Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Special Issue: *Attention in Working Memory*
REVIEW

Attention in working memory: attention is needed but it yearns to be free

Stephen Rhodes and Nelson Cowan
Department of Psychological Sciences, University of Missouri, Columbia, Missouri

Address for correspondence: Stephen Rhodes or Nelson Cowan, Department of Psychological Sciences, University of Missouri, 210 McAlester Hall, Columbia, MO 65211. rhodessp@missouri.edu; cowann@missouri.edu

Recent theoretical development of working memory has emphasized the role of attention in several active processes supporting maintenance. Although this development is certainly welcome and has accounted for a number of phenomena, there are findings that cannot be readily accounted for through the active use of attention in refreshing or removal of information. We review these findings and suggest that, whenever the circumstances allow, participants attempt to reduce the load on attention by making use of stored concepts in long-term memory (LTM) or off-loading new configurations, forming new long-term memories. Newly formed groups and configurations in LTM constitute a list- or array-wide version of the consolidation of information into memory to prevent forgetting in a manner that reduces the need for continued attention to the material. This suggestion leads to a number of interesting questions at the behavioral and neural levels, which we also discuss.

Keywords: attention; working memory; refreshing; consolidation

Introduction

Varying definitions of working memory (WM) that can be found in the literature[1] mostly converge in describing the system as holding a small amount of information to support complex thought. Recent theoretical developments have described a role of attention in the maintenance of information in WM or the clearing of no longer relevant information from it. Several findings, however, remain difficult to reconcile with these proposed theoretical mechanisms. Here, we first selectively outline some of the recent theoretical developments. Second, we outline empirical findings that are difficult for the already-proposed mechanisms to account for. Third, and finally, we supplement the theoretical accounts with the proposal that observers will often try to offset the demands on attention, given the constraints imposed by the task at hand, making use of activated long-term memory (LTM).[2] Specifically, we suggest that long-term storage places a relatively low demand on attention and is brought into play not only passively but actively and strategically. A passive role for LTM in WM tasks is an apparent implication of the statement by Unsworth and Engle that "[i]tems that have been displaced from PM [primary memory] must be retrieved from SM [secondary memory]" (Ref. 3, p. 106). We suggest that people can do more, actively and strategically trying to create structures or groups of items in LTM to lessen the burden on attention. We term this active creation of LTM structures *off-loading* of the information. We suggest that this may occur for lists in situations requiring recall (a form of list-wide consolidation) and for arrays of stimuli that can be grouped into a configuration or new structure in LTM. Furthermore, we suggest that decay and interference can be counteracted with less of a demand on attention when there are more opportunities for off-loading of the memoranda to LTM. This proposal does not necessarily conflict with established ideas, but off-loading would join the arsenal of maintenance processes that also includes rehearsal, refreshing, and removal of distractor information. Presumably, off-loading is a preferred strategy when the task offers opportunities for the discovery of structure or chunks in the items to be remembered. We discuss tentative behavioral and

doi: 10.1111/nyas.13652

neural evidence in favor of this proposition and find exciting opportunities for future discovery.

The role of attention in maintenance

Decay-and-reactivation accounts: rehearsal and refreshing

Early suggestions that short-term memory decays[4] led to proposals as to how this decay could be counteracted. Not all of these proposals emphasize a role of attention. The best known is the phonological loop account of Baddeley and colleagues.[5] According to this account, the capacity of immediate verbal memory is set by the amount of information that can be repeated in around 2 s, before the information has irretrievably decayed. Subsequent investigations, however, suggested that additional mechanisms may be needed. For example, in studies of both developing children and adults, it was found that there are two speed factors that contribute to the prediction of verbal memory span. The two speed measures did not correlate with each other at all, but both correlated well with memory span, together accounting for the age differences in span.[6] One speed measure was rapid recitation, presumably related to the efficacy of verbal rehearsal. The other was the rate of recall of items, with the per-item rate slowing linearly as a function of the list length and being, for any particular list length, slower in younger children.[7] This measure was presumably related to the efficacy of mentally searching through the list to identify the item to be recalled next. Given that a fixed period of recall was not obtained, but rather longer recall of span-length lists in children with higher span, it also was suggested that the search process reactivates items, counteracting their decay.[8,9] In terms of a modeling framework, the assumption at least implicitly was that this reactivation occurred through circulation of the focus of attention among items to be retained for recall.[2]

This concept of covert retrieval has been greatly elaborated in the time-based resource sharing (TBRS) framework of Barrouillet, Camos, and colleagues. According to TBRS, memory for items can decay during diversions of attention but this decay is prevented when items are serially refreshed through attentional focusing on the items.[10] The theory is usually applied to complex span tasks, in which processing episodes, varying in speed from one trial type to the next, are placed between the presentations of items making up a list to be recalled. Span is defined as the length of list that can be recalled after the processing has been completed successfully. Perhaps the strongest evidence for the proposal of decay and refreshing is the *cognitive load function*, a negative linear relation between the proportion of time that attention is occupied by a concurrent processing task and memory span, a relation found in complex span paradigms across a variety of stimulus modalities.[10–13]

That being said, there are asymmetrical dual-task effects on visual and verbal information, with less verbal forgetting,[14] an asymmetry that can be explained in the current TBRS theory inasmuch as verbal information can be retained by both attentional (refreshing) and nonattentional verbal (rehearsal) mechanisms, whereas no analogous nonattentional mechanism exists for nonverbal information.[15]

In summary, two basic concepts have proven quite powerful in accounting for change in memory performance in complex span tasks, which include distraction: decay offset by refreshing, except when the distraction (cognitive load) prevents it, and rehearsal for verbal memoranda, except when the articulatory demands prevent it. Although the latest instantiations of TBRS[16,17] have incorporated additional mechanisms, decay, refreshing, and verbal rehearsal remain central to this framework.

Interference-removal accounts

In contrast to theories positing decay and restoration, an alternative class of models holds that information is lost through various forms of interference rather than passive decay.[18,19] In the recent model devised specifically for the complex span task by Oberauer and colleagues,[20] interference comes from two sources: confusion with other recall targets and distortion of the bindings between items and their serial position caused by superposition—the encoding of distractors into WM in ways that perturb the serial position binding of items to be recalled. To account for the cognitive load function,[11] there is an active removal process in which a single-item focus of attention protects representations of the memoranda by degrading the interfering binding of distractor representations to their serial positions.[21,22] Thus, in this theory, which has no decay process, free time is used to clear distractors out of WM. The model of Oberauer *et al.*[20] reproduces not only the cognitive load function but also more intricate patterns in complex span data, such as patterns of transpositions and serial position functions.

Limitations of refreshing and removal

There has been no clear resolution in the decay + refreshing versus interference + removal debate. Several problems exist for each mechanism. For example, the act of serial refreshing has proven difficult to identify experimentally.[23] Further, in computational instantiations of refreshing, it seems that serial refreshing at the speed typically assumed in the literature (∼50 ms/item) is unable to reproduce patterns of performance seen in humans.[24,25] Rather, a regimen of grouped refreshing may be more appropriate, possibly akin to rotating a multi-item focus of attention around items in the memory list (cf. Ref. 26). It is also a challenge for a decay + restoration account to explain why the relationship between speed and span includes not only speeds that involve mnemonic restoration of memory but also the speed of item identification. In particular, measures of articulation rate and identification time appear to contribute independently to the development of span during childhood.[27,28] It may be that item identification speed is related to the speed with which items can be attentionally refreshed but, to our knowledge, the extant data do not yet substantiate such a proposal.[29]

It is difficult for interference principles to explain some other phenomena, in particular several recent reports pointing to the passive decay of information in at least two modalities for materials that may be difficult to encode well: visual (Refs. 30–32, but see Ref. 33) and auditory.[34,35]

Additionally, in the next section, we present evidence that there is a benefit of silent periods before a distractor has been presented, obviously a problem for a distractor-removal account but leading to the notion of memory consolidation.

The role of WM consolidation

The concept of *WM consolidation*—increased stability of a representation following more protracted initial processing of the item(s)—complicates accounts of WM maintenance but helps to resolve unanswered questions. Consolidation serves to strengthen the representation of memory items, even following a mask.[36] It is disrupted by the processing involved in two-alternative forced-choice tasks, and takes time (>1 s for four letters[37]).

Importantly, more opportunity for consolidation appears to reduce the rate of information loss from WM. Ricker and Cowan[31] presented three abstract characters either sequentially or simultaneously in a probe-recognition procedure and found greater loss of information over 12 s for simultaneous presentation. However, when the amount of time to process each item was matched, allowing equal time for consolidation, the rate of decay was similar between the two study formats.

Consolidation adds a layer of complexity to the TBRS model, as it is hard to disentangle the effect of consolidation from decay and refreshing. It may be that the time after an item is used to refresh the representations of items, or it may be that the time is used to form a better representation that decays more slowly, allows quicker refreshing, or is more resistant to interference. Indeed, proponents of TBRS have added a consolidation mechanism to begin to account for the effects of encoding time, as the simple idea of decay + refreshing was not enough to account for such effects.[16,17]

To take the concept of consolidation further, improved WM consolidation may lead seamlessly to LTM consolidation that also plays an important role in WM tasks. In stark contrast to the typical limit of WM to a handful of items when a small pool of items is used repeatedly from trial to trial, creating strong proactive interference in LTM, the measured capacity can be much larger when the items are unique on every trial, allowing better use of LTM (e.g., Ref. 38). Thus, participants can search arrays to find any of a large number of real, known objects that can be part of the target set at the same time, even 100 of them, though with a gradual transition from what appears to resemble serial retrieval to a mode that seems closer to parallel retrieval.[39,40] When list items do not repeat, the recognition of them does not show a small capacity limit (e.g., Ref. 41), in stark contrast to the limit to about three or four items when the items can recur from trial to trial.[42–45]

Recently, researchers have tried to disentangle consolidation from the maintenance processes we described above (verbal rehearsal, attention-based refreshing, and removal of interference). Work with the complex span paradigm suggests that they all may be independently identifiable processes. Bayliss *et al.*[46] either provided a blank interval before the onset of a processing sequence or burst (i.e., distraction) following each memory item, where it could allow consolidation before processing, or placed the interval after the processing burst. Performance was reliably better with the blank interval before the processing burst. This timing difference occurred even during articulatory suppression, ruling out rehearsal, and it occurred regardless of cognitive load, at odds with what is predicted by the basic notion of decay offset by refreshing, considered alone (see also, Ref. 47).

These findings appear to be fairly problematic for current theories based on the concept of distractor removal, as performance was improved by giving time before the distractors rather than after, useless for removing the distractor-position bindings. As Bayliss and colleagues note, according to the interference model,[20] the encoding of each item should have been complete in all of their conditions, so a mechanism of item strengthening appears to be needed. Of course, these objections do not rule out the possibility that removal of distractors supplements other maintenance processes. The possibilities we outline below may prove helpful in characterizing the possible effects of removal and provide further questions regarding the role of free time during complex span trials.

To summarize, cognitive load effects (e.g.,. Ref. 11) and patterns of interference between multiple sets of memory items (e.g., .Ref. 14) clearly show that attention is important in the maintenance of information in WM. However, as briefly reviewed above, the exact nature of attention's role in active maintenance is unclear, and potential additional factors, such as consolidation or strengthening of memory representations in free time, must be considered.

While attention is clearly important, here we wish to outline an additional factor that may play a role in performance, emphasizing how it might assist WM while minimizing the need for attention. Namely, we propose that, whenever the parameters of the task and stimuli permit, participants make use of mechanisms that are not so attention dependent, supplementing performance through off-loading of information to activated LTM. We make this proposal within a specific framework, the embedded processes model of WM.[2] However, as we will discuss, off-loading is related to other suggestions in the literature and viable within other modeling frameworks too.

Formation of activated LTM representations and freeing up of attention in WM tasks

So far, we have emphasized the mechanisms of refreshing of memoranda and removal of distracting information, both of which may be accomplished only with the help of attention. The attention demands of verbal rehearsal, should it be involved in span, appear to vary depending on maturity or learning of the material.[48] What we now wish to suggest is that it is in participants' interest to do as much maintenance as possible using attention-free aspects of processing, leaving attention free for other demands. The main way we believe that this off-loading of processing out of the focus of attention can occur is through rapid memorization of the memory set, theoretically placing the information in the activated portion of LTM, where it is easy to retrieve for a short time as needed. For example, the seemingly random letters *L, Q,* and *R* presented in a list might be recoded as the consonants within the word *liquor*, making long-term retention easier. Alternatively, even if no familiar pattern is recognized, it may be possible to form a new sequence (*LQR*) that is entered into activated LTM.

According to our framework,[2] the activated portion of LTM is a subset of LTM that is in a state of heightened accessibility but is not currently being processed in the focus of attention. This proposal of off-loading material from the focus of attention is similar to mechanisms within other theories with combined contributions from a primary, attention-based store and a secondary, long-term store searchable at test (e.g., Refs. 3 and 49–51). However, in these accounts, LTM is always formed when items are displaced from WM by concurrent processing (because the number of items to remember exceeds the scope of attention or is lost from temporary buffers). The present suggestion is of a more active process in which schematic structures are formed to improve the representation of the items to be remembered, reducing the need for attention to maintain the information. The structures that are formed can be either recognized from LTM or newly constructed to simplify the representation. Off-loading works for immediate recall, because proactive interference between trials is not an insurmountable problem, as it would be for later recall.

One particular way in which such memorization can occur is through a continued consolidation process that includes properties of the relationships of items to one another to form a configuration, either temporal in nature for lists or spatial in the case of concurrent arrays of stimuli. Next, we explore the tentative evidence for these kinds of set-wide off-loading.

One kind of off-loading: list-wide consolidation?

One important possibility is that the process of consolidation is not limited to the most recently presented item and that consolidation should be redefined to be list-wide. One type of evidence consistent with this proposal is that, in complex span–type studies, reaction times to the processing task slow as more items are added to the memory set.[52–55] This slowing could either reflect a greater demand to refresh the sequence with each new item or, as we suggest, a demand to integrate the new item with representations of those presented previously. Given the demand to retain serial order in these tasks, this kind of list-wide consolidation may prove extremely important in determining the role of attention in maintenance and, consequently, the patterns of decay or interference observed.

Much of the previous work on consolidation has focused on the strengthening of individual items for single-probe tasks.[30,31] As evidence of this potential list-wide strengthening, Vergauwe *et al.*[55] found that the first reaction time in a burst of processing was sensitive to the number of items to be held in memory in a preload procedure in which all memory items were presented before the onset of the processing task. Using trials in which the memory sequence was recalled correctly, they found that each additional memory item, up to four, slowed processing by approximately 250 ms (for convergent evidence, see Ref. 37). Participants may use the time to form a serial chain or to establish bindings between items and their serial positions. Even the latter is likely to be a relational type of information, as the evidence suggests that items are associated not with absolute serial positions but with relative position in the list.[56] This new relational structure may then be off-loaded to the activated portion of LTM, according to the general framework in which we are working,[2] reducing the load on attention.

Some attention is presumably needed to revisit the structure periodically, refreshing it to protect it from interference or possibly decay, improving the structure further by attending to some of its additional details. Such off-loaded information likely also requires a form of controlled search when memory is probed. Thus, while off-loading is assumed to lessen the load on attention, freeing some attention up for other activities, it is not assumed to come at zero cost. We return to this issue in the next section when discussing recent findings regarding performance when two modalities must be retained, but here let us simply reiterate our proposal: there is a cost in sharing attention between two LTM structures, but more can be remembered that way than by trying to hold the same information constantly in the focus of attention.[57]

The consideration of list-wide consolidation as a kind of off-loading leads to an interesting set of questions for future experiments. Namely, is free time equally helpful at all stages in a trial? Bayliss *et al.*[46] appear to show that free time is more beneficial before a burst of processing. To what extent is that benefit specific to points late in the list presentation, after most or all of the list has been presented, permitting a list-wide representation to be formed? To find out, a complex span task could include trials in which more time before processing is either given at the beginning of the memory list or toward the end. If more free time before processing episodes is found to be most beneficial for memory items near the end of the list, the explanation may require consolidation that goes beyond individual items and is instead list-wide. The same pattern of results would be difficult to explain via removal of distractors from the memory representation, as there is currently no reason to expect that this removal is more important toward the end of the memory list.

Off-loading of arrays: consolidation of a spatial configuration?

Moving away from complex span, recent experiments assessing concurrent storage of two modalities have yielded somewhat surprising results pointing toward the consolidation and off-loading of a simultaneously presented array. Cowan *et al.*[57] presented participants with two sets of material: a visual array of colors and a sequence of digits, in either order. In some trial blocks, participants had to remember both sets, whereas in other blocks they had to remember only one modality and ignore the other. Using a simple processing model, they were able to separate and quantify *central* and *peripheral* storage. The central portion was estimated as the number of items sacrificed from WM of one modality when the other modality also had to be retained concurrently, and the peripheral portions were the numbers of items reliably retained in a modality regardless of whether the other modality had to be retained at the same time. The central portion of WM proved to be somewhat smaller than in previous, not-as-well-controlled studies.[58] It appeared to contain only around one item, whereas the peripheral components included around two items per modality. One interpretation of these findings is that participants off-load the first set of memoranda to activated LTM,[2] freeing up the focus of attention to accept the second set with minimal interference. Subsequently, the second set might also be off-loaded. The central portion of WM would reflect the attention cost of maintaining the visual and verbal structures in activated LTM concurrently.

This notion of set-wide off-loading was recently applied in a developmental setting by Cowan *et al.*[59] Specifically, applying the general method of Cowan *et al.*,[57] it was found that peripheral storage elements increased from ages 6 to 13 and into adulthood. The central portion, though, was relatively constant with age or even decreased. This finding raises the intriguing possibility that developmental change is not driven primarily by growth in capacity per se (as suggested previously in Refs. 60 and 61) or knowledge (ruled out in the case of Ref. 62), but by a greater ability to free up attention by increasingly relying on less attention-demanding storage, as in the activated portion of LTM. Developmental mastery of the ability to off-load information may be manifest as improved mechanisms of chunking or organizing information into groups that may be better handled by structures in LTM,[63] similar to the notion of list-wide consolidation discussed above but, for concurrent spatial arrays, in spatial configurations rather than temporal sequences.

The aforementioned developmental findings may provide insight into the possible nature of off-loading. They point to off-loading being a self-initiated, strategic process that participants use (akin to the often-reported strategy of participants linking memoranda to concepts already present in memory, such as remembering letters by linking them to loved ones' initials). Such strategic use might be related to the development of metacognition throughout childhood (e.g., Ref. 64). The proposal is that young children may not understand the need to create effective structures from the stimuli and therefore may lose more information than older individuals when the capacity of attention is exceeded. While the growth in knowledge per se cannot explain the development of WM,[62] growth in the effective use of that knowledge to support active maintenance seems worth investigating.

The initial findings open up other questions for future work. For instance, it has proven difficult to obtain evidence that participants can learn arrays of visual stimuli, such as those used in the studies of central and peripheral WM by Cowan *et al.*[57,59] Strikingly, even if arrays of visual objects are repeated on every single trial (a type of Hebb repetition procedure, in which one looks for improvements in performance due to learning of sets repeated in multiple trials), there seems to be little to no improvement over trials in change detection accuracy.[65] One prediction stemming from the above is that we should be most likely to observe such a learning effect when participants are incentivized to off-load an array to activated LTM to free up attention for a second memory set. The only relevant evidence of which we are aware concerns sequence learning, for which no difference has been found in Hebb repetition learning between simple span and complex span.[66] However, Hebb repetition effects were present for both tasks and, given the requirement to retain order in these tasks, it is possible that participants are engaging in some of the processes described above that serve to consolidate a list into LTM, even in the absence of distraction.

Other behavioral support for LTM involvement and off-loading

One advantage of the proposal of off-loading as a mechanism of WM maintenance is that it is consistent with a flood of other findings indicating that LTM plays a role in WM tasks. We already have noted that known items can be remembered and used far beyond the capacity limit of several items.[38–41] There are also findings suggesting that new long-term memorization plays a role in WM retrieval. In particular, retrieval actually appears to take longer when there is more useful, meaningful off-loaded information to be retrieved. For example, Cowan *et al.*[67] found that, among complex span tasks, the pace of recall was much slower when there was meaningful material to be recalled (in reading and listening span tasks), compared with the recall of random series of sums (in counting span). Suggesting more use of LTM when there are more demands on attention, delayed recall is better for complex span relative to simple span,[68] and the benefit for delayed recall appears to depend on the demand of the task.[69] These results could indicate attempts to off-load memory items to activated LTM in preparation for the concurrent processing task.

Unsworth and Engle[3] proposed that retrieval from secondary memory (i.e., of items displaced from attention) helps to account for individual differences in WM insofar as attention is needed to carry out this retrieval. Although we are proposing that the purpose of off-loading is to reduce the strain on attention, it does make sense that attention is needed for retrieval of the information. Thus, the present proposal and that of Unsworth and Engle are compatible.

Neurological support for off-loading

Considerable recent neuroscientific evidence supports the notion that activated LTM is involved in WM tasks. The concept of long-term activation that has been proposed has moved at least temporarily away from the idea that it consists of persistent neural firing (e.g., Ref. 49) and toward the notion that synaptic weights are at least temporarily changed. Arguing against persistent firing, Lewis-Peacock *et al.*[70] used a task in which multiple types of stimuli had to be retained on the same trial, with cues to use some information right away in an upcoming recognition task (e.g., a word) and sometimes to save other information for another such task later in the trial (e.g., bar orientations). Multivoxel pattern analysis (MVPA) indicated that the type of information needed immediately was active but that the information only needed later in the trial, though still available, was preserved in some form that did not show up as an active pattern. Rose *et al.*[71] found that dormant MVPA patterns could be reactivated using transcranial magnetic stimulation. To relate their physiological theory to the cognitive concept of the activated portion of LTM, the latter could be specified as synaptic weights rather than active patterns of neural firing.

Using event-related potentials, Reinhart and Woodman[72] made a related point about gradual off-loading of a template that was repeated from trial to trial. They presented two targets to be held in WM while searching arrays to find either of those targets. As the same targets were used on multiple trials, the index of active storage in WM (contralateral delay activity; CDA) decreased, while an index of LTM retrieval (a positive-going wave called P170) increased; but the declining CDA was reversible with motivation and attention when a large reward for success was provided. Thus, it seems that observers are able to control the relative use of activated LTM and the focus of attention in response to task demands and rewards, consistent with the proposed strategic use of off-loading.

Finally, Wallis *et al.*[73] used magnetoencephalography, a technique with good spatial and temporal resolution, to show another phenomenon suggesting the strategic use of an activated portion of LTM. When a post-cue is presented to reduce the number of items from an array that have to be held in WM, several indices of attention show a transient burst, with a behavioral benefit that comes after the burst of attention has disappeared. This pattern is just what would be expected if the role of attention was to help somehow in the set-wide consolidation of information into what we have termed the activated portion of LTM.

More work is needed to determine the nature of activated LTM: is it limited to mechanisms other than active neural firing, such as synaptic weights,[71] or would a closer look reveal that items needed later in the trial have been saved with, for example, specialized neural firing in a circuit that includes the hippocampal areas?

Open questions and speculations regarding off-loading

We have argued for a process of off-loading in which structures in LTM (either existing or rapidly created) are used to form representations of the to-be-remembered items in order to reduce the load on the focus of attention at crucial points, freeing attention for other processing. We recognize many open questions regarding the specifics of off-loading (e.g., the capacity limits for such off-loading, how long it takes to off-load information, and how off-loaded information is lost or forgotten). The embedded-processes framework provides some direction and potentially useful constraint for the necessary further research. According to our conceptualization, all off-loaded information was once within the focus of attention, the component from which information is off-loaded. The capacity limitation of the focus of attention thus places a limit on the capacity for off-loading. In the case of briefly presented visual arrays (or lists with rehearsal prevented), this limit would be 3–4 items, the number of items that can be apprehended in such procedures.[42–45]

For sequentially presented lists, the focus of attention is assumed to fill up one by one with each item and, when time is available, a search is presumed to take place for concepts in LTM that can be related to the memoranda. The speed and efficiency of this process will clearly depend on the ease with which an identifiable concept can be retrieved or created, which in turn will depend on the experience and expertise of the observer with the particular material used.[51] For example, a known letter sequence like *FBI* would be quickly identified and off-loaded, whereas, to reuse the example from above, *LQR* would require additional search to produce the token *liquor*. If no pattern is observed, perhaps sometimes a new sequence can be quickly learned by rote, although this would presumably require more time and might form a weaker trace. We presume that the finding of increasingly delayed processing latencies with increasing memory load (e.g., Refs. 52–55) reflects the time taken both to identify concepts or configurations and to carry out any other process needed to consolidate or off-load the information. Further, inasmuch as the focus of attention is limited, we adopt the proposal of previous researchers (e.g., Refs. 3 and 50) that information may also become displaced from the focus to activated LTM by additional memoranda or by distraction. Future work could usefully examine the capacity and temporal constraints of off-loading within a complex span procedure by varying list length as well as the placement of free time within a trial, as suggested above. Similarly, within a dual-modality array procedure (e.g., Refs. 57 and 59), the interarray and poststimulus durations can be varied.

Off-loading serves to reduce the load on attention at crucial times, but it is not assumed to be cost free. Specifically, off-loaded representations are still presumed to suffer decay and interference in the activated portion of LTM. The extent of this information loss will be diminished insofar as a stable structure has been identified or formed from the to-be-remembered material, as exemplified in Ricker and Cowan's finding of decay for abstract symbols but not for letters.[30] To counteract this information loss, the focus of attention is assumed to periodically return to the off-loaded information to refresh it, consistent with recent computational simulations supporting the idea of grouped refreshing.[24,25] An additional source of off-loading cost may come, as Unsworth and Engle[3] have suggested, from the requirement to conduct a controlled search through the contents of activated LTM at test.

The attention cost of off-loading and later retrieval can theoretically come at encoding, maintenance, or retrieval of a stimulus set. We believe that our recent work with dual-set memory showing a central portion of WM[57,59] indicates an effect of attention specifically during maintenance. To illustrate this, let us consider encoding, retrieval, and maintenance in turn. One potential point of conflict between sets could occur if the second set is poorly off-loaded into LTM (what we are calling here an encoding effect) because of the concurrent maintenance of the first set, but this possibility cannot explain why the dual-task cost also falls on the first set. Another possibility is that attention is used at retrieval from activated LTM, but, after the retrieval cue is presented, one set can be forgotten and only one needs to be retrieved, so a dual-task cost might not be predicted for retrieval. In maintenance, however, both sets would have representations needing refreshment or improvement for maximal performance, splitting attention between sets. This need for some attentional involvement is assumed to underlie the one-item central storage portion found by Cowan *et al.*,[57,59] as the focus of attention had to flit between the two sets of information (one visual and one verbal).

In summary, there are various open questions regarding the nature of off-loading. Although the suggestion that LTM plays a role in WM task performance is certainly not new (e.g., Refs. 3 and 49–51), we hope that the observations offered above, with the constraints provided by the embedded processes approach, provoke some new avenues of investigation. While much remains to be learned about the putative mechanism of off-loading, we argue that it can address some of the shortcomings of the established theoretical mechanisms.

How does the concept of off-loading improve our understanding of WM maintenance?

As outlined at the beginning of this review, there are some findings that cannot be fully addressed through the currently dominant theoretical mechanisms of decay offset by refreshing/rehearsal or the prevention of interference via removal of distractors. It is worth asking to what extent the current proposal helps in addressing these limitations. In regard to WM development, the concept of off-loading may help us to understand the finding that measures of identification speed predict the growth of span, independent of other speed measures used to index active maintenance.[27–29] Identification speed may be an index of the strength of LTM representations that can be used for off-loading. Together with recent developmental findings regarding memory for simultaneous arrays of different modalities,[57,59] we may speculate that improvement in the ability to utilize existing structures in LTM to free up attention for other activities underlies some of the development in WM during childhood.

The idea of off-loading may help to explain evidence for a rate of decay that is quite variable between situations, in contrast to the original TBRS assumption. The intermittent finding of decay (see Refs. 30–32,34, and 35) is somewhat a problem for the basic decay + refreshing account, which does not incorporate a means of differential decay (although the TBRS model now has other mechanisms that may allow for this; e.g., Refs. 16 and 17). The idea of off-loading may help because activated LTM following set-wide consolidation might not suffer passive decay to the same degree as items held in the focus of attention. Although Cowan[2,49] proposed that it is activated LTM that decays, subsequent findings suggest that that kind of decay is minimal following good consolidation, whereas it takes vigilance to avoid items dropping out of the focus of attention. Thus, the extent to which items cannot be meaningfully off-loaded from the focus may determine the amount of decay or interference observed. Reports of passive decay have relied on abstract material (such as unfamiliar complex symbols[30–32] or tones differing in timbre[34,35]) for which it is conceivably quite difficult to carry out rapid encoding and consolidation in a manner that draws on concepts in LTM. For memoranda that enable the rapid identification or formation of a long-term trace, we may expect to observe less evidence of decay or interference because of distinct, stable representations in activated LTM.[2] Thus, we do not see the concept of off-loading as abolishing the theoretical concepts of decay + refreshing or interference + removal, but rather modulating the extent to which their effects will be observed.

Off-loading might be of interest as a possible addition to interference + removal models as well. To the extent that off-loading occurs, there may be little need for removal of distractors, as they are presumably not included in the off-loaded representation of the set of memoranda. The slowing of concurrent processing reaction times as more items must be added to the list representation[52–55] points to the possibility of a list-wide consolidation process (see above), which would serve to reconsolidate the list anew with the addition of each new item without distractors in the representation. There is some support for this notion of consolidation without distractors in a recent study assessing the Hebb repetition effect in complex span. Oberauer *et al.*[66] found that participants became increasingly accurate for repeated sequences of memory items and, crucially, this occurred even though the complex span distractors did not repeat. Thus, participants appeared to form a long-term trace without the distractors present, reducing the need for removal of the distractors from the representation.

Conclusions and caveats

We have argued that three often-discussed mechanisms for the maintenance of information in WM (rehearsal, refreshing, and removal of interference) cannot account for all of the relevant evidence, either alone or in combination. First, there have been arguments that rehearsal may play no role;[74] other sorts of verbal interference phenomena could account for the effects of articulatory suppression that have been taken to indicate a role of rehearsal. Second, it has been difficult in some situations to observe attention-based refreshing processes in complex span directly;[23] they have mostly been observed indirectly via the effects on cognitive load. Third, removal of information from distracting processing tasks cannot explain the benefit of a period for consolidation or refreshing before the first processing episode.[46,47] In response to such limitations, we have proposed that there is growing evidence that a method of off-loading information from the focus of attention to the activated portion of LTM takes place and is a maintenance mechanism of choice when it can be used, given that it minimizes the strain on attention (e.g., Refs. 57,59, and 70–73). This strategic off-loading is somewhat different from previous proposals in which information is just displaced from an attention-based store to a long-term store by distraction or overloading (e.g. Refs. 3 and 49–51).

The present proposal of off-loading information on a list-wide or set-wide basis is in its early stages, and we have made several suggestions that point to potentially fruitful areas for future research. To summarize these suggestions: assessing when in complex span trials free time is most productive will be useful to follow up on the findings of Bayliss *et al.*,[46] as the notion of list-wide consolidation predicts an increasing benefit for free time at later points in the list presentation. Hebbian learning effects with or without concurrent distraction are also a potential area of exploration. It may be more likely that researchers will observe learning of a list or array to be remembered (via off-loading) when participants can anticipate that attention must soon be deployed to another upcoming task. Finally, relatively new analysis techniques for neuroimaging data are beginning to shed light on the fate of representations throughout the progression of a trial, and off-loading may explain findings of dormant MVPA patterns.[70–73] It may be that these representations are truly activity silent or, alternatively, the initial perceptual representations, which are typically the focus of MVPA, may have been restructured, via the off-loading process, for storage in regions more typically associated with LTM (e.g., the medial temporal lobes).

Even if this kind of off-loading mechanism can be demonstrated and shown to be an integral part of WM task performance, questions remain regarding the definition of WM[1] and consequently of the tasks that are taken to reflect WM. There are also some questions about whether the broad notion of attention is an oversimplification of different strains of attention thrown together, such as central versus visual attention,[75] or whether there is enough interaction between types of attention so that the broad notion of the focus of attention is ultimately apt.[76,77]

Regarding the relationship between WM and LTM, there have been challenges to the distinction between the two constructs (e.g., Refs. 19 and 78). We assert the distinction between the two (e.g., see Refs. 49,50, and 79) but, at the same time, also acknowledge the need to think of them as richly interacting systems. It seems worthwhile to revisit a point of view articulated by Broadbent in 1971 (Ref. 80, p. 342–343):

> 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.

We endorse Broadbent's view, but in a modified form in which decay from WM is something that occurs when there is insufficient time to establish a useful representation of the identifying characteristics of the information in WM. This modified view emphasizes a distinction between the fate of activated information that has not had the benefit of sufficient processing in the focus of attention[2] and will decay, versus information that has been attended to the point at which a rich off-loading can occur and information is more stabilized.

Acknowledgments

This work was supported by the UK Economic and Social Research Council (ESRC), Grant ES/N010728/1, and NIH Grant R01-HD21338.

Competing interests

The authors declare no competing interests.

References

1. Cowan, N. 2017. The many faces of working memory and short-term storage. *Psychon. Bull. Rev.* **24:** 1158–1170.
2. Cowan, N. 1988. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychol. Bull.* **104:** 163–191.
3. Unsworth, N. & R.W. Engle. 2007. The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* **114:** 104–132.
4. Brown, J. 1958. Some tests of the decay theory of immediate memory. *Q. J. Exp. Psychol.* **10:** 12–21.

5. Baddeley, A.D., N. Thomson & M. Buchanan. 1975. Word length and the structure of short term memory. *J. Verbal Learning Verbal Behav.* **14:** 575–589.
6. Cowan, N., N.L. Wood, P.K. Wood, *et al.* 1998. Two separate verbal processing rates contributing to short-term memory span. *J. Exp. Psychol. Gen.* **127:** 141–160.
7. Cowan, N., J.S. Saults & E.M. Elliott. 2002. The search for what is fundamental in the development of working memory. In *Advances in Child Development and Behavior*. R. Kail & H. Reese, Eds.: 1–49. Elsevier.
8. Cowan, N. 1992. Verbal memory span and the timing of spoken recall. *J. Mem. Lang.* **31:** 668–684.
9. Cowan, N., T. Keller, C. Hulme, *et al.* 1994. Verbal memory span in children: speech timing clues to the mechanisms underlying age and word length effects. *J. Mem. Lang.* **33:** 234–250.
10. Barrouillet, P., S. Bernardin & V. Camos. 2004. Time constraints and resource sharing in adults' working memory spans. *J. Exp. Psychol. Gen.* **133:** 83–100.
11. Barrouillet, P., S. Portrat & V. Camos. 2011. On the law relating processing to storage in working memory. *Psychol. Rev.* **118:** 175–192.
12. Vergauwe, E., P. Barrouillet & V. Camos. 2009. Visual and spatial working memory are not that dissociated after all: a time based resource sharing account. *J. Exp. Psychol. Learn. Mem. Cogn.* **35:** 1012–1028.
13. Vergauwe, E., P. Barrouillet & V. Camos. 2010. Do mental processes share a domain general resource? *Psychol. Sci.* **21:** 384–390.
14. Morey, C.C., R.D. Morey, M. van der Reijden & M. Holweg. 2013. Asymmetric cross-domain interference between two working memory tasks: implications for models of working memory. *J. Mem. Lang.* **69:** 324–348.
15. Camos, V., G. Mora & K. Oberauer. 2011. Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. *Mem. Cognit.* **39:** 231–244.
16. Barrouillet, P., G. Plancher, A. Guida & V. Camos. 2013. Forgetting at short term: when do event-based interference and temporal factors have an effect? *Acta Psychol.* **142:** 155–167.
17. Barrouillet, P. & V. Camos. 2015. *Working Memory: Loss and Reconstruction.* London: Taylor and Francis.
18. Brown, G.D.A. & C. Hulme. 1995. Modeling item length effects in memory span: no rehearsal needed? *J. Mem. Lang.* **34:** 594–621.
19. Nairne, J.S. 2002. Remembering over the short term: the case against the standard model. *Annu. Rev. Psychol.* **53:** 53–81.
20. Oberauer, K., S. Lewandowsky, S. Farrell, *et al.* 2012. Modeling working memory: an interference model of complex span. *Psychon. Bull. Rev.* **19:** 779–819.
21. Ecker, U.K., K. Oberauer & S. Lewandowsky. 2014. Working memory updating involves item-specific removal. *J. Mem. Lang.* **74:** 1–15.
22. Oberauer, K. & S. Lewandowsky. 2016. Control of information in working memory: encoding and removal of distractors in the complex–span paradigm. *Cognition* **156:** 106–128.
23. Vergauwe, E., K.O. Hardman, J.N. Rouder, *et al.* 2016. Searching for serial refreshing in working memory: using response times to track the content of the focus of attention over time. *Psychon. Bull. Rev.* **23:** 1818–1824.
24. Lemaire, B., A. Pageot, G. Plancher & S. Portrat. 2017. What is the time course of working memory attentional refreshing? *Psychon. Bull. Rev.* https://doi.org/10.3758/s13423-017-1282-z.
25. Portrat, S. & B. Lemaire. 2015. Is attentional refreshing in working memory sequential? A computational modeling approach. *Cogn. Comput.* **7:** 333–345.
26. Gilchrist, A.L. & N. Cowan. 2011. Can the focus of attention accommodate multiple separate items? *J. Exp. Psychol. Learn. Mem. Cogn.* **37:** 1484–1502.
27. Case, R., D.M. Kurland & J. Goldberg. 1982. Operational efficiency and the growth of short term memory span. *J. Exp. Child Psychol.* **33:** 386–404.
28. Hitch, G.J., M.S. Halliday & J.E. Littler. 1989. Item identification time and rehearsal rate as predictors of memory span in children. *Q. J. Exp. Psychol.* **41A:** 321–337.
29. Dempster, F.N. 1981. Memory span: sources of individual and developmental differences. *Psychol. Bull.* **89:** 63–100.
30. Ricker, T.J. & N. Cowan. 2010. Loss of visual working memory within seconds: the combined use of refreshable and non-refreshable features. *J. Exp. Psychol. Learn. Mem. Cogn.* **36:** 1355–1368.
31. Ricker, T.J. & N. Cowan. 2014. Differences between presentation methods in working memory procedures: a matter of working memory consolidation. *J. Exp. Psychol. Learn. Mem. Cogn.* **40:** 417–428.
32. Ricker, T.J., L.R. Spiegel & N. Cowan. 2014. Time-based loss in visual short-term memory is from trace decay, not temporal distinctiveness. *J. Exp. Psychol. Learn. Mem. Cogn.* **40:** 1510–1523.
33. Souza, A.S. & K. Oberauer. 2015. Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychon. Bull. Rev.* **22:** 156–162.
34. McKeown, D. & T. Mercer. 2012. Short-term forgetting without interference. *J. Exp. Psychol. Learn. Mem. Cogn.* **38:** 1057–1068.
35. Mercer, T. & D. McKeown. 2014. Decay uncovered in nonverbal short-term memory. *Psychon. Bull. Rev.* **21:** 128–135.
36. Jolicoeur, P. & R. Dell'Acqua. 1998. The demonstration of short-term consolidation. *Cogn. Psychol.* **36:** 138–202.
37. Nieuwenstein, M. & B. Wyble. 2014. Beyond a mask and against the bottleneck: retroactive dual-task interference during working memory consolidation of a masked visual target. *J. Exp. Psychol. Gen.* **143:** 1409–1427.
38. Cowan, N., J.N. Rouder, C.L. Blume & J.S. Saults. 2012. Models of verbal working memory capacity: what does it take to make them work? *Psychol. Rev.* **119:** 480–499.
39. Burrows, D. & R. Okada. 1975. Memory retrieval from long and short lists. *Science* **188:** 1031–1033.
40. Wolfe, J.M. 2012. Saved by a log: how do humans perform hybrid visual and memory search? *Psychol. Sci.* **23:** 698–703.
41. Endress, A.D. & M.C. Potter. 2014. Large capacity temporary visual memory. *J. Exp. Psychol. Gen.* **143:** 548–566.
42. Cowan, N. 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* **24:** 87–185.

43. Luck, S.J. & E.K. Vogel. 1997. The capacity of visual working memory for features and conjunctions. *Nature* **390:** 279–281.
44. Adam, K.C.S., E.K. Vogel & E. Awh. 2017. Clear evidence for item limits in visual working memory. *Cogn. Psychol.* **97:** 79–97.
45. Rhodes, S., N. Cowan, K.O. Hardman & R.H. Logie. Informed guessing in change detection. *J. Exp. Psychol. Learn. Mem. Cogn.* https://doi.org/10.1037/xlm0000495.
46. Bayliss, D.M., J. Bogdanovsa & C. Jarrold. 2015. Consolidating working memory: distinguishing the effects of consolidation, rehearsal and attentional refreshing in a working memory span task. *J. Mem. Lang.* **81:** 34–50.
47. De Schrijver, S. & P. Barrouillet. 2017. Consolidation and restoration of memory traces in working memory. *Psychon. Bull. Rev.* **24:** 1651–1657.
48. Guttentag, R.E. 1984. The mental effort requirement of cumulative rehearsal: a developmental study. *J. Exp. Child Psychol.* **37:** 92–106.
49. Cowan, N. 1995. *Attention and Memory: An Integrated Framework*. New York: Oxford University Press.
50. Davelaar, E.J., Y. Goshen Gottstein, A. Ashkenazi, *et al.* 2005. The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychol. Rev.* **112:** 3–42.
51. Ericsson, K.A. & W. Kintsch. 1995. Long-term working memory. *Psychol. Rev.* **102:** 211–245.
52. Chen, Z. & N. Cowan. 2009. How verbal memory loads consume attention. *Mem. Cognit.* **37:** 829–836.
53. Maehara, Y. & S. Saito. 2007. The relationship between processing and storage in working memory span: not two sides of the same coin. *J. Mem. Lang.* **56:** 212–228.
54. Saito, S. & A. Miyake. 2004. On the nature of forgetting and the processing storage relationship in reading span performance. *J. Mem. Lang.* **50:** 425–443.
55. Vergauwe, E., V. Camos & P. Barrouillet. 2014. The impact of storage on processing: how is information maintained in working memory? *J. Exp. Psychol. Learn. Mem. Cogn.* **40:** 1072–1095.
56. Henson, R.N.A. 1999. Positional information in short term memory: relative or absolute? *Mem. Cognit.* **27:** 915–927.
57. Cowan, N., J.S. Saults & C.L. Blume. 2014. Central and peripheral components of working memory storage. *J. Exp. Psychol. Gen.* **143:** 1806–1836.
58. Saults, J.S. & N. Cowan. 2007. A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *J. Exp. Psychol. Gen.* **136:** 663–684.
59. Cowan, N., Y. Li, B. Glass & J.S. Saults. Development of the ability to combine visual and acoustic information in working memory. *Dev. Sci.* https://doi.org/10.1111/desc.12635.
60. Cowan, N. 2016. Working memory maturation: can we get at the essence of cognitive growth? *Perspect. Psychol. Sci.* **11:** 239–264.
61. Pascual Leone, J. & J. Smith. 1969. The encoding and decoding of symbols by children: a new experimental paradigm and a neo Piagetian model. *J. Exp. Child Psychol.* **8:** 328–355.
62. Cowan, N., T.J. Ricker, K.M. Clark, *et al.* 2015. Knowledge cannot explain the developmental growth of working memory capacity. *Dev. Sci.* **18:** 132–145.
63. Halford, G.S., N. Cowan & G. Andrews. 2007. Separating cognitive capacity from knowledge: a new hypothesis. *Trends Cogn. Sci.* **11:** 236–242.
64. Fandakova, Y., D. Selmeczy, S. Leckey, *et al.* 2017. Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proc. Natl. Acad. Sci. USA* **114:** 7582–7587.
65. Logie, R.H., J.R. Brockmole & A.R. Vandenbroucke. 2009. Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Vis. Cogn.* **17:** 160–179.
66. Oberauer, K., T. Jones & S. Lewandowsky. 2015. The Hebb repetition effect in simple and complex memory span. *Mem. Cognit.* **43:** 852–865.
67. Cowan, N., T.N. Towse, Z. Hamilton, *et al.* 2003. Children's working-memory processes: a response-timing analysis. *J. Exp. Psychol. Gen.* **132:** 113–132.
68. McCabe, D.P. 2008. The role of covert retrieval in working memory span tasks: evidence from delayed recall tests. *J. Mem. Lang.* **58:** 480–494.
69. Camos, V. & S. Portrat. 2015. The impact of cognitive load on delayed recall. *Psychon. Bull. Rev.* **22:** 1029–1034.
70. Lewis-Peacock, J.A., A.T. Drysdale, K. Oberauer & B.R. Postle. 2012. Neural evidence for a distinction between short-term memory and the focus of attention. *J. Cogn. Neurosci.* **24:** 61–79.
71. Rose, N.S., J.J. LaRocque, A.C. Riggall, *et al.* 2016. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354:** 1136–1139.
72. Reinhart, R.M. & G.F. Woodman. 2014. High stakes trigger the use of multiple memories to enhance the control of attention. *Cereb. Cortex* **24:** 2022–2035.
73. Wallis, G., M. Stokes, H. Cousijn, *et al.* 2015. Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory. *J. Cogn. Neurosci.* **27:** 2019–2034.
74. Lewandowsky, S. & K. Oberauer. 2015. Rehearsal in serial recall: an unworkable solution to the nonexistent problem of decay. *Psychol. Rev.* **122:** 674–699.
75. Souza, A.S. & K. Oberauer. 2017. The contributions of visual and central attention to visual working memory. *Atten. Percept. Psychophys.* **79:** 1897–1916.
76. Awh, E., E.K. Vogel & S.H. Oh. 2006. Interactions between attention and working memory. *Neuroscience* **139:** 201–208.
77. Foster, J.J., D.W. Sutterer, J.T. Serences, *et al.* 2017. Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychol. Sci.* **28:** 929–941.
78. Oberauer, K. & H.Y. Lin. 2017. An interference model of visual working memory. *Psychol. Rev.* **124:** 21–59.
79. Talmi, D., C.L. Grady, Y. Goshen Gottstein & M. Moscovitch. 2005. Neuroimaging the serial position curve: a test of single-store versus dual-store models. *Psychol. Sci.* **16:** 716–723.
80. Broadbent, D.E. 1971. *Decision and Stress*. London: Academic Press.