (1989) Position and order: Information in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition* 15:859–67. [SEA]

James, W. (1890) *The principles of psychology.* Henry Holt. [aNC]

Jarvella, R. J. (1971) Syntactic processing of connected speech. *Journal of Verbal Language and Verbal Behavior* 10:409–16. [EMP]

Javitt, D. C., Strous, R., Grochowski, S., Ritter, W. & Cowan, N. (1997) Impaired precision, but normal retention, of auditory sensory (Aechoic@) memory information in schizophrenia. *Journal of Abnormal Psychology* 106:315–24. [rNC]

Jensen, E. M., Reese, E. P. & Reese, T. W. (1950) The subitizing and counting of visually presented fields of dots. *Journal of Psychology* 30:363–92. [JJ]

Jensen, O. & Lisman, J. E. (1996) Novel lists of 7 ± 2 known items can be reliably stored in an oscillatory short-term memory network: Interaction with long-term memory. *Learning and Memory* 3:257–63. [OJ]

(1998) An oscillatory short-term memory buffer model can account for data on the Sternberg task. *Journal of Neuroscience* 18:10,688–10,699. [OJ]

Jensen, O. & Tesche, C. D. (2000) Frontal midline activity in the theta band (6–8 Hz) increases with memory load in a short-term memory task: A parametric MEG study. European Research Conference, May 2000, Grenada. [OJ]

Jevons, W. S. (1871) The power of numerical discrimination. *Nature* 3:281–82. [aNC]

Jiang, Y., Olson, I. R. & Chun, M. M. (2000) Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26:683–702. [KAE]

Jinks, A. & Laing, D. G. (1999) Temporal processing reveals a mechanism for limiting the capacity of humans to analyze odor mixtures. *Cognitive Brain Research* 8:311–25. [rNC]

Johnson, J., Fabian, V. & Pascual-Leone, J. (1989) Quantitative hardware-stages that constrain language development. *Human Development* 32:245–71. [JPL]

Johnston, J. C., McCann, R. S., & Remington, R. W. (1995) Chronometric evidence for two types of attention. *Psychological Science* 6:365–69. [GFW]

Jones, D., Farrand, P., Stuart, G. & Morris, N. (1995) Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21:1008–18. [aNC]

Jou, J. (1998) Two retrieving modes in memory of the U.S. state names. Unpublished data. [JJ]

Jou, J. & Aldridge, J. W. (1999) Memory representation of alphabetic position and interval information. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25:680–701. [JJ]

Just, M. A., Carpenter, P. A. & Hemphill, D. D. (1996) Constraints on processing capacity: Architectural or implementational? In: *Mind matters: A tribute to Allen Newell,* ed. D. M. Steier & T. M. Mitchell. Erlbaum. [CPB]

Kareev, Y. (1995) Through a narrow window: Working memory capacity and the detection of covariation. *Cognition* 56:263–69. [aNC, SAH]

(2000) Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review* 107:397–402. [rNC]

Kareev, Y., Lieberman, I. & Lev, M. (1997) Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General* 126:278–87. [aNC]

Kaufman, E. L., Lord, M. W., Reese, T. W. & Volkmann, J. (1949) The discrimination of visual number. *American Journal of Psychology* 62:498–525. [aNC, JJ]

Kawai, N. (in press) Ordering and planning in sequential responding to arabic numbers by a chimpanzee. *Psychologia.* [NK]

Kawai, N. & Matsuzawa, T. (2000a) A conventional approach to chimpanzee cognition: Response to M. D. Hauser. *Trends in Cognitive Science* 4:128–29. [NK]

(2000b) Numerical memory span in a chimpanzee. *Nature* 403:39–40. [NK, rNC]

(2001) Reproductive memory processes in chimpanzees: Homologous approaches to research on human working memory. In: *Primate origins of human cognition and behavior,* ed. T. Matsuzawa. Springer-Verlag. [NK]

Keller, T. A. & Cowan, N. (1994) Developmental increase in the duration of memory for tone pitch. *Developmental Psychology* 30:855–63. [rNC]

Keller, T. A., Cowan, N. & Saults, J. S. (1995) Can auditory memory for tone pitch be rehearsed? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21:635–45. [rNC]

Killeen, P. R. & Weiss, N. A. (1987) Optimal timing and the Weber function. *Psychological Review* 94:455–68. [SG]

Kintsch, W. & van Dijk, T. A. (1978) Toward a model of text comprehension and production. *Psychological Review* 85:363–94. [aNC]

Kirschfeld, K. (1992) Oscillations in the insect brain: Do they correspond to theoretical (γ-waves of vertebrates? *Proceedings of the National Academy of Sciences* 89:4764–68. [aNC]

Klahr, D. (1973) Quantification processes. In: *Visual information processing,* ed. W. G. Chase. Academic Press. [JJ]

Klapp, S. T. & Netick, A. (1988) Multiple resources for processing and storage in short-term working memory. *Human Factors* 30:617–32. [aNC]

Kleinberg, J. & Kaufman, H. (1971) Constancy for short-term memory: Bits and chunks. *Journal of Experimental Psychology* 90:326–33. [rNC]

Kobatake, E. & Tanaka, K. (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology* 71:856–67. [PRR}

Koehler, O. (1954) Vorbedingungen und Vorstufen unserer Sprache bei Tieren. *Zoologischer Anzeiger Supplement* 18:327–41. [DT]

Kolb, B. & Whishaw, I. Q. (1996) *Fundamentals of human neuropsychology.* W. H. Freeman and Company. [JPL]

Konishi, M. (1989) Bird song for neurobiologists. *Neuron* 3:541–49. [DT]

Kroodsma, D. E. & Miller, E. H. (1996) *Ecology and evolution of acoustic communication in birds.* Cornell University Press. [DT]

Laming, D. (1992) Analysis of short-term retention: Models for Brown-Peterson experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18:1342–65. [PM]

Lane, P. C. R., Gobet, F. & Cheng, P. C.-H. (2000a) Learning perceptual chunks in a computational model of problem solving with diagrams. In: *Proceedings of the Third International Conference on Cognitive Modeling.* Universal Press. [PCL]

(2000b) Learning-based constraints on schemata. In: *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society.* Erlbaum. [PCL]

LaPointe, L. B. & Engle, R. W. (1990) Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16:1118–33. [aNC]

Lavie, N. & Driver, J. (1996) On the spatial extent of attention in object-based visual selection. *Perception and Psychophysics* 58:1238–51. [GD]

Lee, C. L. & Estes, W. K. (1981) Item and order information in short-term memory: Evidence for multilevel perturbation processes. *Journal of Experimental Psychology: Human Learning and Memory* 7:149–69. [rNC]

Leventhal, A. G., Thompson, K. F., Liu, D., Zhou, Y. & Ault, S. J. (1995) Concomitant sensitivity to orientation, direction, and color of cells in layers 2, 3, and 4 of monkey striate cortex. *Journal of Neuroscience* 15:1808–18. [PRR]

Levy, B. A. (1971) Role of articulation in auditory and visual short-term memory. *Journal of Verbal Learning and Verbal Behavior* 10:123–132.

Lewicki, P., Czyzewska, M. & Hoffman, H. (1987) Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology* 13:523–30. [aNC]

Libet, B., Alberts, W. W., Wright, E. W., Delattre, L., Levin, G. & Feinstein, B. (1964) Production of threshold levels of conscious sensation by electrical stimulation of human somatosensory cortex. *Journal of Neurophysiology* 27:546–78. [RDO]

Lisman, J. E. & Idiart, M. A. P. (1995) Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science* 267:1512–15. [aNC, OJ, PMM, AR, PRR, HT, MU, rNC]

Livingstone, M. S. (1996) Oscillatory firing and interneuronal correlations in squirrel monkey striate cortex. *Journal of Neurophysiology* 75:2467–85. [PRR]
Logan, G. D. (1988) Toward an instance theory of automatization. *Psychological Review* 95:492–527. [aNC, SAH]
Logan, G. D. & Klapp, S. T. (1991) Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17:179–95. [aNC]
Logie, R. H. & Baddeley, A. D. (1987) Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 13:310–26. [aNC, JJ]
Logie, R. H., Gilhooly, K. J. & Wynn, V. (1994) Counting on working memory in arithmetic problem solving. *Memory and Cognition* 22:395–410. [aNC]
Longoni, A. M., Richardson, J. T. E. & Aiello, A. (1993) Articulatory rehearsal and phonological storage in working memory. *Memory and Cognition* 21:11–22. [aNC]
Lovett, M. C., Reder, L. M. & Lebiere, C. (1997) Modeling individual differences in a digit working memory task. In: *Proceedings of the Nineteenth Conference of the Cognitive Science Society.* Erlbaum. [NAT]
Luck, S. J. & Vogel, E. K. (1997) The capacity of visual working memory for features and conjunctions. *Nature* 390:279–81. [aNC, SAH, AR, PRR, WS, rNC]
(1998) Response to visual and auditory working memory capacity by N. Cowan. *Trends in Cognitive Sciences* 2:78–80. [aNC]
MacGregor, J. N. (1987) Short-term memory capacity: Limitation or optimization? *Psychological Review* 94:107–108. [aNC]
Mahoney, J. V. & Ullman, S. (1988) Image chunking defining spatial building blocks for scene analysis. In: *Computational processes in human vision: An interdisciplinary perspective,* ed. Z. Pylyshyn. Intellect. [PRR]
Mandler, G. (1967) Organization and memory. In: *The psychology of learning and motivation, vol. 1,* ed. K. W. Spence & J. T. Spence. Academic Press. [aNC]
(1975) Memory storage and retrieval: Some limits on the reach of attention and consciousness. In: *Attention and performance V,* ed. P. M. A. Rabbit & S. Dornic. Academic Press. [aNC]
(1985) *Cognitive psychology: An essay in cognitive science.* Erlbaum. [aNC]
Mandler, G. & Shebo, B. J. (1982) Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General* 111:1–22. [aNC, BLB, JJ, rNC]
March, L., Cienfuegos, A., Goldbloom, L., Ritter, W., Cowan, N. & Javitt, D. C. (1999) Normal time course of auditory recognition in schizophrenia, despite impaired precision of the auditory sensory (Aechoic@) memory code. *Journal of Abnormal Psychology* 108:69–75. [rNC]
Marler, P. (1976) Sensory templates in species-specific behavior. In: *Simpler networks and behavior,* ed J. C. Fentress. Sinauer Associates. [DT]
(1991) Differences in behavioural development in closely related species: Bird song. In: *The development and integration of behaviour,* ed P. Bateson. Cambridge University Press. [DT]
Marr, D. (1982) *Vision.* Freeman. [rNC]
Marsh, R. L., Sebrechts, M. M., Hicks, J. L. & Landau, J. D. (1997) Processing strategies and secondary memory in very rapid forgetting. *Memory and Cognition* 25:173–81. [PM]
Martin, M. (1978) Memory span as a measure of individual differences in memory capacity. *Memory and Cognition* 6:194–98. [SEA]
(1980) Attention to words in different modalities: Four-channel presentation with physical and semantic selection. *Acta Psychologica* 44:99–115. [aNC]
Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review* 79:124–45. [rNC]
Matsuzawa, T. (1985) Use of numbers by a chimpanzee. *Nature* 315:57–59. [NK]
Maunsell, J. H. R. (1995) The brain's visual world: Representation of visual targets in cerebral cortex. *Science* 270:764–69. [PRR]
May, P., Tiitinen, H., Ilmoniemi, R. J., Nyman, G., Taylor, J. G. & Näätänen, R. (1999) Frequency change detection in human auditory cortex. *Journal of Computational Neuroscience* 6:97–118. [HT]
McCann, R. & Johnston, J. C. (1992) Locus of the single-channel bottleneck in dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance* 18:471–64. [TS]
McElree, B. (1996) Accessing short-term memory with semantic and phonological information: A time-course analysis. *Memory and Cognition* 24:173–87. [BM]
(1997) Working memory and the focus of attention. Paper presented at the Psychonomic Society Conference, Philadelphia, Pa, November. [BM]
(1998) Attended and nonattended states in working memory: Accessing categorized structures. *Journal of Memory and Language* 38:225–52. [BM, rNC]
(in press) Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* [BM]
McElree, B. & Dosher, B. A. (1989) Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General* 118:346–73. [BM]
(1993) Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General* 122:291–315. [BM]
McGeoch, J. A. (1932) Forgetting and the law of disuse. *Psychological Review* 39:352–70. [aNC]
McKone, E. (1995) Short-term implicit memory for words and nonwords. *Journal of Experimental Psychology: Learning, Memory and Cognition* 21:1108–26. [aNC, SEA, EM]
McKone, E. & Dennis, C. (2000) Short-term implicit memory: Visual, auditory, and cross-modality priming. *Psychonomic Bulletin and Review* 7:341–46. [EM]
McLean, R. S. & Gregg, L. W. (1967) Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology* 74:455–59. [aNC]
Melton, A. W. (1963) Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior* 2:1–21. [aNC]
Mewhort, D. J. K., Campbell, A. J., Marchetti, F. M. & Campbell, J. I. D. (1981) Identification, localization, and "iconic memory": An evaluation of the bar-probe task. *Memory and Cognition* 9:50–67. [rNC]
Meyer, D. E. & Kieras, D. E. (1997) A computational theory of executive processes and multiple-task performance: Part 1. Basic mechanisms. *Psychological Review* 104:3–65. [aNC, TS, rNC]
Michon, J. (1978) The making of the present: A tutorial review. In: *Attention and Performance VII,* ed. J. Requin. Erlbaum. [SG]
Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63:81–97. [aNC, SEA, BLB, CPB, GG, SG, JJ, BJS, HT, NAT, rNC]
Milner, P. M. (1974) A model for visual shape recognition. *Psychological Review* 81:521–35. [aNC]
(1999) *The autonomous brain.* Erlbaum. [PMM]
Miltner, W. H. R., Braun, C., Arnold, M., Witte, H. & Taub, E. (1999) Coherence of gamma-band EEG activity as a basis for associative learning. *Nature* 397:434–36. [aNC]
Miyake, A. & Shah, P. (1999) *Models of working memory: Mechanisms of active maintenance and executive control.* Cambridge University Press. [ADB]
Miyashita, Y. (1993) Inferior temporal cortex: Where visual perception meets memory. *Annual Review of Neuroscience* 16:245–63. [PMM]
Monsell, S. (1978) Recency, immediate recognition memory, and reaction time. *Cognitive Psychology* 10:465–501. [rNC]
Moran, J. & Desimone, R. (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229:782–84. [PMM]
Morra, S. (1989) Untitled presentation given at the Workshop on Working Memory, Schonloo, The Netherlands, April 21–24. [SM]
(1990) Why another model of working memory? Paper presented at the Fourth Conference of the European Society for Cognitive Psychology, Como, Italy, September 15–19. [SM]
(1994) Issues in working memory measurement: Testing for M capacity. *International Journal of Behavioral Development* 17:143–59. [SM, JPL]
(2000) A new model of verbal short-term memory. *Journal of Experimental Child Psychology* 75:191–227. [SM, JPL]
Morra, S., Mazzoni, G. & Sava, D. (1993) Esiste un loop articolatorio di capacità limitata? [Is there a time-limited articulatory loop?] Proceedings of the SIPs Research Division Conference, Rome, September 29-October 2, 94–95. [SM]
Morra, S., Moizo, C. & Scopesi, A. (1988) Working memory (of the M. operator) and the planning of children's drawings. *Journal of Experimental Child Psychology* 46:41–73. [SM]
Morra, S., Pascual-Leone, J., Johnson, J. & Baillargeon, R. (1991) Understanding spatial descriptions: Experimental test of a mental capacity model. In: *Mental images in human cognition,* ed. R. H. Logie & M. Denis. North Holland. [SM]
Morra, S. & Stoffel, C. (1991) A new model of verbal short term memory and its development. Paper presented at the International Conference on Memory, Lancaster, U. K., July 15–19. [SM]
Morra, S., Vigliocco, G. & Penello, B. (in press) M capacity as a lifespan construct: A study of its decrease in aging subjects. *International Journal of Behavioral Development.* [SM]
Moscovitch, M. & Umiltà, C. (1990) Modularity and neuropsychology: Implications for the organization of attention and memory in normal and brain-damaged people. In: *Modular processes in dementia,* ed. M. Schwartz. MIT/Bradford. [SM]
Murdoch Jr., B. B. & Hockley, W. E. (1989) Short-term memory for associations. In: *The psychology of learning and motivation, vol. 24,* ed. G. H. Bower. Academic Press. [PM]
Murray, D. J. (1968) Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology* 78:679–84. [aNC]
(1995) *Gestalt psychology and the cognitive revolution.* Harvester Wheatsheaf. [DJM]
Murray, D. J., Boudreau, N., Burggraf, K. K., Dobell, L., Guger, S. L., Leask, A.,

Stanford, L., Tate, T. L. & Wheeler, M. (1999) A grouping interpretation of the modality effect in immediate probed recognition. *Memory and Cognition* 27:234–45. [DJM]
Murray, D. J., Burhop, J., Centa, S., Chande, N., Oinonen, K., Thomas, T., Wilkie, T. & Farahmand, B. (1998) A partial matching theory of the mirror effect in immediate probed recognition. *Memory and Cognition* 26:1196–1213. [DJM]
Muter, P. (1980) Very rapid forgetting. *Memory and Cognition* 8:174–79. [PM]
(1995) Very rapid forgetting: Reply to Cunningham, Healy, Till, Fendrich and Dimitry. *Memory and Cognition* 23:383–86. [PM]
Nairne, J. S. (1990) A feature model of immediate memory. *Memory and Cognition* 18:251–69. [rNC]
(1991) Positional uncertainty in long-term memory. *Memory and Cognition* 19:332–40. [aNC, JJ, rNC]
(1992) The loss of positional certainty in long-term memory. *Psychological Science* 3:199–202. [aNC, JJ]
Nairne, J. S., Neath, I. & Serra, M. (1997) Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin and Review* 4:541–45. [SM]
Neath, I. (1998) *Human memory: An introduction to research, data, and theory.* Brooks/Cole. [aNC]
Neath, I. & Nairne, J. S. (1995) Word-length effects in immediate memory: Overwriting trace decay. *Psychonomic Bulletin and Review* 2:429–41. [aNC, DJM, rNC]
Neely, J. H. (1991) Semantic priming effects in visual word recognition: A selective review of current findings and theories. In: *Basic processes in reading: Visual word recognition,* ed. D. Besner & G. W. Humphreys. Erlbaum. [SM]
Neumann, O. (1987) Beyond capacity: A functional view of attention. In: *Perspectives on perception and action,* ed. H. Heuer & A. Sanders. Erlbaum. [TS]
Newman, J. B., Baars, B. J. & Cho, S.-B. (1997) A neural Global Workspace model for conscious attention. *Neural Networks* 10:1195–1206. [BJB]
Newport, E. L. (1988) Constraints on learning and their role in language acquisition: Studies of the acquisition of the American Sign Language. *Language Sciences* 10:147–72. [EMP]
(1990) Maturational constraints on language learning. *Cognitive Science* 14:11–28. [aNC, EMP]
Nicolson, R. & Fawcett, A. (1991) Decomposing working memory: New evidence from memory span. Paper presented at the International Conference on Memory, Lancaster, U. K., July 15–19. [SM]
Nissen, M. J. & Bullemer, P. (1987) Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology* 19:1–32. [aNC]
Norman, D. A. & Bobrow, D. G. (1975) On data-limited and resource-limited processes. *Cognitive Psychology* 7:44–64. [rNC]
Norman, D. A. & Shallice, T. (1986) Attention to action: Willed and automatic control of behavior. In: *Consciousness and self-regulation, vol. 4,* ed. R. J. Davidson, G. E. Schwartz & D. Shapiro. Plenum. [TS]
Nottebohm, F. (1993) The search for neural mechanisms that define the sensitive period for song learning in birds. *Netherlands Journal of Zoology* 43:193–234. [DT]
Oades, R. D., Dittmann-Balcar, A., Zerbin, D. & Grzella, I. (1997) Impaired attention-dependent augmentation of MMN in nonparanoid versus paranoid schizophrenic patients: A comparison with obsessive-compulsive disorder and healthy subjects. *Biological Psychiatry* 41:1196–1210. [RDO]
Oades, R. D., Zerbin, D., Dittmann-Balcar, A. & Eggers, C. (1996) Auditory event-related potential (ERP) and difference-wave topography in schizophrenic patients with/without active hallucinations and delusions: A comparison with young obsessive-compulsive disorder (OCD) and healthy subjects. *International Journal of Psychophysiology* 22:185–214. [RDO]
Oberauer, K. (in press) Removing irrelevant information from working memory: A cognitive aging study with the modified Sternberg task. *Experimental Psychology: Learning, Memory and Cognition.* [rNC]
Okada, R. & Burrows, D. (1978) The effects of subsidiary tasks on memory retrieval from long and short lists. *The Quarterly Journal of Experimental Psychology* 30:221–33. [JJ]
Oram, M. W. & Perrett, D. I. (1992) Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology* 68:70–84. [PRR]
Paivio, A., Yuille, J. C. & Madigan, S. A. (1969) Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monograph Supplement* 76:1–25. [JSN]
Pascual-Leone, J. (1970) A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica* 63:301–45. [SM, JPL]
(1987) Organismic processes for neo-Piagetian theories: A dialectical causal account of cognitive development. *International Journal of Psychology* 22:531–70. [SM, JPL]
(1995) Learning and development as dialectical factors in cognitive growth. *Human Development* 38:338–48. [JPL]
(1998) To appraise developmental difficulty or mental demand, relational complexity is not enough. *Behavioral and Brain Sciences* 12:843–44. [JPL]
(in press) Reflections on working memory: Are the two models complementary? *Journal of Experimental Child Psychology.* [JPL]
Pascual-Leone, J. & Baillargeon, R. (1994) Developmental measurement of mental attention. *International Journal of Behavioral Development* 17:161–200. [SM, JPL]
Pascual-Leone, J. & Johnson, J. (1991) The psychological unit and its role in task analysis. In: *Criteria for competence: Controversies in the assessment of children's ablities,* ed. M. Chandler & M. Chapman. Erlbaum. [SM]
Pascual-Leone, J., Johnson, J. & Morris, S. (in preparation) A quantitative model of mental attention in a mental span task. [JPL]
Pashler, H. (1988) Familiarity and visual change detection. *Perception and Psychophysics* 44:369–78. [aNC, rNC]
(1991) Shifting visual attention and selecting motor responses: Distinct attentional mechanism. *Journal of Experimental Psychology: Human Perception and Performance* 17:1023–40. [GFW]
(1994) Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin* 16:220–44. [TS, GFW]
Penney, C. G. (1980) Order of report in bisensory verbal short-term memory. *Canadian Journal of Psychology* 34:190–95. [aNC]
Pepperberg, I. M. (1993) A review of the effects of social interaction on vocal learning in African grey parrots (*Psittacus erithacus*). *Netherlands Journal of Zoology* 43:104–24. [DT]
Peterson, L. R. & Johnson, S. T. (1971) Some effects of minimizing articulation on short-term retention. *Journal of Verbal Learning and Verbal Behavior* 10:346–54. [aNC]
Peterson, L. R. & Peterson, M. J. (1959) Short-term retention of individual verbal items. *Journal of Experimental Psychology* 58:193–98. [aNC, PM]
Peterson, S. A. & Simon, T. J. (2000) Computational evidence for the subitizing phenomenon as an emergent property of the human cognitive architecture. *Cognitive Science* 24:93–122. [NAT, rNC]
Petrides, M. (1996) Lateral frontal cortical contribution to memory. *Seminars in the Neurosciences* 8:57–63. [BR]
Pinker, S. (1979) Formal models of language learning. *Cognition* 7:217–83. [EMP]
(1994) *The language instinct.* Penguin Books. [EMP]
Platt, J. R. (1964) Strong inference. *Science* 146:347–53. [rNC]
Pollack, I. (1952) The information of elementary auditory displays. *Journal of the Acoustical Society of America* 24:745–49. [BLB]
Pollack, I. & Ficks, L. (1954) Information of elementary multi-dimensional auditory displays. *Journal of the Acoustical Society of America* 26:155–58. [BLB]
Pollack, I., Johnson, I. B. & Knaff, P. R. (1959) Running memory span. *Journal of Experimental Psychology* 57:137–46. [aNC, rNC]
Posner, M. I. (1969) Abstraction and the process of recognition. In: *Psychology of learning and motivation, vol. 3,* ed. G. H. Bower & J. T. Spence. Academic Press. [aNC]
Posner, M. L. & Konick, A. W. (1966) On the role of interference in short-term retention. *Journal of Experimental Psychology* 72:221–31. [PM]
Posner, M., Snyder, C. & Davidson, B. (1980) Attention and detection of signals. *Journal of Experimental Psychology: General* 109:160–74. [aNC]
Postman, L. & Phillips, L. W. (1965) Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology* 17:132–38. [aNC]
Pothos, E. M. (1998) Aspects of generalisation. Unpublished PhD thesis, University of Oxford. [EMP]
Potter, M. C. (1976) Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory* 2:509–22. [GFW]
Poulton, E. C. (1954) Eye-hand span in simple serial tasks. *Journal of Experimental Psychology* 47:403–10. [aNC]
Prabhakaran, V., Narayanan, K., Zhao, Z. & Gabrieli, J. D. E. (2000) Integration of diverse information in working memory within the frontal lobe. *Nature-Neuroscience* 3:85–90. [BR]
Pylyshyn, Z. (1989) The role of location indices in spatial perception: A sketch of the FINST spatial-index model. *Cognition* 32:65–97. [GD]
Pylyshyn, Z., Burkell, J., Fisher, B., Sears, C., Schmidt, W. & Trick, L. (1994) Multiple parallel access in visual attention. *Canadian Journal of Experimental Psychology* 48:260–83. [aNC]
Pylyshyn, Z. W. & Storm, R. W. (1988) Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision* 3:179–97. [aNC, BJS]
Raaijmakers, J. G. W. & Shiffrin, R. M. (1981) Search of associative memory. *Psychological Review* 88:93–134. [aNC]
Rafal, R. D. (1997) Balint syndrome. In: *Behavioral neurology and neuropsychology,* ed. T. Feinberg & M. Farrah. McGraw-Hill. [BJS]
Raffone, A. & Wolters, G. (in press) A cortical mechanism for binding in visual working memory. *Journal of Cognitive Neuroscience.* [AR]
Raghavachari, S., Caplan, J. B., Kirschen, M., Kahana, M., Madsen, J. R. & Lisman, J. E. (1999) The Sternberg task evokes theta oscillations in human intracranial recordings. *Society for Neuroscience Abstracts* 25:1143. [OJ]
Rainer, G., Asaad, W. F. & Miller, E. K. (1998) Selective representation of relevant

information by neurons in the primate prefrontal cortex. *Nature* 393:577–79. [PRR]

Ratcliff, R. (1978) A theory of memory retrieval. *Psychological Review* 85:59–108. [rNC]

Reber, P. J. & Kotovsky, K. (1997) Implicit learning in problem solving: The role of working memory capacity. *Journal of Experimental Psychology: General* 126:178–203. [aNC]

Reisberg, D., Rappaport, I. & O'Shaughnessy, M. (1984) Limits of working memory: The digit-digit span. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10:203–21. [aNC]

Rensink, R. A. (1999) The magical number one, plus or minus zero. *Investigative Ophthalmology and Visual Science* 40:52. [RAR]

(2000a) Visual search for change: A probe into the nature of attentional processing. *Visual Cognition* 7:345–76. [RAR, BJS]

(2000b) Seeing, sensing, and scrutinizing. *Vision Research* 40:1469–87. [RAR]

(2000c) Differential grouping of features. *Perception* 29 supplement. [RAR]

Rensink, R. A., Deubel, H. & Scheneider, W. X. (2000) The incredible shrinking span: Estimates of memory capacity depend on interstimulus interval. *Investigative Ophthalmology and Visual Science* 41:425. [RAR]

Rensink, R. A., O'Regan, J. K. & Clark, J. J. (1997) To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science* 8:368–73. [aNC, RAR]

Richman, H. B., Gobet, F., Staszewski, J. J. & Simon, H. A. (1996) Perceptual and memory processes in the acquisition of expert performance: The EPAM model. In: *The road to excellence,* ed. K. A. Ericsson. Erlbaum. [PCL]

Richman, H. B., Staszewski, J. J. & Simon, H. A. (1995) Simulation of expert memory using EPAM IV. *Psychological Review* 102:305–30. [aNC, KAE]

Robertson, L., Treisman, A., Friedman-Hill, S. & Grabowecky, M. (1997) The interaction of spatial and object pathways: Evidence from Balint's syndrome. *Journal of Cognitive Neuroscience* 9:295–317. [BJS]

Robin, N. & Holyoak, K. J. (1995) Relational complexity and the functions of prefrontal cortex. In: *The cognitive neurosciences,* ed. M. S. Gazzaniga. MIT Press. [JPL]

Rodriquez, E., George, N., Lachaux, J.-P., Martinerie, J., Renault, B. & Varela, F. J. (1999) Perception's shadow: long-distance synchronization of human brain activity. *Nature* 397:430–33. [aNC]

Roelfsema, P. R., Lamme, V. A. F. & Spekreijse, H. (2000) The implementation of visual routines. *Vision Research* 40:1385–1411. [PRR]

Roelfsema, P. R. & Singer, W. (1998) Detecting connectedness. *Cerebral Cortex* 8:385–96. [PRR]

Roitblat, H. L., Bever, T. G., Helweg, D. A. & Harley, H. E. (1991) On-line choice and the representation of serially structured stimuli. *Journal of Experimental Psychology: Animal Behavior Processes* 17:55–67. [DT]

Rubin, D. C. & Wenzel, A. E. (1996) One hundred years of forgetting: A quantitative description of retention. *Psychological Review* 103:734–60. [PM]

Ryan, J. (1969) Grouping and short-term memory: Different means and patterns of groups. *Quarterly Journal of Experimental Psychology* 21:137–47. [aNC]

Rypma, B., Prabhakaran, V., Desmond, J. E., Glover, G. H. & Gabrieli, J. D. E. (1999) Load-dependent roles of prefrontal cortical regions in the maintenance of working memory. *NeuroImage* 9:216–26. [BR]

Sanders, A. F. (1968) Short term memory for spatial positions. *Psychologie* 23:1–15. [aNC]

Sanders, A. F. & Schroots, J. J. F. (1969) Cognitive categories and memory span: III. Effects of similarity on recall. *Quarterly Journal of Experimental Psychology* 21:21–28. [aNC]

Saults, J. S. & Cowan, N. (1996) The development of memory for ignored speech. *Journal of Experimental Child Psychology* 63:239–61. [rNC]

Scarborough, D. L. (1971) Memory for brief visual displays: The role of implicit speech. Paper presented to the Eastern Psychological Association, New York, NY. [aNC]

Schneider, W. X. (1999) Visual-spatial working memory, attention, and scene representation: A neuro-cognitive theory. *Psychological Research* 62:220–36. [WS]

Schneider, W. & Detweiler, M. (1987) A connectionist/control architecture for working memory. In: *The psychology of learning motivation, vol. 21,* ed. G. H. Bower. Academic Press. [aNC]

Schneider, W. & Shiffrin, R. M. (1977) Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review* 84:1–66. [aNC, rNC]

Schneider, W. X., Wesenick, M. B., Deubal, H. & Bundesen, C. (1999) A study of visuospatial working memory. *Perception* 28:5. [RAR]

Scholl, B. J. (in press) Objects and attention: The state of the art. *Cognition: Special Issue on Objects and Attention.* [BJS]

Scholl, B. J. & Leslie, A. M. (1999) Explaining the infant's object concept: Beyond the perception/cognition dichotomy. In: *What is cognitive science?,* ed. E. Lepore & Z. Pylyshyn. Blackwell. [BJS]

Scholl, B. J. & Pylyshyn, Z. W. (1999) Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology* 38:259–90. [BJS]

Scholl, B. J., Pylyshyn, Z. W. & Feldman, J. (in press) What is a visual object? Evidence from multiple object tracking. *Cognition.* [BJS]

Schubert, T. (1999) Processing differences between choice and simple reaction tasks affect outcome of bottleneck localisation experiments. *Journal of Experimental Psychology: Human Perception and Performance* 25:408–25. [TS]

Schumacher, E., Seymour, T. L., Glass, J. M., Lauber, E. J., Kieras, D. E. & Meyer, D. E. (1998) Virtually perfect timesharing in dual-task performance: Behavioural evidence for independent parallel processing in the human brain. *Journal of Cognitive Neuroscience* supplement:106. [TS]

Schweickert, R. & Boruff, B. (1986) Short-term memory capacity: Magic number or magic spell? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12:419–25. [aNC, SM, JSN]

Schweickert, R., Hayt, C., Hersberger, L. & Guentert, L. (1996) How many words can working memory hold? A model and a method. In: *Models of short-term memory,* ed. S. E. Gathercole. Psychology Press. [aNC]

Schyns, P., Goldstone, R. L. & Thilbaut, J.-P. (1998) The development of features in object concepts. *Behavioral and Brain Sciences* 21:1–54. [RDO]

Sebrechts, M. M., Marsh, R. L. & Seamon, J. C. (1989) Secondary memory and very rapid forgetting. *Memory and Cognition* 17:693–700. [PM]

Service, E. (1998) The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology* 51A:283–304. [aNC, SM]

Shah, P. & Miyake, A. (1996) The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General* 125:4–27. [aNC]

Shallice, T. & Warrington, E. K. (1970) Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology* 22:261–73. [BJS]

Shastri, L. & Ajjanagadde, V. (1993) From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences* 16:417–94. [aNC, HT]

Shibuya, H. & Bundesen, C. (1988) Visual selection from multielement displays: Measuring and modeling effects of exposure duration. *Journal of Experimental Psychology: Human Perception and Performance* 14:160–68. [WS]

Shiffrin, R. M. (1993) Short-term memory: A brief commentary. *Memory and Cognition* 21:193–97. [aNC, rNC]

Shiffrin, R. M. & Schneider, W. (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review* 84:127–90. [aNC, rNC]

Simon, H. A. (1974) How big is a chunk? *Science* 183:482–88. [aNC]

Simon, T. J. & Vaishnavi, S. (1996) Subitizing and counting depend on different attentional mechanisms: Evidence from visual enumeration in afterimages. *Perception and Psychophysics* 58:915–26. [aNC, BJS]

Simons, D. J. (1996) In sight, out of mind: When object representations fail. *Psychological Science* 7:301–305. [RAR]

Simons, D. J. & Levin, D. T. (1998) Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review* 5:644–49. [aNC]

Sirevaag, E. J., Kramer, A. F., Coles, M. G. H. & Donchin, E. (1989) Resource reciprocity: An event-related brain potentials analysis. *Acta Psychologica* 70:77–97. [aNC]

Slak, S. (1970) Phonemic recoding of digital information. *Journal of Experimental Psychology* 86:398–406. [DJM]

Slemecka, N. J. (1985) Ebbinghaus: Some associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11:414–35. [JJ, rNC]

Smith, E. E. & Jonides, J. (1995) Working memory in humans: Neuropsychological Evidence. In: *The cognitive neurosciences,* ed. M. Gazzaniga. MIT Press. [ADB]

(1999) Storage and executive processes in the frontal lobes. *Science* 283:1657–61. [BR]

Solso, R. L. (1995) *Cognitive Psychology,* 4th ed. Allyn and Bacon. [PM]

Sperling, G. (1960) The information available in brief visual presentations. *Psychological Monographs* 74:Whole No. 498. [aNC, SAH, RDO, TS, rNC]

(1967) Successive approximations to a model for short-term memory. *Acta Psychologica* 27:285–92. [aNC]

Spitz, H. H. (1973) The channel capacity of educable mental retardates. In: *The experimental psychology of mental retardation,* ed. D. K. Routh. Aldine. [BLB]

Spitzer, M. (1997) A cognitive neuroscience view of schizophrenic thought disorder. *Schizophrenia Bulletin* 23:29–50. [RDO]

Stadler, M. (1989) On learning complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15:1061–69. [aNC]

Steele, C. M. & Josephs, R. A. (1990) Alcohol myopia: Its prized and dangerous effects. *American Psychologist* 45:921–33. [rNC]

Steriade, M., McCormick, D. A. & Sejnowski, R. J. (1993) Thalamocortical oscillations in the sleeping and aroused brain. *Science* 262:679–685. [BJB]

Sternberg, S. (1966) High-speed scanning in human memory. *Science* 153:652–54. [aNC, BJB, GG, JJ, DJM, rNC]
(1969a) The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica* 30:276–315. [BR]
(1969b) Memory scanning: Mental processes revealed reaction-time experiments. *American Scientist* 57:421–57. [JJ]
(1975) Memory scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology* 27:1–32. [JJ]
Stewart, L. & Pascual-Leone, J. (1992) Mental capacity constraints and the development of moral reasoning. *Journal of Experimental Child Psychology* 54:251–87. [JPL]
Stigler, J. W. (1984) "Mental abacus": The effect of abacus training on Chinese children's mental calculation. *Cognitive Psychology* 16:145–76. [KAE]
Straube, E. & Oades, R. D. (1992) *Schizophrenia: Empirical research and finding.* Academic Press. [RDO]
Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. (1999) Global and fine information encoded by single neurons in the temporal visual cortex. *Nature* 400:869–72. [RDO]
Taatgen, N. A. (1999a) Cognitief modelleren, een nieuwe kijk op individuele verschillen. *Nederlands tijdschrift voor de psychologie* 54:167–77. [NAT]
(1999b) A model of learning task-specific knowledge for a new task. In: *Proceedings of the Twenty-first Annual Meeting of the Cognitive Science Society.* Erlbaum. [NAT]
Tallon-Baudry, C., Kreiter, A. & Bertrand, O. (1999) Sustained and transient oscillatory responses in the gamma and beta bands in a visual short-term memory task in humans. *Visual Neuroscience* 16:449–59. [OJ]
Tan, L. & Ward, G. (in press) A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition.* [SEA]
Tehan, G. & Humphreys, M. S. (1996) Cuing effects in short-term recall. *Memory and Cognition* 24:719–32. [aNC, JSN, rNC]
Terrace, H. (1987) Chunking by a pigeon in a serial learning task. *Nature* 325:149–51. [DT]
Thompson, C. P., Cowan, T. M. & Frieman, J. (1993) *Memory search by a memorist.* Erlbaum. [DJM, JW]
Tiitinen, H. & May, P. (preprint) Auditory transient and sustained responses as a function of interstimulus interval. [HT]
Tiitinen, H., May, P., Reinikainen, K. & Näätänen, R. (1994) Attentive novelty detection in humans is governed by pre-attentive sensory memory. *Nature* 372:90–92. [HT]
Tiitinen, H., Sinkkonen, J., Reinikainen, K., Alho, K., Lavikainen, J. & Näätänen, R. (1993) Selective attention enhances the auditory 40-Hz transient response in humans. *Nature* 364:59–60. [aNC, HT]
Todt, D., Cirillo, J., Geberzahn, N. & Schleuss, F. (2000) The role of hierarchy levels in vocal imitations of songbirds. *Cybernetics and Systems: An International Journal* 32:1–27. [DT]
Tomonaga, M. & Matsuzawa, T. (2000) Sequential responding to arabic numbers with wild cards by the chimpanzee (*Pan troglodytes*). *Animal Cognition* 3:1–11. [NK]
Toms, M., Morris, N. & Ward, D. (1993) Working memory and conditional reasoning. *Quarterly Journal of Experimental Psychology* 46A:679–99. [aNC]
Tononi, G. & Edelman, G. M. (1998) Consciousness and complexity. *Science* 282:1846–51. [BJB]
Townsend, J. T. (1976) Serial and within-stage independent parallel model equivalence on the minimum completion time. *Journal of Mathematical Psychology* 14:219–38. [rNC]
Towse, J. N. & Houston-Price, C. M. T. (in press) Reflections on the concept of the central executive. In: *Working memory in perspective,* ed. J. Andrade. Psychology Press. [JNT]
Treisman, A. M. & Gelade, G. (1980) A feature integration theory of attention. *Cognitive Psychology* 12:97–136. [PRR, rNC]
Treisman, A. & Gormican, S. (1988) Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review* 95:15–48. [WS]
Trick, L. M. & Pylyshyn, Z. W. (1993) What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance* 19:331–51. [aNC, GD, TS]
(1994a) Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review* 101:80–102. [aNC, SEA, BJS]
(1994b) Cueing and counting: Does the position of the attentional focus affect enumeration? *Visual Cognition* 1:67–100. [aNC]
Tsodyks, M., Pawelzik, K. & Markram, H. (1998) Neural networks with dynamic synapses. *Neural Computation* 10:821–35. [MU]
Tulving, E. & Colotla, V. (1970) Free recall of trilingual lists. *Cognitive Psychology* 1:86–98. [aNC]
Tulving, E. & Patkau, J. E. (1962) Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canadian Journal of Psychology* 16:83–95. [aNC]
Tulving, E. & Patterson, R. D. (1968) Functional units and retrieval processes in free recall. *Journal of Experimental Psychology* 77:239–48. [aNC]
Tulving, E. & Pearlstone, Z. (1966) Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior* 5:381–91. [aNC]
Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information processing analysis of masking with patterned stimuli. *Psychological Review* 80:1–52. [rNC]
Uller, C., Carey, S., Huntley-Fenner, G. & Klatt, L. (1999) What representations might underlie infant numerical knowledge? *Cognitive Development* 14:1–43. [BJS]
Usher, M. & Cohen, J. (1999) Short term memory and selection processes in a frontal-lobe model. In: *Connectionist models in cognitive neuroscience,* ed. D. Heinke, G. W. Humphries & A. Olsen. Springer-Verlag. [MU]
Usher, M. & McClelland, J. L. (1995) *Time course of perpetual choice.* (Tech. Rep. No. PDP. CNS.95.5). Pittsburgh, PA: Carnegie Mellon University. [MU]
Vallar, G. & Baddeley, A. D. (1982) Short-term forgetting and the articulatory loop. *Quarterly Journal of Experimental Psychology* 34A:53–60. [aNC]
(1984) Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. *Journal of Verbal Learning and Verbal Behavior* 23:151–61. [BR]
Van Oeffelen, M. & Vos, P. (1982) A probabilistic model for the discrimination of visual number. *Perception and Psychophysics* 32:163–70. [SEA]
Van Strien, J. W. & Bouma, A. (1990) Selective activation effects on concurrent verbal and spatial memory loads in left-handed and right-handed adults. *Brain and Cognition* 14:81–91. [SM]
Varela, F. J. (1995) Resonant cell assemblies: A new approach to cognitive functions and neuronal synchrony. *Biological Research* 28:81–95. [AR]
Vogel, E. K. (2000) *Selective storage in visual working memory: Distinguishing between perceptual-level and working memory-level mechanisms of attention.* University of Iowa Press. [GFW]
Vogel, E. K., Luck, S. J. & Shapiro, K. L. (1998) Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance* 24:1656–74. [aNC, GFW]
Von der Malsburg, C. (1981) *The correlation theory of brain function.* (Internal Report 81–2). Göttingen, NL: Max Planck Institute for Biophysical Chemistry. [AR]
(1995) Binding in models of perception and brain function. *Current Opinion in Neurobiology* 5:520–26. [aNC, PRR]
Warren, R. M. & Obusek, C. J. (1972) Identification of temporal order within auditory sequences. *Perception and Psychophysics* 12:86–90. [CPB, rNC]
Warren, R. M., Obusek, C. J., Farmer, R. M. & Warren, R. P. (1969) Auditory sequence: Confusion of patterns other than speech or music. *Science* 164:586–87. [CPB]
Watkins, M. J. (1974) Concept and measurement of primary memory. *Psychological Bulletin* 81:695–711. [aNC]
Watkins, M. J. & Watkins, O. C. (1974) Processing of recency items for free recall. *Journal of Experimental Psychology* 102:488–93. [PM]
Watkins, O. C. & Watkins, M. J. (1975) Build-up of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory* 1:442–52. [aNC]
Watson, S. E. & Kramer, A. F. (1999) Object-based visual selective attention and perceptual organization. *Perception and Psychophysics* 61:31–49. [GD]
Waugh, N. C. & Norman, D. A. (1965) Primary memory. *Psychological Review* 72:89–104. [aNC, JJ, BR]
Wegner, D. M., Quillian, F. & Houston, C. E. (1996) Memories out of order: Thought suppression and the disturbance of sequence memory. *Journal of Personality and Social Psychology* 71:680–91. [CPB]
Welford, A. T. (1980) The single-channel hypothesis. In: *Reaction times,* ed. A. T. Welford. Academic Press. [TS]
Wharton, R. M. (1974) Approximate language identification. *Information and Control* 26:236–55. [EMP]
Wickelgren, W. A. (1964) Size of rehearsal group and short-term memory. *Journal of Experimental Psychology* 68:413–19. [aNC]
(1966) Phonemic similarity and interference in short-term memory for single letters. *Journal of Experimental Psychology* 71:396–404. [aNC]
Wickelgren, W. A., Corbett, A. T. & Dosher, B. A. (1980) Priming and retrieval from short-term memory: A speed-accuracy tradeoff analysis. *Journal of Verbal Learning and Verbal Behavior* 19:387–404. [BM]
Wickelgren, W. A. & Norman, D. A. (1966) Strength models and serial position in short-term memory. *Journal of Mathematical Psychology* 3:316–47. [DJM]
Wickens, C. D. (1984) Processing resources in attention. In: *Varieties of attention,* ed R. Parasuraman & D. R. Davies. Academic Press. [aNC]
Wickens, D. D., Moody, M. J. & Dow, R. (1981) The nature and timing of the

retrieval process and of interference effects. *Journal of Experimental Psychology: General* 110:1–20. [aNC]

Wilkes, A. L. (1975) Encoding processes and pausing behaviour. In: *Studies in long-term memory,* ed. A. Kennedy & A. Wilkes. Wiley. [aNC]

Wilson, B. A. & Baddeley, A. D. (1988) Semantic episodic and autobiographical memory in a postmeningitic amnesic patient. *Brain and Cognition* 8:31–46. [ADB]

Woodman, G. F. & Luck, S. J. (1999) Electrophysiological measurement of rapid shifts of attention during visual search. *Nature* 400:867–69. [GFW]

Woodman, G. F., Vogel, E. K. & Luck, S. J. (in press) Visual search remains efficient when visual working memory is full. *Psychological Science.* [GFW, rNC]

Wynn, K. (1992) Addition and subtraction by human infants. *Nature* 358:749–50. [BJS]

Xu, Y. (submitted) Integrating color and shape in visual short-term memory for objects with parts. [BJS]

Yamaguchi, S., Yamagata, S. & Kobayashi, S. (2000) Cerebral asymmetry of the "top-down" allocation of attention to global and local features. *Journal of Neuroscience* 20:1–5. [RDO]

Yantis, S. (1992) Multi-element visual tracking: Attention and perceptual organization. *Cognitive Psychology* 24:295–340. [aNC, SEA]

Yantis, S. & Johnson, D. (1990) Mechanisms of attentional priority. *Journal of Experimental Psychology: Human Perception and Performance* 16:812–25. [BJS]

Young, R. M. (1996) Functionality matters: Capacity constraints and Soar. In: *Mind matters: A tribute to Allen Newell,* ed. D. M. Steier & T. M. Mitchell. Erlbaum. [CPB]

Young, R. M. & Lewis, R. L. (1999) The Soar cognitive architecture and human working memory. In: *Models of working memory: Mechanisms of active maintenance and executive control,* ed. A. Miyake & P. Shah. Cambridge University Press. [CPB]

Zhang, G. & Simon, H. A. (1985) STM capacity for Chinese words and idioms: Chunking and acoustical loop hypotheses. *Memory and Cognition* 13:193–201. [aNC, SEA]