A Working Memory Workout: How to Expand the Focus of Serial Attention From One to Four Items in 10 Hours or Less

Paul Verhaeghen, John Cerella, and Chandramallika Basak
Syracuse University

Five individuals participated in an extensive practice study (10 1-hr sessions, 11,000 trials total) on a self-paced identity-judgment n-back task (n ranging from 1 to 5). Within Session 1, response time increased abruptly by about 300 ms in passing from n = 1 to n > 1, suggesting that the focus of attention can accommodate only a single item (H. Garavan, 1998; B. McElree, 2001). Within Session 10, response time was dramatically reduced and increased linearly with n for n ≤ 4, with a slope of about 30 ms. The data suggest that working memory consists of a focus of attention governed by a limited-capacity search, expandable through practice, and a content-addressable region outside the focus of attention.

Recent researchers have suggested that working memory operations that involve only a single item have privileged access (Garavan, 1998; McElree, 2001; McElree & Dosher, 1989; Oberauer, 2002; Verhaeghen & Basak, in press; Voigt & Hagendorf, 2002). These findings have led to the idea that a focus of attention, containing a single item, may lie at the core of the working memory system. Data from an earlier experiment of ours that exemplify this pattern are depicted in Figure 1 (Verhaeghen & Basak, in press, Experiment 1, college-age sample). The task is a version of the identity-judgment n-back task (after McElree, 2001), in which n was varied from 1 to 5 (for a full description of the task, see the Method section). The critical data are given by the response times (left panel). As shown in the figure, the Response Time × N trace is close to a step function, with a fast response at n = 1 and then a jump to slower and statistically equal response times for values of n > 1.

How are data such as these to be interpreted? In recent theories of working memory, researchers propose that working memory is subdivided into concentric regions that differ in the accessibility of the information stored. Cowan’s (1988, 1995, 1999, 2001) model has probably been most influential in this regard. In the model, Cowan proposes a hierarchical two-tier structure for working memory, distinguishing a zone of immediate access, labeled the focus of attention, from a larger, activated portion of long-term memory in which items are stored in a readily available but not immediately accessible state. (In accordance with the terminology introduced by McElree in, e.g., McElree, 2001, accessibility of an element in working memory is defined by the time needed to retrieve it; availability is defined by the probability that the element is retrieved correctly.)

The focus of attention is typically considered to be capacity limited, and to contain a fixed number of items; the activated portion of long-term memory is not thought to be capacity limited.
We are faced, then, with opposing expectations. On one hand, the handful of recent studies cited in the first paragraph point to a focus of attention that can accommodate no more than a single item. On the other hand, there is a large and venerable body of literature supporting the claim that the focus of attention can hold four items comfortably. We set up the present experiment to try to reconcile the two outcomes. Before we turn to the rationale and description of our study, we review the evidence for a narrow focus in more detail and take up some existing explanations for the discrepancy between narrow-focus and wide-focus results.

Evidence for a Narrow Focus of Attention in Working Memory

Prior to Verhaeghen and Basak’s (in press), at least four studies, using different paradigms, have converged on the conclusion that the width of the focus of attention is not four items but one item. The decisive variable is item accessibility, that is, the number of items that can be retrieved without necessitating a cost in response time. Step or threshold effects in the Response Time/Working Memory active load function, such as seen in Figure 1, are taken as an indication that the focus of attention is restricted to the number of items prior to the step—in the case of Figure 1, this is only a single item.

In one of the precedents to the Verhaeghen and Basak (in press) result, Garavan (1998) asked participants to keep a separate running count of triangles and squares that appeared in a random sequence on a computer screen. Presentation was self-paced, allowing for the recording of response times. Response times were about 500 ms slower when the stimulus shape was changed from the previous trial than when both successive stimuli had the same shape. Garavan interpreted the results as showing that participants kept separate mental counters for each stimulus shape, and that switching between counters in working memory was a time-consuming process. This, in turn, suggests that only one of the counters was immediately accessible for updating and hence resided in the focus of attention, whereas the other counter was held in temporary storage outside the focus of attention. Further research using this paradigm has shown that the cost associated with switching between counters is largely due to retrieval demands (Voigt & Hagendorf, 2002).

In another study, using a speed–accuracy methodology in a probe-recognition task with serially presented nine-word lists, McElree (1998) found that retrieval speed for the last item was faster than retrieval speed for any of the other items presented (see McElree & Dosher, 1989; Wickelgren, Corbett, & Dosher, 1980, for earlier demonstrations of this effect). McElree interpreted this result as implying that the last item processed resided in the focus of attention, whereas the items presented previously were stored in a less accessible state. In addition, in this study, items were categorized semantically, with three items for each category; all items from the category last presented were retrieved faster than the other items in the list. McElree (1998) interprets this result as indicative of some amount of clustering in the stored representations in working memory.

In a third study, again using a speed–accuracy procedure, McElree (2001) found that access times in an identity-judgment n-back task were faster for the n = 1 condition than they were in either the n = 2 or the n = 3 conditions. His interpretation was that in the n > 1 conditions, the items presented could not all be accommodated in the focus of attention, necessitating slow search processes on a significant portion of the trials. A model was fitted to the data, depicting performance as a mixture of a fast matching process...
working memory had privileged access only to the single item last items had intervened (Oberauer, 2002). This again suggests that faster when an item had just been updated than it did when other arithmetic operations. Oberauer found that updating occurred digit numbers stored in working memory one by one using simple updating task in which participants had to update a set of single-digit numbers stored in working memory one by one using simple arithmetic operations. Oberauer found that updating occurred faster when an item had just been updated than it did when other items had intervened (Oberauer, 2002). This again suggests that working memory had privileged access only to the single item last updated.

Theoretical Resolutions

There have been two attempts to reconcile the conflicting findings concerning the size of the focus of attention at a theoretical level. One proposal focused on the requirements of the task (the serial-attention explanation; Garavan, 1998), and the other focused on structural characteristics of working memory (the tripartite-architecture explanation; Oberauer, 2002).

Garavan’s (1998) explanation draws on the specifics of the tasks involved. In all of the narrow-focus procedures, attention is directed serially to different elements that are either stored in working memory or are being encoded into working memory. The serial requirement probably necessitates a controlled switch of attention to successive items. For example, in the Garavan running-count task, a switch in stimulus shape necessitates access to the stored count for each shape and a controlled update of that count. In the McElree (1998) probe-recall task, attention is directed to each successive item, which is then encoded explicitly into working memory. In the McElree (2001) n-back task, the item shown on the screen has to be explicitly compared with an item stored in working memory; the stored representation then has to be either strengthened or updated. In the Oberauer (2002) task, attention needs to be directed to the stored representation of the appropriate item, which then needs to be consciously updated. Thus, what distinguishes narrow-focus outcomes from wide-focus outcomes may be the requirement to shift attention serially within either the stored representations or the to-be-encoded items in the narrow-focus contexts. Oberauer uses the term object switching for this attentional process, and Garavan uses the term attention switching; we (see also Verhaeghen & Basak, in press, and Voigt & Hagen- dorf, 2002) prefer to use the term focus switching—the switching of the focus of attention from one item to another. It is this focus switching that precipitates the characteristic jump in response time when an active working memory load of two (or more) rather than of one must be maintained.

Oberauer (2002) has advanced a different kind of explanation. He argues that the conflicting results are not due to the processes involved but that, rather, the two outcomes derive from different memory structures. Oberauer argues for the existence of a concentric tripartite store with functionally distinct regions. In effect, the tripartite store combines the architecture proposed by Cowan (e.g., 1995) with the architecture proposed by McElree (2001) and Garavan (1998). The three regions are characterized by their increasing accessibility to cognitive processes. The first and outermost tier, the activated part of long-term memory, serves mainly to store information over brief periods; the processes operating in this region concern encoding and maintenance. The second tier, lying within the activated region, is a capacity-limited region of direct access corresponding to Cowan’s focus of attention, where a limited number of items are stored that are likely to be selected for subsequent processing. The third and innermost tier, which lies within the region of direct access, is the focus of attention, corresponding to the focus of attention proposed by McElree and Garavan and containing only a single chunk of information, namely, the chunk actually selected as the object of the next cognitive operation.

Oberauer’s (2002) theory was designed to accommodate data from a variant of the memory-updating task. In his version of the task, two sets of single-digit numbers were presented, and set size was either one or three. One set of numbers was marked for subsequent updating (using mental arithmetic operations); the other set was marked for storage only. First and foremost, Oberauer observed the focus-switching effect indicative of a narrow focus of attention: When the current item matched the last item updated, response times were faster than when a different item needed to be accessed. Second, Oberauer found that speed of updating was dependent on the size of the to-be-updated set and not on the size of the to-be-stored set. His interpretation was that the two sets were stored in distinct regions of working memory. The storage-only set was placed in the activated part of long-term memory. The to-be-updated set needed to be held in the more accessible region of direct access, ready to be retrieved into the focus of attention. Because items compete for activation in the region of direct access, retrieval times in that store were dependent on the number of items stored.

The Size of the Focus of Attention: A Structural Limitation or an Allocation Artifact?

The limitations imposed on the focus of attention by both the serial-attention explanation and the tripartite-architecture explanation are hard and fast. Both theories treat a focus of one as built into the working memory machinery. Oberauer’s (2002) position is that the focus of attention “holds at any time the one chunk that is actually selected as the object of the next cognitive operation” (p. 412). McElree (2001) states that “focal attention is able to maintain one temporally extended event only” (p. 820). Garavan (1998) claims that “this subset [i.e., the attentional focus as a subset of short-term memory] contains just one item” (p. 275). The single exception is that this event may contain multiple items if they can be coded into a chunk that forms a unitary processing epoch (Garavan, 1998), for instance a set of categorically related words (McElree, 1998). The limitations imposed on the region of direct access, on the other hand, have been taken to be more fluid and to depend on constraints such as overwriting or crosstalk (Oberauer, 2002). It seems to us that the position that the focus of attention is limited to a single item is excessively rigid even under circumstances that require serial switching. One of the basic suppositions of attentional theory has been that attentional capacity can be allocated flexibly across the perceptual–cognitive field, and that this allocation follows the needs of the participant, the demands of
the task, or both (e.g., Kahneman, 1973; Wickens, 1984). The only constraint is that the amount to be allocated falls within the resource limits of the individual at the time at which the task is performed (limits that vary between individuals and, also, within individuals from instant to instant and over development). The distribution of attention has been shown to be quite flexible. For instance, with instructions on strategy use and practice with the task, even hard-wired limitations in the attentional system such as the bottleneck associated with time delays in double stimulus–response mapping paradigms (the psychological refractory period) can be made to disappear (Meyer & Kieras, 1999; Schumacher et al., 2001).

These considerations point to a third resolution of the narrow-focus versus wide-focus controversy: Perhaps the reported values represent the two ends of an underlying continuum of resource allocation. Let us entertain the hypothesis that the focus of attention can vary in size: It shrinks to one item when all resources are channeled to the processing of a single item, and it expands up to a size of four when several items can be processed in parallel. On this view, a focus of one is not due to a structural limitation of the cognitive system but, rather, to the way attention is distributed over the task and stimuli. At least one mathematical model has been proposed that allows the capacity of the focus of attention to fluctuate flexibly between one and four items (Usher, Cohen, Haarman, & Horn, 2001). The underlying mechanism is situated in the level of lateral inhibition: Decreasing levels of inhibition widen the focus of attention to its maximum value of four.

A core assumption of our proposal is that of a trade-off between resources applied to processing and those applied to storage (presumably, the latter are dedicated to keeping item activation levels high). This assumption underlies the literature on working memory span tasks (operation span, listening span, reading span, and the like). For instance, Engle, Kane, and Tuholski (1999) argue that individual differences in operation span are due to more efficient processing of the concurrent task, and that “this processing efficiency frees up resources” (p. 110) that can then be applied to item maintenance. Likewise, Daneman and Carpenter (1980) assume that resources applied to the processing of sentences in the reading span task reduce capacity for storage, and vice versa. This notion is also intrinsic to Cowan’s portrayal of working memory as an embedded-process system (Cowan, 1999).

**Practice Manipulations**

Garavan (1998, Experiment 2) used an extended practice manipulation as a test for a resource-allocation explanation of the focus-of-one outcome in his symbol-counting task. In his study, a substantial focus-switch cost remained after practice; Garavan interpreted this as evidence for the fixed, structural nature of the size limitation in serial attention.

We designed the experiment reported below as another test of the resource-allocation hypothesis. We took up Garavan’s (1998) practice manipulation and applied it to a simpler task. In our experiment, participants were given extended practice on an n-back task with the expectation that practice would affect resource allocation. We reasoned that extended practice would lead to automaticity in some of the task components; this, in turn, would free up resources that could then be applied toward storage inside the focus of attention. An empirical test of this reasoning seemed straightforward: If the limit on the size of the focus of attention is structural, then we would expect that the n-back step function (with the step at \( n = 2 \)) will remain intact over the course of practice (although the height of the step may be diminished). If, however, the location of the step shifts over the course of practice (perhaps settling, finally, at \( n = 5 \)), this would suggest that a focus of one is due not to a structural limitation, but to a resource limitation. In keeping with the mainstay of the literature, we assumed that the maximum size of the focus of attention is likely to be four, for reasons admittedly unknown.

For the present study, we made a few changes in the McElree (2001) paradigm, following Verhaeghen and Basak (in press). Response times were participant-controlled, rather than experimenter-controlled (i.e., time–accuracy methodology). Fewer trials were thereby needed in any condition. This is critical in a practice study in order for one to capture successive phases in the evolution of a participant’s performance. This also allowed us to vary \( n \) over a wider range, from 1 to 5. If the focus of attention indeed expands with continued practice on the task, a wide range of \( n \) might be needed to document the change.

We introduced a second change to reduce a demanding control requirement. In the standard version of the task, stimuli are presented one at a time in a fixed location on the computer screen. Keeping track of the position of the stimulus in this stimulus series is paramount for comparing every nth item, and this demand may be confounded with the working memory load of the task. We tried to minimize this keeping-track demand by presenting our stimuli one at a time in virtual columns on the screen. Each column was defined by both location and stimulus color; the number of columns was equal to \( n \) (see also Hartley, 2002). The instruction to the participant was to compare the current item with the item last presented in the same column and color. This change makes our task, in effect, an identity-matching \( n \)-column 1-back task.

The collection of participant-controlled response times allowed us to examine the shape of response time distributions, as well as any changes in shape as training progressed. We used the ex-Gaussian model to analyze the distributions. In brief, in the ex-Gaussian model, it is assumed that each response time can be represented as the sum of a Gaussian or normally distributed random variable and an independent exponentially distributed random variable. The ex-Gaussian distribution is described by three parameters: \( \mu \) and \( \sigma \) are the mean and standard deviation of the normal distribution, and \( \tau \) is the mean of the exponential distribution. \( \mu \) and \( \sigma \) determine the location of the leading edge of the distribution; \( \tau \) reflects slow responses at the tail of the distribution (the skew).

The ex-Gaussian decomposition provides more information from the set of response times than the usual summary measures of mean and variance. (Note that the observed mean of the distribution equals the sum of \( \mu \) and \( \tau \); its variance equals the sum of \( \sigma^2 \) squared and \( \tau^2 \) squared; its skew equals two times \( \tau \) cubed divided by the standard deviation cubed.) The three parameters can dissociate, that is, they can be affected differentially by experimental manipulations (such as a Stroop manipulation; Spieler, Balota, & Faust, 1996) or by different tasks performed on identical materials (Hockley, 1984). In our previous study with the same task (Verhaeghen & Basak, in press, Figure 1 of this article), we found that \( \tau \) increased from \( n = 1 \) to \( n = 2 \) and remained flat over the subsequent values of \( n \), whereas the values of \( \mu \) and \( \sigma \) were constant over the whole range of \( n \). The step function visible in Figure 1, therefore, was induced entirely by changes in
tau, or the skew of the distribution. Differential changes in tau were available from the present study as a further diagnostic device to identify practice effects on the size of the focus of attention.

Accessibility Versus Availability

Published work in the focus-of-one tradition has centered on Response Time × Load functions, which define the accessibility of items. Accuracy × Load functions have received less emphasis. In his Figure 1, McElree (2001) predicted an accuracy function that decreases monotonically and smoothly over n rather than on a step. This is because item availability (i.e., the probability that an item is available for processing) is portrayed as susceptible to decay, to interference, or to both. Because both storage time (which governs decay) and the number of items in working memory (which creates interference) are governed by the overall n, a monotonic decrease in accuracy is predicted. The decrease is expected to be smooth, because there is no reason to assume that items stored outside the focus of attention will lose activation (whether due to decay or interference) more rapidly than an item inside the focus of attention. In addition, McElree predicts that the decrease will be negatively accelerated, that is, that accuracy will decline toward a horizontal asymptote.

To the best of our knowledge, the Verhaeghen and Basak (in press, Experiment 1) study was the first to examine the availability function over a large range of n. Their accuracy data are reproduced in Figure 1, right-hand panel. These results conform to McElree’s (2001) first-order prediction: A smooth decline over n is evident. Contrary to McElree’s more detailed predictions, the decline was not negatively accelerated; rather, it was positively accelerated. This trend would seem to rule out two possible explanations for the decline in accuracy. It shows that the availability function cannot be the outcome of error compounding over successive values of n. Error compounding is exponential; that is, if P is the accuracy obtained for n = 1, the compounded value for n items is \(P^n\). This process generates a negatively accelerated function over n (Schweickert, 1985). The finding of positive acceleration is also incompatible with traditional mechanisms of forgetting, which likewise yield negatively accelerating functions with a horizontal asymptote (Rubin & Wenzel, 1996).

Verhaeghen and Basak (in press, Experiment 1) also collected data from a sample of older adults that differed from the data of the college-age sample in an interesting way. In older adults, the data from a sample of older adults that differed from the data of the young sample in an interesting way. In addition, learning effects were seen, that is, response times will drop, possibly including the height of the step function, but the location of the step will stay fixed at n = 2.

The second and third scenarios (Patterns B and C) show two patterns that might emerge under the focus-of-four hypothesis. In addition to the learning effects, these scenarios incorporate an expanded focus of attention, with a step located at n = 5. The two patterns differ with regard to the dynamics of the expanded focus. Pattern B depicts the focus of attention as content addressable, like the region of direct access. In that case, a true step function will be seen: a horizontal segment up to the breakpoint (which indicates the size limit of the focus of attention) and a horizontal segment thereafter.

Pattern B is elegant in its process symmetry, but it is plagued by a question: If the focus of attention were content addressable, it is hard to see why its energy requirement would be any greater than the region of direct access; why, therefore, should it not be fully expanded for all tasks at all stages of practice? In Pattern C, we answered this question by assigning to the focal store a more effortful retrieval process: Items stored in the focus of attention are accessed through a limited-capacity parallel search. This process is known to generate a linear rise in response time over n, with a shallow, positive slope (Townsend & Ashby, 1983). Pattern C shows this linear ramp over the range n = 1–4, followed by a discontinuity at n = 5. The breakpoint would be signaled by an increase in latency larger than that extrapolated from the linear ramp.

Pattern C draws on Cowan’s (2001) premise that the zone of immediate access is delimited by the number of items that can be processed in parallel. In support of his bound (four items plus or minus one), Cowan (1999) cites Trick and Pylyshyn (1994) on the subbating span and Fisher (1984) on parallel processing in consistently mapped visual search; both are processes characterized by shallow, positive load functions. As a supplement to Cowan, we note another memory process with this linear signature, short-term memory scanning (Sternberg, 1966). Mathematical models of this process are highly developed and converge on the schema of parallel, self-terminating, limited-capacity memory access (Murdock, 1971; Ratcliff, 1978; Van Zandt & Townsend, 1993). Applied to working memory tasks, this process has appealing properties. If access occurs in parallel, but capacity-demanding, channels, the number of supportable channels can be supposed to be a function of the total resources available to the working memory system. If a task is very demanding, the residual may allow only a single active channel. As a task is automated, resources are freed, allowing the system to open more channels to simultaneous search.

Model C traces storage limits in the focus of attention to retrieval characteristics: The number of items is equal to the number of channels that can be searched in parallel. It lacks the
symmetry of Model B: The focus of attention and the region of
direct access differ in their retrieval process: The region of direct
access uses a slow content-addressable system; the focus of atten-
tion uses a rapid parallel system. If the dynamic signatures of
subitizing and memory search (Sternberg, 1966; Trick & Pyly-
shyn, 1994) extend to the \( n \)-back task, the expanded focus of
attention should show a shallow Response Time \( \times N \) slope of
about 40–100 ms per item, as illustrated in Pattern C. Hockley
(1984, Experiment 1) has shown that the ex-Gaussian signature of
the Sternberg search process is a linear increase in \( \tau \) over
memory load while \( \mu \) and \( \sigma \) remain constant.

With regard to item availability (i.e., item accuracy), we ex-
pected a replication of the Verhaeghen and Basak (in press) results
for the first session, that is, a monotonically declining, positively
accelerated curve over increasing values of \( n \). After practice, for
Pattern A, we again expected performance to be nearly perfect for
\( n = 1 \), with a monotonic decline thereafter (which may be atten-
uated by practice). As noted above, our current assumption was
that items residing in the focus of attention are immune to inter-
ference effects. If this held true, we would expect the following
results for Patterns B and C: near-perfect performance for values
of \( n \) between 1 and 4, and possibly a decline for \( n = 5 \).

Method

Participants

Five adults, all of whom were right-handed women, participated in the
experiments. Four participants were graduate students from disciplines
outside psychology; 1 was a research assistant employed in the psychology
department. They were paid $10 per hour for their participation.

Task and Procedure

As explained above, we modeled the task after an \( n \)-back task used by
McElree (2001). In McElree’s task, participants indicated whether the item
currently presented on the screen was identical to the item presented \( n \)
positions back. In our version, stimuli were shown on the screen in \( n \) virtual
columns; columns were additionally defined by stimulus color. Figure 3
shows a black-and-white rendition of a sample stimulus set for one trial (in
this case \( n = 4 \)) as it would appear on the computer screen if all items
remained visible. The digits shown were 6-mm tall, the horizontal sepa-
ration between columns was 1.3 cm, and the vertical separation between
rows was 1.2 cm. In practice, only one digit was shown at any one time;
the order of presentation was a reading pattern: left to right and top to
bottom. For the first row, a new digit was presented every 2,000 ms; from

![Figure 2. Expected pattern of response times before practice and three hypothesized patterns of response time after practice. Pattern A should occur if the focus of attention is immutably fixed at a size of one; Pattern B is expected if the focus expands and is content addressable; Pattern C would be obtained if the focus is expandable and subject to limited-capacity parallel search.](image)

<table>
<thead>
<tr>
<th>3</th>
<th>5</th>
<th>8</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

![Figure 3. An example of a trial in the 4-back version of the task if all stimuli remained onscreen. In the experiment, stimuli were shown one at a time in a reading pattern (left to right, then on the next line, etc.); each column was depicted in a different color. The first row was presented at a 2 s per item pace; presentation of subsequent stimuli was participant-paced. The response required was a judgment whether the digit currently projected was identical to the digit previously shown one row higher in the same column.](image)
the second row on, participants pressed either of two keys to indicate their answer. The / key stood for “yes” (i.e., identical) and was marked with a piece of green tape; the Z key stood for “no” (i.e., different) and was marked with a piece of red tape. Participants were instructed to be both fast and accurate. As soon as the key was pressed, the next stimulus appeared. Participants were encouraged to choose a comfortable viewing distance from the screen.

Each stimulus set (a trial) contained a total of 20 to-be-responded-to items. After each trial, the participant received feedback about both total accuracy and average response time over the run of 20 items. Within each session, 11 trials (yielding a total of 275 response times) were presented for each value of n (n varied from 1 to 5), distributed as follows: first 6 trials for n = 1, then 6 for n = 2, 6 for n = 3, 6 for n = 4, 11 for n = 5, 5 for n = 4, 5 for n = 3, 5 for n = 2, and 5 for n = 1. The first trial for each of the values of n in the first half of each session was considered a warm-up trial and was discarded from further analysis. For each trial, half of the stimuli were identical to the item n back, and the other half were not. The exact composition of each trial was determined by an online algorithm that used a random seed. All participants completed ten 1-hr sessions over the course of 5 consecutive days, at a rate of two sessions a day, one scheduled in the morning and the other in the afternoon.

Response Time Analysis

For all response time analyses (including ex-Gaussian decomposition), we used correct responses only. We discarded response times for the second row, that is, when the comparison items were the items initially presented by the experimenter. We removed response times less than 100 ms and greater than 5,000 ms from the data set. In total, we discarded 53 data points out of a total of 45,079 through this trimming procedure. We used the quantile maximum probability estimation procedure (QMPE) proposed by Heathcote and Brown (in press) to fit the ex-Gaussian model to the response time distributions; we used the maximum number of quantiles option for fitting. Monte Carlo simulation has indicated that the QMPE algorithm is more efficient and less biased than other methods in use. We used an alpha level of .05 for all statistical tests.

Results

The response time and accuracy data from Sessions 1 and 10 are depicted in Figure 4 for each individual; the average data from every session are depicted in Figure 5. These plots show performance as a function of memory load, with “yes” and “no” responses combined. Table 1 breaks down the response time data by session and by n separately for “yes” and “no” responses.

The First Session

Mean accuracy and response time data for Session 1 are presented in the first two columns of Figure 4 for each participant separately and for the average of the 5 participants.

Response times. In a two-way analysis of variance (ANOVA), response times were found to be affected both by n (n = 1–5), F(1, 4) = 11.53, MSE = 9,730.73; and trial type, F(1, 4) = 50.56, MSE = 15,955.12 (“yes” response time = 859 ms, “no” response time = 1,113 ms). At least part of the increase in response times in the “no” trials can be attributed to the memory-updating requirement. That is, when the item n back is different from the item currently in focus (i.e., a “no” trial), the item stored in the nth memory slot back needs to be replaced with the new item. No updating is required when the two items are identical (i.e., a “yes” trial). Trial type did not interact reliably with n, F(4, 16) = 1.96, MSE = 7,417.37, indicating that the memory-updating process and the focus-switching process were independent of each other. Because of the mutual independence of the two processes, we decided to pool data across “yes” and “no” responses for all further analyses.

The findings from the response time analyses show the telltale signature of the predicted architectural boundary (see Figure 2, Pattern A): There is a sharp increase in response time from n = 1 to n = 2 (from 800 ms to 1,039 ms), followed by a flat curve from n = 2 to n = 5. The increase from n = 1 to n = 2 was significant for all 5 participants, as indicated by nonoverlapping 95% confidence intervals. We investigated the flatness of the individual curves for n > 1 using regression analysis. Participant A showed a significant slope of 71 ms/n, SE = 15, F(1, 704) = 23.53; for the other 4 participants, slopes were not significantly different from zero. When all data were pooled, the slope for the resulting regression line was 6 ms/n, and this slope was not significantly different from zero, SE = 8.07, F(1, 3372) = 0.50.

Ex-Gaussian decomposition. The third column of Figure 4 shows the results of the ex-Gaussian decomposition of response times for each of the 5 participants, as well as the parameters averaged over individuals. (We defer to a chi-squared assessment of the goodness of the ex-Gaussian fits in the Practice Effects Over Sessions section). Turning our attention first to the leading edge of the distribution (i.e., mu, the mean of the Gaussian component, and sigma, its standard deviation), no evidence for a step function is apparent. There were no significant differences between the values for mu and sigma at n = 1 versus n = 2, F(1, 4) = 1.41, MSE = 1,196.28; and F(1, 4) = 1.28, MSE = 350.23. Regression analysis showed nonsignificant slopes over the full range of n values for the mu parameter for all 5 participants (slopes were 17, −8, 5, 10, and 15 ms/n, respectively). After pooling the parameter values across participants, we found the slope for mu to be a nonsignificant 8 ms/n. For sigma, we found a significant 20 ms/n slope for Participant B, but slopes for all other participants were not significantly different from zero (3, 3, 1, and −1 ms/n, respectively). After pooling the parameter values across participants, we found the slope for sigma to be a nonsignificant 5 ms/n.

The tail of the distribution, however, shows a clear step function. For the tau parameter, a sharp increase from n = 1 to n = 2 can be noted (from 278 ms to 493 ms), followed by a flat curve from n = 2 to n = 5. The difference between the value of tau at n = 1 and at n = 2 was significant, F(1, 4) = 9.18, MSE = 12,544.40. We investigated the flatness of the curves for n > 1 using regression analysis. None of the 5 participants showed a slope that was significantly different from zero (slope values were 47, −23, 0, −41, and 7 ms/n, respectively). After pooling the parameter values across participants, we found that the slope for tau for n > 1 was −2 ms/n. This value was not significantly different from zero.
Figure 5. Average accuracy (A), response time (B) and ex-Gaussian decomposition of response times (C) for all 10 sessions. Error bars denote standard errors calculated at the group level.
The linear component of this model was significant, 
\[ F(1, 4) = 11.62, \quad \text{MSE} = 0.002; \] and \[ F(4, 16) = 11.01, \quad \text{MSE} = 0.001. \] Both the linear and the quadratic component of the effect of \( n \) were significant, \( F(1, 4) = 11.62, \quad \text{MSE} = 0.002; \) and \( F(1, 4) = 18.48, \quad \text{MSE} = 0.00009. \) (The cubic and fourth-order components were not significant.)

### The Final Session

Mean accuracy and response time data for Session 10 are presented in Figure 4 for each participant separately and for the average of the 5 participants.

#### Response times.
As is evident in column 5 of Figure 4, by the end of the study, there was no longer an indication of privileged access in the \( n = 1 \) condition. Rather, the data appeared to conform to Pattern C of Figure 2. Response times increased linearly over \( n = 1–4 \) with an additional increment, or breakpoint, at \( n = 5 \). In a repeated measures ANOVA testing for the main effect of \( n \) with the average response times of each individual participant as the data points, we obtained a significant effect, \( F(4, 16) = 33.83, \quad \text{MSE} = 670.42, \) with slower response times for larger values of \( n \). The linear component of this model was significant, \( F(1, 4) = 47.43, \quad \text{MSE} = 1,829.14; \) the quadratic component was close to significance, \( F(1, 4) = 7.64, \quad \text{MSE} = 333.98, \quad p = .051. \) Regression analysis yielded significant slopes for all 5 participants, with values of 55, 49, 39, 55, and 21 ms/\( n \), respectively, smallest \( F(1, 902) = 51.82, \quad \text{MSE} = 82,502.42. \) When all data were pooled, the slope for the resulting regression line was 41 ms/\( n \). This slope value was significantly larger than zero (\( SE = 2.24), F(1, 4516) = 326.80. \)

To test for the hypothesis of slower access in the \( n = 5 \) condition relative to the regression line describing response times for \( n = 1 \) to \( n = 4 \), we fitted a linear model to the \( n < 5 \) portion of the curve using all available data points and tested whether the response times at \( n = 5 \) were adequately predicted by the regression line. The resulting regression line had a slope of 33 ms/\( n \) (\( SE = 2.68), \) which is significantly larger than zero, \( F(1, 3761) = 157.45. \) The presence of a linear effect in the \( n = 1 \) to \( n = 4 \) portion of the curve was further confirmed in a repeated measures ANOVA testing for the effects of \( n, F(3, 12) = 41.60, \quad \text{MSE} = 228.15, \) with slower response times for larger values of \( n \). The linear component of this model was significant, \( F(1, 4) = 68.55, \quad \text{MSE} = 407.04, \) whereas the quadratic component clearly was not, \( F(1, 4) = 0.06, \quad \text{MSE} = 82.05. \) Response time at \( n = 5 \) was underpredicted by 38 ms (a value about equal to the slope of the regression line), and this deviation was significant, \( t(754) = 3.32. \)
To further confirm the presence of a breakpoint at \( n = 4 \) rather than earlier in the sequence, we repeated this procedure using the \( n < 4 \) portion of the curve to fit the regression. This regression line was virtually identical to the line derived for the \( n < 5 \) portion (slope was now 31 ms/n, \( SE = 3.69 \)), and no significant deviation was found at the \( n = 4 \) point (deviation = 9 ms), \( t(848) = 1.09 \). This statistically confirms the presence of a breakpoint at \( n = 4 \). We should note that this quadratic trend is in the direction opposite that predicted from the McElree (2001) model, that is, the data curve traveled upward rather than showing the threshold-and-cruise pattern seen in the first session.

In an \( N \times \) Trial Type ANOVA, response times were found to be slower for “no” trials (541 ms) than they were for “yes” trials (483 ms), \( F(1, 4) = 40.94, \text{MSE} = 1,002.60 \). As for the first session, trial type did not interact reliably with \( n \), \( F(4, 16) = 1.33, \text{MSE} = 407.69 \), indicating that the shape of response time (RT) over \( n \) function was identical for “yes” and “no” responses.

There was also a substantial speedup at all \( n \) after 9 sessions of practice. The average response time (averaged over \( n \) and over participants) was 583.70 ms in Session 1 (\( SE = 33.56 \)) versus 511.93 ms in Session 10 (\( SE = 31.03 \)); this difference was significant, \( F(1, 4) = 545.08, \text{MSE} = 5,103.92 \). Indeed, there was no overlap between the five response times of Session 1 and the five response times of Session 10, smallest \( t(4) = 4.73 \).

Ex-Gaussian decomposition. Ex-Gaussian decomposition uncovered a pattern characteristic of Sternberg-like search (Hockley, 1984, Experiment 1): a linear increase in tau and constancy in \( \mu \) and \( \sigma \). The latter is shown in Figure 4 as flat slopes for \( \mu \) and \( \sigma \). For \( \mu \), the slopes were 8, -4, -7, -5, and 16 ms/n for the 5 participants, respectively, and 4 ms/n after the data were pooled (\( SE = 4.21 \)). None of these slopes were significantly different from zero. Likewise, none of the slopes for \( \sigma \) (-4, -1, 3, -4, and 1 ms/n for the individual participants and -2 ms/n for the pooled data, \( SE = 1.31 \)) were significantly different from zero.

The \( \tau \) parameter yielded significant slopes for Participants A–D, namely 43, 55, 24, and 57 ms/n, respectively. The slope for Participant E was small, 7 ms/n, and not significant. The slope for the pooled data was 38 ms/n, and this slope was significantly larger than zero. The \( \tau \) slope is close to the 43 ms per item slope reported by Hockley (1984) for \( \tau \) in his Sternberg memory-scanning task. This lends further credence to the hypothesis that the retrieval process within the widened focus of attention is a parallel self-terminating process with limited capacity.

Because the \( \tau \) parameter seems to show the same discontinuity at \( n = 5 \) as the raw response time data, we used the same procedure to test for the presence of a breakpoint at \( n = 5 \). The regression line over the \( n < 5 \) section had a smaller slope than the regression line obtained over the full range of \( n \) values, 28 ms/n (\( SE = 10.52 \)); the slope is significantly larger than zero). At the \( n = 5 \) point, the deviation from this line was quite large (45 ms, or more than 1.5 times the slope; the largest absolute deviation at the other points was -5 ms). The deviation was not significantly different from zero, \( t(4) = 0.96 \), but note the small number of degrees of freedom.

Accuracy. In the final session, the decrease in accuracy over \( n \) is only noticeable for \( n = 4 \) and \( n = 5 \). A repeated measures ANOVA with the average accuracy of each individual participant as data points yielded a main effect of \( n \), \( F(4, 16) = 4.62, \text{MSE} = 0.001 \). Both the linear and the quadratic component of the effect of \( n \) were significant (.05 < \( p < .10 \)), \( F(1, 4) = 4.97, \text{MSE} = 0.002 \); and \( F(1, 4) = 4.78, \text{MSE} = 0.001 \); the cubic and fourth-order components were not significant (\( p > .15 \)). For all participants, proportion correct at \( n = 4 \) and \( n = 5 \) was significantly lower than 1, as indicated by significant \( t \) tests. Two participants (A and B) reached perfect accuracy for \( n \leq 3 \).

Practice Effects Over Sessions

Figure 5 provides the practice data: mean response times, ex-Gaussian parameters, and accuracy over all 10 sessions as a function of \( n \), averaged over the 5 participants.

Response time. In an ANOVA with \( n \) and session as the within-subject variables, both main effects were significant. Response time speeded up over sessions, \( F(9, 36) = 91.28, \text{MSE} = 6,455.21 \); this speed-up had a significant linear component, \( F(1, 4) = 409.74, \text{MSE} = 10,187.32 \); as well as a significant quadratic component, \( F(1, 4) = 58.58, \text{MSE} = 15,573.59 \). Thus, response times dropped with practice, and, as is usual in practice or learning studies, the largest effects occurred early in practice. The cubic component was also significant, \( F(1, 4) = 12.54, \text{MSE} = 9,109.71 \), indicating some bumpiness in the data. To be specific, response times at Session 4 seemed to be elevated above the quadratic trend. We do not think that this effect is substantive.

Response time increased with \( n \), \( F(4, 16) = 53.06, \text{MSE} = 5,511.48 \). The \( N \times \) Session interaction was also significant, \( F(36, 144) = 3.97, \text{MSE} = 1,031.07 \). Thus the accessibility function changed its shape, as well as dropped, from session to session.

Ex-Gaussian decomposition. Figure 5 shows the ex-Gaussian parameters averaged over the 5 participants as a function of sessions. These values were based on the analysis of 250 response time distributions (5 participants \( \times \) 5 sessions \( \times \) 10 sessions), each containing approximately 200 observations. We calculated a chi-squared goodness-of-fit statistic for each of the ex-Gaussian solutions using the method recommended by Van Zandt (2000). We divided the observed distribution into 20 quantiles, and we calculated the expected frequency for each quantile from the estimated ex-Gaussian density function (nominally 10, given 200 observations total). We obtained the chi-squared statistic by summing over the observed and expected values. Of the 250 chi-squared values, 14 were larger than the criterion that is often taken to indicate a satisfactory fit for models of this sort, 3 times the chi-square degrees of freedom (Bollen, 1989; Kline, 1998; for our model, the degrees of freedom were 16: 20 quantiles minus 3 estimated parameters minus 1). This is a failure rate of 5.6% over the 250 models and is almost exactly the rate expected by chance. The 14 failures were distributed fairly evenly across participants, across ns, and across sessions. We concluded that the ex-Gaussian decomposition provided an adequate description of the response time distributions.

As for the values of the fitted parameters, we found that \( \mu \) was sensitive only to session (i.e., performance speeds up over the course of practice), \( F(9, 36) = 21.11, \text{MSE} = 4,993.22 \). The speed-up over sessions has a significant quadratic component, \( F(1, 4) = 18.29, \text{MSE} = 7,742.03 \), as well as a significant linear component, \( F(1, 4) = 41.47, \text{MSE} = 19,022.21 \); that is, the largest learning effects occurred early in practice. Neither the effect of \( n \), \( F(4, 16) = 1.78, \text{MSE} = 3,952.82 \); nor the \( N \times \) Session interaction were significant, \( F(36, 144) = 1.26, \text{MSE} = 370.348 \), indicating that \( \mu \) stayed flat over \( n \) over the course of practice.

For \( \sigma \), we found a significant decrease over sessions, \( F(9, 36) = 10.09, \text{MSE} = 560.07 \). Only the linear component was
significant, $F(1, 4) = 28.80$, $MSE = 1,376.56$; the quadratic component had a $p$ value of .08; $F(1, 4) = 5.54$, $MSE = 1,464.76$. The effect of $n$ on sigma was not significant, $F(4, 16) = 5.24$, $MSE = 410.98$, indicating flatness of sigma over $n$. The session main effect was qualified by a significant but small $MSE$ interaction effect, $F(36, 144) = 1.75$, $MSE = 146.69$. This interaction effect seems to be due to the value of sigma at $n = 1$, which changes less over sessions than the value of sigma for $n > 1$.

For $tau$, we found a speed-up over sessions, $F(9, 36) = 38.98$, $MSE = 5,227.86$; and a general increase with $n$, $F(4, 16) = 17.58$, $MSE = 13,063.66$. The $N \times Session$ interaction was also significant, $F(36, 144) = 3.37$, $MSE = 1,298.92$, and its shape echoed the shape of the changes in response times. The figure indicates that practice effects are smaller for $n = 1$ than they are for $n > 1$. All components for the effect of $n$ on $tau$, up until the fifth order component, reached significance, smallest $F(1, 4) = 10.66$, $MSE = 1,862.25$. Despite the apparently complex pattern of the data, it is clear from the figure that the largest effects were found early in practice.

**Accuracy.** For accuracy, there was no main effect of session, $F(9, 36) = 1.99$, $MSE = 0.001$; thus overall, accuracy did not change with practice. We obtained a main effect of $n$, $F(4, 16) = 7.24$, $MSE = 0.004$; as well as a significant $N \times Session$ interaction, $F(36, 144) = 2.28$, $MSE = 0.000$. Thus the availability function changed its shape from session to session.

**Discussion**

Our practice study was directed toward a basic theoretical question, the structure of working memory. To be specific, we asked whether the size of the focus of attention (the immediately accessible region of working memory) is limited to a single item, as reported in a series of recent studies (Garavan, 1998, McElree, 1998, 2001; Oberauer, 2002). The alternative view, the consensus of a wider literature (most notably Cowan, 1995, 1999, 2001), is that the focus of attention can hold four items.

We hypothesized that the two values, one versus four, may represent the endpoints of an underlying continuum of resource allocation. On this hypothesis, the focus of attention can vary in size: It shrinks to one item when all resources must be committed to an item, and it expands up to a size of four when several items can be processed in parallel. Here a focus of one is due not to any structural limitation of the cognitive system but rather to the attentional demands of the task and the stimuli.

We tested the resource-allocation hypothesis using a modified version of the $n$-back task, which has been shown previously (McElree, 2001; Verhaeghen & Basak, in press, Experiment 1) to yield a narrow, single-item focus of attention. Five participants were given extensive practice on the task in ten 1-hr sessions. We conjectured that practice would lead to the automatization of some of the processing steps. Freed attentional resources might then be reallocated to an expanded focus of attention. Possible outcomes are shown in Figure 2, in the form of three alternative Response Time $\times N$ or accessibility functions. Alternative A shows no change in the breakpoint of the function, which is still located at $n = 2$, despite the expected speedup in response times overall. Alternatives B and C show the breakpoint shifted to $n = 5$, indicating an expansion of the focus of attention from one item to four items. In Alternative B, the leading segment (over the range of 1–4) is flat, indicating content addressability of items stored within the focus. In Alternative C, the leading segment is ramped, suggesting a parallel search process with limited capacity.

In brief, the results of the first session replicated the focus-of-one finding, and the results of the final session conformed to Pattern C. This shows that (a) the focus of attention can be expanded with practice, even in tasks initially requiring serial attention; and (b) the retrieval dynamics within the focus of attention are different from those in the region of working memory surrounding it.

**Performance at the Onset of Practice: The McElree Strut**

The findings from the response time analyses agree with the predictions from the McElree (2001) focus-of-one model and replicate an earlier experiment of ours in which we used the same methodology on a larger sample of participants (Verhaeghen & Basak, in press, Experiment 1, see Figure 1). The accessibility functions (i.e., the response time data in Figure 4) show the telltale signature of an architectural boundary: There is an abrupt increase in response time from $n = 1$ to $n = 2$ (from 800 ms to 1,039 ms) followed by a flat segment from $n = 2$ to $n = 5$. This step function we call the *McElree strut*. This configuration was statistically significant for all 5 participants. It suggests a focus of attention of one item. When more items are stored, those residing in working memory outside the focus of attention must be retrieved, at a cost of about 0.25 s in response time.

The flatness of the segment from $n = 2$ to $n = 5$ suggests two things. It shows that the step effect is not a consequence of increased memory load per se. If the effect were dependent on load alone, response times would continue to rise. It also demonstrates that the elements contained in the portion of working memory outside the focus of attention are not subject to a limited-capacity parallel or to a serial search process, both of which would give rise to a ramped segment rather than a flat segment (Ashby, Tein, & Balakrishnan, 1993; Hockley, 1984). Rather, the items outside the focus of attention appear to be directly content addressable, much in the way elements stored in long-term memory are. Our results, together with those of Verhaeghen and Basak (in press), show that this content-addressable area outside the focus of attention is rather large; it can contain at least four items.

The accuracy data in Figure 4 also replicate those of Verhaeghen and Basak (in press, Experiment 1, see Figure 1). In both that study and in this one, we found a monotonic, positively accelerated decline in accuracy over $n$. This outcome is contrary to the negatively accelerated decline predicted by McElree (2001, his Figure 1) but accords with data of Atkinson and Shiffrin (1968, Experiment 4). They observed a positively accelerated accuracy function (over short lags) similar to that found here and concluded that an accelerated trend of this sort will be seen whenever stimulus items have a very high probability of entering the temporary memory buffer. Atkinson and Shiffrin heightened this probability by requiring overt rehearsal in their experiment. Our task did not use overt rehearsal, but the $n$-back task requirement of an explicit response to each item obviously assures that each item will be attended to and stored.

Atkinson and Shiffrin (1968) attributed the accelerated trend to item interference and embodied their argument in a quantitative model of working memory. Following Atkinson and Shiffrin, we suggest that the similar trend in our study may result from the combination of error-free storage in the focus of attention and a steady decline for items stored outside the focus of attention, on
the assumption that the items outside the focus of attention are vulnerable to interference from all items, including those residing inside the focus. (The positive acceleration in Atkinson and Shiffrin’s data was actually the first segment of a larger sigmoid function. It seems likely to us that \( n \)-back data would trace a similar sigmoid if the range of \( n \) were extended beyond 5.)

Finally, note that the decline in accuracy with \( n \) cannot plausibly be attributed to a speed-accuracy tradeoff, both because latency did not increase as accuracy declined and because of McElree’s (2001) finding that the asymptote of the speed-accuracy function declined with \( n \), even with unlimited processing time.

**Performance at the End of Practice: The Sternberg Slide**

After 10 hr of practice with the task, the pattern of response times differed considerably from that of the first session (Figure 4). There was no longer evidence for privileged access in the \( n = 1 \) condition. Rather, the response time increased linearly over the range \( n = 1 \) to \( n = 4 \), with a shallow slope of 30 ms/n. We call this steady ramp function the Sternberg slide. There is a breakpoint at \( n = 5 \): The increment here is about twice as large as that extrapolated from the linear trend. These data are at odds with focus-of-one predictions on two accounts. First, it appears that with extended practice, participants were able to expand the focus of attention to accommodate four items. Second, retrieval from the widened focus appears not to be achieved through a content-addressable process.

This \( n = 1-4 \) accessibility function matches the outcome reported by Sternberg (1966) for a memory-scanning task. In the Sternberg task, participants are shown lists of items of varying length (the memory set). After the items have disappeared from the display, a recognition probe is given; the participant’s task is to indicate whether the probe was present in the memory set. (From our perspective, the Sternberg task can be described as a discrete-trial inclusive \( n \)-back task.)\(^1\) The Sternberg task typically yields results that are linear over set size with a slope of about 40 ms. Sternberg’s interpretation of this result was in terms of a serial, exhaustive (i.e., not self-terminating) search of the memory set. His interpretation has since been challenged (for reviews, see McElree & Dosher, 1989; Van Zandt & Townsend, 1993; for a series of distribution-based analyses, see Ashby et al., 1993). Current views interpret Sternberg-type search as a parallel, self-terminating process with very limited capacity.

The accessibility function recorded from Session 10 suggests that retrieval within the focus of attention may use a similar search process. Further confirmation comes from the Ex-Gaussian response time analysis (Figure 4). Hockley (1984, Experiment 1) investigated the ex-Gaussian signature of the Sternberg memory-scanning task. The linear load function was due entirely to change in the skew of the distribution, that is, \( \mu \) and \( \sigma \) were constant over set size, and \( t_{au} \) increased linearly. This was the same pattern found here. Our slope was very close to the 43 ms per item slope reported by Hockley for \( t_{au} \) in memory scanning.

The accuracy data also differed from those obtained in Session 1. Accuracy stayed high over the range \( n = 1 \) to \( n = 3 \) and dropped at \( n = 4 \) or \( n = 5 \). We interpret this as indicating that performance is error-free for items stored in the focus of attention and declines steadily for items stored outside the focus of attention. The convergent confirmation of the response time and accuracy data lends further credence to the claim that the focus of attention had expanded to four items by the end of practice.

---

1. Two reviewers pointed out a corollary to our claim of a malleable focus: An inclusive version of the \( n \)-back task, one in which participants are asked whether the item on the screen matches any of the items seen on the previous \( n \) trials, should yield the same focus as the exclusive version. In practice, however, the inclusive analog to the continuous \( n \)-back task is very awkward. We used a columnized version of \( n \)-back precisely because it has been shown to help participants to track the required lag. In the inclusive version, participants must lump together, not keep separate, the \( n \) stimuli. Pilot testing showed that crossing over columns in this way was very difficult. Our resource-allocation theory would then predict that the additional load of tracking item status might drain resources that otherwise might be used to support an enlarged focus. One simple way to salvage the inclusive version of the task would be to drop the continuous updating requirement and present \( n \) items followed by a probe. This is exactly the classic, discrete-trial Sternberg memory-scanning paradigm. Findings from this paradigm show an enlarged focus from the outset of training, in keeping with the reduced bookkeeping requirement and in support of the resource-allocation prediction. We consider memory-scanning data further in the Discussion section.

---

**Summary and Implications: The Malleable Structure of Working Memory**

Our data suggest that the focus of attention can be expanded to hold four items with practice, even in a task requiring serial attention, which initially limits processing to a single item. Four items is just the size postulated by mainstream working memory theories (e.g., Cowan, 2001). In our experiment, four seemed to be the maximum size attainable: Inspection of Figure 5 suggests that the focus was already fully expanded by the fifth session, and neither response times nor accuracy changed much from then on, implying a performance that was firmly at asymptote on the tenth session.

These practice effects counter claims that the size of the focus of attention is fixed at one, be it through structural separation (the tripartite-architecture hypothesis) or through the constraints of serial attention (the serial-attention hypothesis). Rather, the results support what we have called the allocation hypothesis, namely a variable focus of attention that can fluctuate between one and four, depending on the capacity demand of the embedded task. As the embedded task (in our case, a numerical identity-judgment task) becomes more automatized over practice (as evidenced in our study by the large decrease in response time for \( n = 1 \), which was reduced to about half of its original value), more attentional resources can be allocated to storage inside the focus of attention, thereby increasing its functional size.

The current study also provided evidence for differential retrieval dynamics in the focus of attention and the region of direct access. We attribute the differences to the operation of different retrieval processes in the two stores, content addressability in the region of direct access and parallel search in the focus of attention. The latter result is new. Evidence came from the ex-Gaussian decomposition of response times, which showed that the linear increase in response time over \( n \) in the expanded focus of attention was due to an increase in the skew of the distribution. This matches results obtained from ex-Gaussian decomposition of the Sternberg (1966) memory-scanning data, which is understood to involve a parallel and self-terminating search process (Murdock, 1971; Ratcliff, 1978; Van Zandt & Townsend, 1993). The value we obtained for the response time by \( n \) slope in the final session, 30 ms, is not only close to the typical slope of memory scanning but also to the...
typical slope obtained from subitizing research, which has likewise been ascribed to limited-capacity parallel processes (Trick & Pylyshyn, 1994). The striking resemblance in response time by load slopes of these three operations—our version of n-back, memory scanning, and subitizing—and the identical signature of the n-back and memory-scanning tasks in ex-Gaussian space suggest that the processing involved in all three tasks may be traceable to the same modus operandi within the focus of attention. This is a hypothesis that merits more research.

Accepting this hypothesis helps to resolve an anomaly in focus-of-one theorizing not hitherto recognized: If working memory outside the focus of attention is content addressable, tasks such as memory scanning should show flat profiles over working memory loads greater than one rather than positive slopes. Our perspective amends this expectation: We would expect to see positively sloped access functions over small load values and flat profiles over larger load values. Several memory-scanning studies have varied n over a range large enough to test this prediction, and we reexamined the data from these studies. Burrows and Okada (1975) varied list length from one to twenty. To our sight, the load function was not linear; rather, it showed a break. The initial segment, for \( n \leq 6 \), had a slope of about 40 ms per item; for \( n > 6 \), the slope was about 15 ms per item. Comparable results can be seen in results of Theios (1975). Data reported by Roebber and Kaernbach (2004) showed a breakpoint at \( n = 4 \) with a slope of 35 ms per item before the break and 5 ms per item thereafter (largest list length was 15). These two-segmented functions have a ready interpretation from our point of view. The first segment stems from an expanded focus of attention, and the second stems from a content-addressable store. The slope of the latter remains somewhat above zero as a result of averaging over focus differences.

Neuroimaging results provide further evidence for a dissociation between the focus of attention and the region of direct access: When small amounts of information (fewer than 6 items) are retained in working memory, the ventrolateral prefrontal cortex is active; for retention of larger amounts of information, the dorsolateral prefrontal cortex is further activated (Rypma, Berger, & D’Esposito, 2002; Rypma & D’Esposito, 1999), indicating a possible distinction between the focus of attention and the region of direct access. It remains to be explained, however, why in memory scanning the size of the focus seems to be five (plus or minus one) items rather than four (plus or minus one). We know of no study, however, designed to pinpoint the location of the breakpoint in the scanning task.

Although the present experiment rules out the strong version of the tripartite-architecture hypothesis (the claim that the focus of attention holds only a single item), the results are not incompatible with looser versions of that hypothesis. For instance, it is possible that Oberauer’s (2002) region of direct access is indeed restricted to four items, and comprises within itself the focus of attention, which, depending on allocation of attentional resources, can hold between 1 and 4 items. This hypothesis might in fact explain the distinctive response time patterns in Sessions 2 and 3. It is possible that as the focus of attention widens and the focus-switching effect diminishes in impact, elements stored in the region of direct access also become more accessible, thereby causing a larger drop in the response times for values of \( n = 4 \) or smaller. On this score, more detailed measurements are needed.

Another alternative explanation for our results needs to be considered. As noted in the introduction to this article, McElree (1998, 2001) proposed item clustering as a mechanism for sur-mounting focus-of-one limitations. If successive digits in our procedure were stored as a cluster rather than as distinct items, the focus could still be construed as containing a single item after practice—namely, the cluster. There are several reasons why this alternative seems unlikely. First, if this were the case, the updating and comparison process should now apply to the whole cluster. We would expect that updating and comparing a whole cluster would take longer than updating a single item, but this is not seen in the data. On the contrary, response times decreased sharply over the course of practice. Second, if clustering were the operating mechanism behind the data, the cluster size would seem to be limited to four items. One would still have to explain this limit (see Ericsson & Kintsch, 1995, for an overview of studies in which cluster size in working memory was shown to be virtually unlimited). Third, the retrieval dynamics that emerged in the latter phase of practice matched those obtained in studies of other, nonserial-attention, working memory processes, notably subitizing and memory scanning. A unified account of these phenomena would be preferable to one drawing on both clustering and a focus of four as piecemeal explanations. Under our theory, McElree’s (1998) finding of clustering within three-element semantic categories (see the introduction to this article) may have resulted from an expanded focus of attention, with the expansion perhaps due to semantic similarity. We note that the retrieval dynamics in McElree’s data (his Figure 3) are in accordance with our proposal of parallel search, giving rise to a ramped function within the focus of attention. We admit that this interpretation of the McElree findings is post hoc, and more research is needed.2

Finally, it is not clear whether our results will generalize to different tasks or to other variations of the n-back task. There is at

---

1 A reviewer suggested yet another alternative to our explanation of an expansion of focus. The reviewer argued that the focal span may be fixed at one item but that participants learn to switch focus to the relevant item in anticipation of the next stimulus. What changes over practice is the frequency of anticipatory switching. In the first session, one assumes that the focus rests randomly or drifts from one item to another, whereas in the final session, the focus is more often on target because of an anticipatory switch. The predictions of this model for Session 1 indeed entail a jump in RT from \( n = 1 \) to \( n = 2 \) as found in our data. The drift-model predictions over the full range of \( n = 1 \) to \( n = 5 \), however, depart from the data. To be specific, the model predicts a negatively accelerating curve of RT versus \( n \). When \( n = 1 \), no competing items are in storage, and consequently, no switching is necessary. When \( n = 2 \), there is a 50% chance that the correct item is in focus, and on 50% of the trials, a focus switch needs to be executed; therefore, the \( n = 2 \) time will be equal to the \( n = 1 \) time plus one half of the focus-switch time. When \( n = 3 \), there is a one-third chance of a correct focus, and a need for a switch on two thirds of the trials, so the \( n = 3 \) time will be equal to the \( n = 1 \) time plus two thirds of the focus-switch time, and so on. This prediction fails to make good contact with the Session 1 data. Rather than the flat slope found in both this study and the Verhaeghen and Basak (in press, Experiment 1) study, it predicts substantial longer response times than observed for \( n > 2 \). Consider now Session 10. The effect of practice may be embodied in an efficiency factor \( e \leq 1 \), which reduces the Session 1 switching probabilities. This configuration underpredicts response times for \( n > 3 \), and it does not predict the observed spline node at \( n = 4 \), with significant elevation of \( n = 5 \) above the regression line for \( n = 1 \). Of course, the failure of this particular implementation of a drift model does not establish the validity of our expanding-focus model; indeed other increasing-switch-efficiency models may be found to fit the data. (Interested readers can obtain more details from either of the authors.)
least one counterexample: Garavan (1998, Experiment 2) did not succeed in expanding the focus of attention in his participants through extensive training. He compared response times on non-switch trials and switch trials in a symbol-counting task. This is analogous to our \( n = 1 \) versus \( n = 2 \) manipulation. Garavan was not successful in reducing the switch costs to levels comparable to ours. There were several differences between the tasks that may be responsible for the difference in outcomes. Symbol counting times and switch costs were much longer than ours, suggesting that Garavan’s task was more difficult. After 12 sessions, the response times for the switch trials appear to be still declining (Garavan’s Figure 4). Perhaps, then, Garavan’s task was more resistant to automatization, and still more practice would have been necessary to free up substantial amounts of attention in his study. Further research is obviously needed to determine the limits of malleability in working memory.

In summary, our research suggests the following conclusions: (a) working memory contains at its core a zone of privileged access; (b) depending on the task and on the allocation of resources (which is partially a function of experience with the task), this zone can hold between one and four items; (c) items within this zone appear to be retrieved through a retrieval mechanism that is not content addressable, probably a parallel, self-terminating, limited-capacity search; and (d) items stored in working memory beyond this zone of direct access appear to be subject to direct, content-addressable retrieval. Given the adaptability of the configuration, the stroll down memory lane can be either a McElree strut or a Sternberg slide.

References


Schweickert, R. (1985). Separable effects of factors on speed and accuracy:


Received October 13, 2003
Revision received June 6, 2004
Accepted June 8, 2004

Call for Nominations

The Publications and Communications (P&C) Board has opened nominations for the editorships of *Clinician’s Research Digest, Emotion, JEP: Learning, Memory, and Cognition, Professional Psychology: Research and Practice*, and *Psychology, Public Policy, and Law* for the years 2007–2012. Elizabeth M. Altmaier, PhD; Richard J. Davidson, PhD, and Klaus R. Scherer, PhD; Thomas O. Nelson, PhD; Mary Beth Kenkel, PhD; and Jane Goodman-Delahunty, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2006 to prepare for issues published in 2007. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations also are encouraged.

Search chairs have been appointed as follows:

- **Clinician’s Research Digest**: William C. Howell, PhD
- **Emotion**: David C. Funder, PhD
- **JEP: Learning, Memory, and Cognition**: Linda P. Spear, PhD, and Peter Ornstein, PhD
- **Professional Psychology**: Susan H. McDaniel, PhD, and J. Gilbert Benedict, PhD
- **Psychology, Public Policy, and Law**: Mark Appelbaum, PhD, and Gary R. VandenBos, PhD

Candidates should be nominated by accessing APA’s EditorQuest site on the Web. Using your Web browser, go to [http://editorquest.apa.org](http://editorquest.apa.org). On the Home menu on the left, find Guests. Next, click on the link “Submit a Nomination,” enter your nominee’s information, and click “Submit.”

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Karen Sellman, P&C Board Search Liaison, at ksellman@apa.org.

The deadline for accepting nominations is **December 10, 2004**, when reviews will begin.