Working-memory consolidation

Ricker, Timothy J.; Nieuwenstein, Mark R.; Bayliss, Donna M.; Barrouillet, Pierre

Published in:
Annals of the New York Academy of Scieces

DOI:
10.1111/nyas.13633

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

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

Working memory consolidation: insights from studies on attention and working memory

Timothy J. Ricker,[1,2] Mark R. Nieuwenstein,[3] Donna M. Bayliss,[4] and Pierre Barrouillet[5]

[1]College of Staten Island, City University of New York, Staten Island, New York. [2]The Graduate Center, City University of New York, New York, New York. [3]University of Groningen, Groningen, the Netherlands. [4]University of Western Australia, Western Australia, Australia. [5]Université de Genève, Geneva, Switzerland

Address for correspondence: Timothy J. Ricker, College of Staten Island, City University of New York, 2800 Victory Blvd, Staten Island, NY 10314. Timothy.Ricker@csi.cuny.edu

Working memory, the system that maintains a limited set of representations for immediate use in cognition, is a central part of human cognition. Three processes have recently been proposed to govern information storage in working memory: consolidation, refreshing, and removal. Here, we discuss in detail the theoretical construct of working memory consolidation, a process critical to the creation of a stable working memory representation. We present a brief overview of the research that indicated the need for a construct such as working memory consolidation and the subsequent research that has helped to define the parameters of the construct. We then move on to explicitly state the points of agreement as to what processes are involved in working memory consolidation.

Keywords: consolidation; attentional blink; working memory; short-term memory; attention

Introduction

Attention plays an important role in working memory performance by creating and maintaining a set of representations accessible for immediate use in cognition. The mechanism or set of mechanisms through which attention affects working memory is less obvious. In the following pages, we explain the role of attention in creating a stable working memory state. This process is referred to as working memory consolidation. We start by giving a brief overview of research that demonstrates the existence of a working memory consolidation process. We then lay out in detail the properties of the consolidation process that are broadly agreed upon across research groups and the properties that are still under debate.

The concept of working memory consolidation: a brief historical overview

In contemplating the workings of mind, philosophers in the 16th–19th centuries often commented on the intimate link between attention and memory.[1] Indeed, it takes only a brief moment of introspection to notice that the information we perceive is available only fleetingly unless we pay attention to it. However, understanding exactly how attention contributes to memory is not an unequivocal matter, as many ideas have been proposed with regard to the role of attention in memory formation. For instance, William James[2] proposed that attention enables the selection of perceived information for representation in a more durable form of secondary memory, whereas Waugh and Norman[3] echoed Broadbent's[4] early filter theory in proposing that attentional selection occurs prior to perception, and Shiffrin[5] proposed that perceived information is automatically "dumped" into short-term memory, after which attention can selectively accentuate relevant information by enabling its rehearsal. While these early views thus differed in terms of their proposal about the locus of selection (i.e., before perception, during perception, or after short-term memory storage), they converged in assuming that rehearsal constitutes the process by which attended information can be remembered. Thus, of relevance to the current purposes, the early views on attention and memory typically did not assume a need for a

doi: 10.1111/nyas.13633

Ann. N.Y. Acad. Sci. 1424 (2018) 8–18

separate process of short-term or working memory consolidation. Instead, such consolidation was held to be synonymous with attention and rehearsal (see also, Duncan, who used the phrase "consolidation in short-term memory" to denote Shiffrin's[5] proposal for selective accentuation by rehearsal in short-term memory).

The notion that there exists a separate process of working memory consolidation finds its origins in studies on the attentional blink (AB). An important antecedent to these studies can be found in the work by Duncan,[6] who was one of the first to investigate the limitations in processing a masked visual target for a delayed report. Specifically, Duncan presented participants with a brief and masked display that contained letter distractors and either one or two digit targets, and compared target detection. The results showed that participants often missed a target when the display included two targets, whereas they performed relatively well when the display included only a single target. On the basis of these results, Duncan proposed that target selection occurred after an initial categorization of the items in the display (thus explaining why distractors did not influence target detection as much as the presence of another target), and he proposed that the limitation in detecting the second of two targets could be explained in terms of a limited-capacity processing stage that was needed for conscious report of the targets. In proposing this account, Duncan thus assumed a distinction between selection and a subsequent, capacity-limited stage of processing that led to a consciously accessible, reportable memory trace, and he assumed no role for rehearsal in mediating performance.

Following Duncan's[6] seminal discovery, Broadbent and Broadbent[7] set out to investigate the time course of the capacity-limited process that was identified by Duncan. To this end, Broadbent and Broadbent conducted a series of experiments in which participants were asked to report the identities of two visual targets that appeared in a rapid serial visual presentation (RSVP) of distractors. The results showed that identification of the first target (T1) was associated with a U-shaped performance curve for identifying the second target (T2) across target onset asynchronies of 80–720 ms, suggesting that the processing required for report of T1 interfered with the occurrence of such processing for T2 for more than half a second. In accounting for this effect, which was later dubbed the AB,[8] Broadbent[a] and Broadbent proposed that it reflected a limitation in the process of identification, the idea being that the identification of T1 took approximately 720 ms, thereby causing interference with the identification of T2 across this period of time.

In a subsequent study, Chun and Potter[9] were the first to hint at the possibility that the AB might reflect the time course of a consolidation process whereby the identity of a target comes to be represented in a durable, consciously accessible form in short-term or working memory. Specifically, Chun and Potter interpreted the AB in the context of Potter's earlier work, which had shown that participants can easily detect the presence of a target picture in an RSVP sequence of other pictures, even when they were only provided with a general description of this picture beforehand (e.g., a picnic scene).[10–13] Based on this result, Potter proposed that the gist or meaning of a visual stimulus is activated very rapidly, thus enabling target detection even when items are shown at a very rapid pace. In line with this view, Chun and Potter proposed that items in RSVP are rapidly identified in a first stage of processing, with the resulting information being highly vulnerable to the effects of masking. Accordingly, the ability to report a target's identity at the end

[a]To explain why report of T2 was often spared from the AB when T2 appeared directly after T1, Chun and Potter further proposed that the detection of a potential target in Stage 1 triggers a transient but sluggish attentional response by which a target represented in Stage 1 could be selected for capacity-limited processing in Stage 2. Since this attentional response was assumed to be somewhat sluggish, it would allow T2 to be selected for Stage 2 in the slipstream of T1, with the result being that both targets could be consolidated together, at the cost of losing information about the order in which the two targets appeared (i.e., when T1 and T2 appear in direct succession, they are often both identified correctly, but reported in an incorrect order). Finally, Chun and Potter also offered an explanation for the finding—first reported by Raymond et al.[8]—that omitting the item trailing T1 (i.e., the T1 mask) resulted in strong attenuation of the AB. Specifically, Chun and Potter reasoned that the presence of such a mask would increase the difficulty of T1 identification, thereby increasing the time it would take to consolidate T1 in working memory, and thus increasing the magnitude and duration of the AB.

of the trial would require a second stage of processing that was assumed to involve a capacity-limited consolidation process whereby the target's identity can be consolidated into a more durable representation in short-term or working memory. Thus, the time course of the AB was interpreted to reflect the time course of T1 consolidation, the idea being that the consolidation of T2 could only begin after consolidation of T1 was completed.

Chun and Potter's[9] proposal for a two-stage account found support in the results of Vogel and colleagues[14,15] who used the logic of event-related potentials (ERPs) to examine how the AB affects the processing of T2. Specifically, in an elegant series of experiments, Vogel *et al.* manipulated the nature of T2 in such a way that it would be expected to elicit distinctive ERPs, such as the P1 component reflecting early visual processing, the N400 component reflecting access to meaning, and the P3 component. The examination of the P3 was of particular interest to test the two-stage account because previous work had shown that the magnitude of this component is associated with whether or not a to-be-encoded stimulus can later be remembered, thus suggesting a link between the P3 and the successful consolidation of information in memory[16] (for recent reviews and discussion of the link between the P3 and WMC, see Refs. 17 and 18). Indeed, Vogel *et al.* found that the AB was not associated with modulation of the P1 or N400, whereas it was associated with suppression of the P3, thus suggesting that visual processing and the extraction of T2's meaning (Stage 1) were not affected by the AB, whereas an ensuing stage of consolidating information into working memory (i.e., Stage 2) indeed appeared to be blocked for T2 during the AB. In further support of this view, Vogel and Luck[19] later showed that if T2 is not followed by any masking distractors, it will elicit a delayed P3 component when shown during the AB, thus corroborating the idea that during the AB, working memory consolidation of T2 is postponed until consolidation of T1 has been completed.[20]

Working memory consolidation within working memory paradigms

The consolidation effects found in the above-described studies on the AB and dual-task interference have also been shown to influence performance in a variety of other tasks commonly used in studies on working memory. Seminal work by Jolicoeur and Dell'Acqua[21] systematically examined dual-task interference between a visual working memory task and a 2-alternative forced choice (2-AFC) task. This study provided further support for Chun and Potter's[9] proposal of a consolidation-bottleneck and expanded upon it by demonstrating that the bottleneck should be considered a central attentional bottleneck. Jolicoeur and Dell'Acqua asked participants to memorize a display of either one or three letters or symbols and to perform a speeded 2-AFC task for a trailing auditory target. Across seven experiments, they demonstrated that response times for the 2-AFC task showed a psychological refractory period (PRP) effect, meaning that when the auditory target was presented with a shorter stimulus onset asynchrony (SOA) to the memory set, reaction time on the 2-AFC task was slower. Furthermore, the results also showed that this PRP effect was stronger when more items had to be remembered and it did not occur when participants were instructed to ignore the memory array. Taken together, these findings supported Chun and Potter's bottleneck account in demonstrating that the requirement to consolidate the first target produced interference with a shortly following target. They also expand upon Chun and Potter's ideas by demonstrating that consolidation causes a postponement of response selection for a trailing target. Accordingly, Jolicoeur and Dell'Acqua concluded that working memory consolidation imposes a central attentional bottleneck, meaning that the bottleneck not only applies to the process of consolidation itself, but also to other mental operations that rely on central processing mechanisms.

Following the work of Jolicoeur and Dell'Acqua,[21] several investigations confirmed that engaging in working memory consolidation delays or slows the execution of other attention demanding tasks, while also expanding our knowledge about the consolidation process. Stevanovski and Jolicoeur[22] used the same basic dual-task paradigm as Jolicoeur and Dell'Acqua, a visual-array change-detection task, in which several visual items must be remembered for a later recognition test, with a 2-AFC task presented at variable delays after memory array presentation. Critically, in experiment 3, Stevanovski and Jolicoeur also prevented articulatory rehearsal and phonological recoding with the use of articulatory suppression. This study replicated previous findings of an effect of secondary task SOA on secondary

task reaction time (i.e., the PRP effect) even when the engagement of verbal processing was not possible. This finding was a critical demonstration that consolidation is not synonymous rehearsal, in contrast with the assumptions of early models of short-term memory.[5,6] On the other hand, when verbal rehearsal was prevented, Stevanovski and Jolicoeur did not replicate the finding that increasing the memory set size increased the PRP effect. Instead, they propose that Jolicouer and Dell'Acqua found an increase in the PRP effect with increasing memory set size due to the contribution of verbal recoding. When more items were presented, more verbal recoding was required and secondary task performance was increasingly delayed. Together these results suggest consolidation can proceed in parallel for multiple items presented simultaneously.

Although these working memory studies show that the processing involved in memory creation delays other attention demanding tasks, they do not show that this delay is due to the memory creation process. In order to confirm that the delayed processing of secondary tasks is due to the consolidation of the memory, one would want to interrupt the consolidation process and observe impaired memory performance as a result. Nieuwenstein and Wyble[23] did this using a visual-array memory task with a 2-AFC secondary task similar to those used in previous studies, but with a higher degree of difficulty in both tasks. They observed deficits in memory accuracy and secondary task reaction time that increased as the SOA between the tasks decreased. A decrease in memory performance following a decrease in secondary task SOA was also observed by Bayliss *et al.*[24] and De Schrijver and Barrouillet,[25] confirming an impairment in memory performance when consolidation is disrupted. These two studies differed from previous work in that they used a complex span task as the primary memory task, in which memory items are presented sequentially with secondary distraction tasks interspersed between the presentations of each memory item. Thus, the studies of Bayliss *et al.* and De Schrijver and Barrouillet provide converging evidence for the existence of a working memory consolidation process beyond visual array and RSVP paradigms.

In defining the role of consolidation in memory formation, an important question is whether effects ascribed to consolidation can be distinguished from effects that might be caused by operations of maintenance that occur after an item has been consolidated (i.e., rehearsal and other means by which items in working memory can be refreshed). A number of findings from both AB and working memory span tasks suggest that consolidation and maintenance indeed reflect distinct processes. For example, Bayliss *et al.*[24] demonstrated that providing an opportunity for consolidation in a complex span task, by inserting an unfilled delay interval immediately after each memory item presentation, led to better memory performance than when there was no delay between each memory item and a secondary distraction task. This effect did not interact with manipulations of another attentional-based process, attentional refreshing, the process of maintaining memory items by cycling them in and out of the focus of attention,[26–28] suggesting that consolidation and attentional refreshing are separable processes. Bayliss *et al.* also showed that the improved memory performance observed when an opportunity for consolidation was provided was not due to phonological recoding or articulatory rehearsal, as the same benefit was evident even when participants engaged in articulatory suppression. De Schrijver and Barrouillet[25] replicated these effects when using articulatory suppression (though a different pattern was evident when articulatory suppression was not used), but argued instead that the processes of consolidation and attentional refreshing can be substituted for one another, based on the finding that performance was similar in conditions that equated overall time available for attention-demanding processes. Thus, the differentiation of these two processes is still a matter of some debate.

At least three studies have demonstrated that consolidation has downstream effects on later maintenance. Ricker and Cowan[29] demonstrated that limiting consolidation time results in greater rates of time-based forgetting during a subsequent retention interval. Similarly, De Schrijver and Barrouillet[25] show that decreasing the time available for consolidation increases the amount of dual-task interference caused by a persistent secondary task that occurs at a constant pace throughout the memory trial. In the same way, Barrouillet *et al.*[30] observed that allowing more time for consolidating memory items made them more resistant to interference created by an attention-demanding secondary task. Together, these studies appear to demonstrate that working memory consolidation reduces the fragility

of a memory trace, increasing its resistance to later forgetting.

Recent findings by Ricker and Hardman[31] suggest the working memory and AB literatures have much in common. Ricker and Hardman employed a visual working memory task in which participants were presented with a series of memory items varying in their orientation and were required to reproduce the angle of orientation of all items at the end of the presentation sequence. Unsurprisingly, when the time for consolidation between each item varied from 200 to 1200 ms, shorter SOAs between memory items resulted in poorer performance. When manipulating only a single SOA on a given trial, Ricker and Hardman were able to demonstrate that the item preceding the SOA was unaffected by its length. The observed decrease in memory performance was present only for the item after the consolidation period. Through mathematical modeling of response errors, Ricker and Hardman were able to show that increased error rates following shorter SOAs were driven by a decrease in the likelihood of having the item in mind during recall and not by a change in the quality of the memory maintained. This constellation of findings presents strong evidence that an AB is driving memory deficits in working memory procedures with short consolidation times.

Our present understanding of working memory consolidation

In this section, we outline our present understanding of working memory consolidation. By addressing several specific questions, we hope to clarify the process of consolidation and differentiate it from maintenance mechanisms. This should allow for clearer testing, verification, and falsification of the proposed mechanisms. In some cases, there is active debate over how consolidation should be implemented within working memory models. We describe where there are points of consensus and where there is disagreement in our varying viewpoints.

Definition

Working memory consolidation refers to the transformation of transient sensory input into a stable memory representation that can be manipulated and recalled after a delay.[9,21,24,29] This process should also convert fleeting internally generated representations into stable working memory traces, but this has not yet been confirmed experimentally. The process of consolidation can be distinguished from more basic processes involved in the initial perception and identification of a stimulus. This is demonstrated by the findings that providing an opportunity for consolidation results in better memory performance even when ongoing perceptual processes are prevented by the presentation of a mask,[23,25,29,31] that effects that appear to reflect a consolidation process are independent of differences in the perceptual characteristics of stimuli,[32] and that identification of a stimulus does not always result in its consolidation into a durable form.[33,34]

A characteristic of consolidation that is important for distinguishing it from maintenance processes in working memory is that it refers to the *initial* attentional processing directed at a new sensory trace to establish that trace in working memory. In this sense, consolidation may reflect the entry of activated long-term memory representations into the focus of attention,[29] the binding of perceptual and semantic features to a token representation,[35,36] or the initial binding of an item to its context.[37] Once attention is directed elsewhere, any further attentional processing directed toward the memory trace would not act to consolidate the memory representation, but instead, would act to refresh or boost the activation of the established memory trace (the nature of this process is the subject of other papers within this special issue). Although there is some debate as to whether the attentional processing involved in consolidation and refreshing is "substitutable,"[25] the finding that free time presented immediately following the presentation of a memory item is more beneficial for memory performance than the same amount of free time presented later in the trial, after an interleaved processing activity,[24] suggests that the period immediately following the presentation of memory items is unique in some way.

It is also important to differentiate consolidation from masking effects. Masking a memory stimulus immediately after its presentation has been proposed to end the consolidation process.[38,39] It is clear that masking does interfere with memory performance in a time-dependent manner, with shorter SOAs between memory item and mask presentation leading to lower performance levels.[39–42] The fast time course over which masking deficits

dissipate and the gradual reduction in the size of this deficit over time do seem to imply that they are related to working memory consolidation, but previous work has shown that masking does not prevent working memory consolidation.[21,23,25,29,31] Instead, masking likely interacts with consolidation either by lengthening the time needed for working memory consolidation[35,36,43] or by interfering with the unconsolidated representation.[21,44] Together the data seem to indicate that a consolidated representation should show smaller masking effects and that those representations that survive masking are not necessarily consolidated.

One plausible characterization of the consolidation process is that it involves a strengthening or stabilizing of the memory trace through modifications of synaptic activation or connectivity. This could be likened to the process of synaptic consolidation, in which activated synapses are tagged, resulting in cellular changes that act to modify and strengthen the synaptic configuration (see Dudai for a review and Poo *et al.* for a more recent discussion of plausible neural mechanisms underlying memory formation).[45,46] Much progress has been made in understanding the neurobiological processes underlying the formation of long-term memory. Whether similar neurobiological processes underlie the shorter and more transient memory traces involved in working memory remains to be seen.

To be clear, working memory consolidation is not proposed to work through the same mechanisms as long-term consolidation. Although long-term memories undergo consolidation, the long-term process is not necessarily related to working memory consolidation. The relationship between these two processes is likely to be purely linguistic, although it is not impossible that they are related on some level. For example, consolidation into working memory may be a necessary precursor to long-term memory formation.[44,47]

Basic mechanism and representation

In the general sense, researchers working on questions related to consolidation agree that it is the process of transitioning a sensory representation into a working memory representation, that it requires attention, and that it is relatively rapid in its completion. Beyond this though, agreement on a specific mechanism or set of mechanisms to account for the body of research on working memory consolidation has yet to coalesce. Researchers have produced several potential theoretical mechanisms differentiated largely by how attention is restricted during consolidation.

The early conceptualization of consolidation within working memory proposed by Jolicoeur and Dell'Acqua[21] relied on a strict attentional bottleneck model in which early sensory and perceptual processes proceeded in parallel, but the process of consolidation, which required capacity-limited central mechanisms, proceeded serially, effectively delaying the central processing required for any subsequent task until consolidation had finished. Thus, an attentional bottleneck model accounts for findings from this and other studies demonstrating delayed response times to a subsequent processing task presented at short intervals following the presentation of a to-be-remembered item.[21,22,24,48]

The attentional bottleneck explanation of consolidation fits nicely within the time-based resource-sharing (TBRS) account of working memory that describes dual-task performance more generally. The process of consolidation, as with any other operation on representations within the TBRS model, is assumed to be sequential in nature, operating on one representation at a time. Empirical evidence for this assumption comes from a study by Vergauwe *et al.*[49] in which participants had to maintain series of letters of different lengths for a 12-second delay filled by a parity judgment task on digits appearing successively on screen. Vergauwe *et al.* observed that processing time was longer for the first compared with the subsequent digits, a postponement attributed to the consolidation process of the letters to be recalled before entering the parity task. Interestingly, in line with the hypothesis of sequentiality, this postponement increased with the number of memory items (see Stevanovski and Jolicoeur, discussed previously).

In dual-task paradigms that focus on response times to a secondary task, memory performance is typically high and unaffected by changes in SOA[21], particularly for low memory loads. However, in the working memory span paradigms used by Bayliss *et al.*[24] and De Schrijver and Barrouillet,[25] memory performance is affected by manipulation of the delay interval between the presentation of a memory item and a subsequent processing task. Thus, whereas evidence based on response times to a secondary

processing task suggests that ongoing consolidation delays subsequent task processing, evidence based on memory performance assumes that consolidation can be interrupted by the presentation of a secondary processing task.[23] The fact that we see effects of secondary task processing (Task 2) on memory performance (Task 1) that are affected by the timing of the onset of the distractor processing activity suggests that these two components of the task are not processed in a strict serial "all-or-none" manner where Task 1 must be completed in its entirety before processing of Task 2 can begin.

Models that allow for parallel central processing can accommodate both sets of findings (e.g., central capacity sharing models).[50,51] These models also assume that central processing resources are limited, but in contrast to strict attentional bottleneck models, the processing required for Task 1 and Task 2 can proceed in parallel. Thus, any central processing required by each task must share the limited processing capacity available. The implication from this is that if an item is still being consolidated when a subsequent processing task is presented, then the available capacity for central processing would be shared at that point. This is likely to lead to less than optimal consolidation of the memory item, and will also lead to delayed response times to the subsequent processing task, as it will take longer to complete the same amount of processing given that the available resources are now being shared. A capacity sharing model can account for studies demonstrating effects consistent with consolidation in both memory performance and secondary task processing response times,[23] or in both memory performance and secondary task accuracy,[24] within the same paradigm. Importantly, the proportion of the available capacity allocated to each task can vary, and may be influenced by task demands,[52] which can also account for variation in the effects attributable to consolidation under some conditions. That is, the allocation of attentional resources for consolidation appears to be under strategic control.

This raises the question of whether the limited resources are truly "shared" or whether individuals are strategically switching their attention from consolidation, effectively interrupting the consolidation process, to engage with the subsequent processing task. A task-switching account would be able to account for the more general finding in the literature that dual-task costs are persistent and difficult to eradicate,[53] as well as the finding that effects are sometimes seen in more than one task component. If we assume that individuals do not always switch consistently, sometimes choosing to interrupt consolidation to engage with processing, sometimes opting to finish consolidation before starting processing, then the fact that most studies tend to average performance across trials could lead to the pattern of effects described above. However, arguments against a task-switching model have also been made.[23] Thus, whether consolidation is best accounted for by a capacity sharing model or a task-switching model is yet to be determined.

Which type of attention is engaged?

The initiation of consolidation appears to be under conscious control and is assumed to require central attention (often called, executive, general, or domain-general attention). Analogously, within the TBRS model, consolidation would be considered to occur within the executive loop. Central attention is theorized to be the same resource as is used for inhibitory control,[54,55] response selection,[56,57] and attention-based refreshing[26–28] as well as a number of other cognitive processes. Because it requires central attention, working memory consolidation can be blocked by any other cognitive process that prevents the shifting of attention toward the consolidation target.[9,31]

Perhaps the strongest evidence for the role of conscious control of central attention in the consolidation process comes from the paradigms of Jolicoeur and colleagues,[21,22] in which they employ an encode condition that requires consolidation, but also an ignore condition, which involves an identical stimulus presentation but participants are not required to encode the memory item. In the encode condition, there are delays to RTs as a function of the delay interval between the memory item and the secondary processing task that exceed any delay related to the ignore condition, suggesting that consolidation does not happen automatically. Further support comes from Chen and Wyble,[33,34] who showed that when asked to report the location of a letter in an array of digits, participants were not be able to report the identity of that letter if they did not know they would be asked about it. This suggests selective consolidation of the to-be-remembered feature (i.e., location), without consolidation of the feature that was used to locate and attend the target (i.e., its

identity). They replicated this finding across a number of feature combinations, demonstrating that this was a general principle not bound to specific contexts. An important qualification to the conclusion that the initiation of consolidation is under conscious control stems from findings that show that stimuli that capture attention because of their intrinsic emotional salience appear to be consolidated into memory even when there is no task-requirement to do so.[58]

Although there is consensus surrounding the role of attention in the initiation of consolidation, there is debate over whether attention can be intentionally disengaged from the process on a moment-to-moment basis once it has begun. There is some evidence that attention can be switched to new stimuli during ongoing consolidation,[23,36,59–61] but other evidence suggests that during consolidation, attention cannot be switched to a new stimulus until consolidation is complete.[9,21,31]

Time course

The time course of consolidation estimated across the various studies has tended to vary. This may be influenced by the particular paradigm used to investigate consolidation. Studies on the AB suggest that consolidation of a single letter or digit may be completed within approximately 500 ms.[9,15,35,62] On the other hand, paradigms that have examined the point at which memory performance is no longer influenced by a secondary processing task, indicating that any consolidation is likely to be finished, have consistently shown slightly longer estimates. For example, Nieuwenstein and Wyble[23] showed that memory performance for a single novel character was unaffected by a secondary task after 1000 ms suggesting that consolidation finished somewhere between 500 and 1000 ms. Similarly, Ricker and Hardman[31] show that memory for a novel angle is adversely affected by the presentation of the next memory stimulus at 400 ms, but unaffected at 800 ms.

Using a complex span paradigm seems to produce slightly longer estimates again, with Bayliss *et al.*[24] demonstrating that memory performance was affected by a secondary processing task presented 700 ms after the onset of the memory stimulus (experiment 1). De Schrijver and Barrouillet[25] also had participants perform a complex span task and manipulated consolidation times before the onset of the first distracter item, but with each letter followed by a mask. The results revealed that even long consolidation times (up to 2000 ms) still had a beneficial effect on recall performance, especially when the distracter task involved a high cognitive load, that is, when the secondary task had to be performed at a fast pace, reducing the time available for refreshing. Similar findings also come from Barrouillet *et al.*[30] who found that accuracy continued to improve with increases in consolidation time beyond 1500 ms. The slightly longer estimates from studies using complex span paradigms may, in part, reflect the increased demands of these tasks. Greater overall difficulty could produce an overall nonspecific slowing affecting all cognitive processing. This type of slowing was evident in experiment 4 of Bayliss *et al.*, in which response times to a secondary task increased with memory load, but the rate of consolidation remained the same.

In contrast, studies that have examined the point at which RTs on a secondary task are no longer affected by a memory load, again indicating that any consolidation processes are likely to have finished, produce more variable estimates. In their seminal paper, Jolicoeur and Dell'Aqcua[21] suggested a time course for consolidation of around 500 ms. However, in the study by Stevanovski and Jolicoeur,[22] in which verbal rehearsal was controlled, an inflection point in the RT function was evident at around 500 ms for a single item, but RTs continued to decrease up to the maximum SOA of 1600 ms. An even longer estimate was provided by Bayliss *et al.*,[24] in which participants completed a serial recall task with a secondary tone discrimination task interspersed between each item at varying delay intervals. For the first item in the list (i.e., Serial position 1), RTs to the tone task showed an inflection point around 1000 ms from the onset of the memory stimulus, whereas RTs following every subsequent item continued to decrease up to the maximum SOA of 2000 ms. Again, this paradigm is more complex than those used by Jolicoeur, which may have contributed to the RT slowing. Taken together, these results suggest that a reasonable estimate of the time course of consolidation is likely to be between 500 and 1000 ms for a single item, but that the actual time point at which consolidation is finished within any given task may vary depending on the task demands.

The episodic simultaneous type-serial token model of consolidation

Perhaps one of the most detailed accounts of working memory consolidation can be found in the simultaneous type-serial token (STST) model of the AB,[62] and its successor, the episodic STST (eSTST) model.[35,36] In an attempt to explain several robust findings from studies on the AB, the STST and eSTST accounts provide a theory of what happens when a masked visual target needs to be consolidated into working memory for a delayed judgment or report. The model accomplishes this task by means of a neural-network architecture in which mechanisms of attentional selection and working memory consolidation are dissected and implemented by putative algorithms that have some neurobiological plausibility.

Specifically, the model assumes that working memory consolidation involves a binding process whereby a type (i.e., a representation of the to-be-remembered information about a target) is bound to a token (i.e., a working memory placeholder that can sustain its activation over time and that enables the retrieval of the associated type). The initiation of this binding process occurs when the activation level of a type reaches a threshold level of activation. The selection of targets for consolidation is implemented by assuming that the activation of a target type is more likely to reach the threshold for initiating consolidation because targets are assumed to trigger a transient attentional boost that enhances their activation. The completion of the ensuing binding process is assumed to take several hundred milliseconds even for a simple stimulus such as a single letter or a digit, and it is assumed that this binding process can occur for multiple items in parallel, as long as they were attended.

To accommodate the intricate patterns of interference and spared performance (i.e., the finding that T2 report is spared from the AB when it is precued or part of a continuous sequence of targets)[59–61,63,64] seen in studies of the AB, the model assumes that the selection of targets for consolidation depends on a competitive interaction between consolidation-driven inhibition and target-driven excitation of attentional enhancement. This competitive interaction entails that a rapid sequence of targets can sustain attentional enhancement despite the ongoing consolidation of earlier items, thereby enabling the selection of successive targets for parallel consolidation. In contrast, a brief temporal gap in a target sequence can cause an AB because the momentary absence of new target input will allow consolidation-driven inhibition to suppress attentional enhancement, initiating the AB. The model's assumptions are supported by the fact that it produces accurate predictions and simulations of both behavioral and electrophysiological data.

Although the eSTST model is detailed and can account for a variety of findings, it was not developed with the goal of describing consolidation within working memory paradigms. Thus, it is still unknown whether it generalizes to explain consolidation across a variety of working memory paradigms, which should be a priority of future research. If its assumptions hold and provide good fits to the existent data outside of the RSVP context, it could provide a concise description of a variety of findings in both RSVP and working memory.

Concluding remarks

The last 20 years have produced a steady increase in our knowledge about working memory consolidation. Far from simply being dumped into working memory, stimuli must instead undergo a rapid, yet attention demanding, consolidation process to enter working memory. We have learned that initiation of consolidation is under conscious control and requires the use of central attention. Although at first glance stimulus masking effects seem theoretically similar to consolidation effects, masking does not end working memory consolidation. Some aspects are not universally agreed upon, but several compelling evidence-based hypotheses exist. For example, it is not yet certain whether attention is limited by a bottleneck during consolidation, whether attention is a limited capacity but divisible resource during consolidation, or if the eSTST model can be extended to govern the deployment of attention during consolidation beyond conditions similar to the RSVP paradigm. There is much yet to learn about working memory consolidation, but answers continue to accumulate at a steady pace.

Acknowledgments

This manuscript is the result of collaborative debate and discussion at the workshop, "The Crossroads of Attention in Working Memory: Consolidation, Refreshing, & Removal." The ideas related in this manuscript are a condensed version of our

conclusions. All authors took part in these discussions. All authors took part in the planning and writing of the manuscript. T.J. Ricker curated and penned the first and final drafts of the manuscript. All authors performed critical revisions of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Vu, K.-P. 2004. *Attention: Theory and Practice.* Thousand Oaks, CA: SAGE Publications, Inc.
2. James, W. 1890. *The Principles of Psychology.* Vol. 1. New York: Holt.
3. Waugh, N.C. & D.A. Norman. 1965. Primary memory. *Psychol. Rev.* **72:** 89–104.
4. Broadbent, D.E. 1958. *Perception and Communication.* New York, NY: Pergamon Press.
5. Shiffrin, R.M. 1975. The locus and role of attention in memory systems. In *Attention & Performance V*. P. Rabbit & S. Dornic, Eds.: 168–193. New York, NY: Academic Press.
6. Duncan, J. 1980. The demonstration of capacity limitation. *Cogn. Psychol.* **12:** 75–96.
7. Broadbent, D.E. & M.H. Broadbent. 1987. From detection to identification: response to multiple targets in rapid serial visual presentation. *Atten. Percept. Psychophys.* **42:** 105–113.
8. Raymond, J.E., K.L. Shapiro & K.M. Arnell. 1992. Temporary suppression of visual processing in an RSVP task: an attentional blink? *J. Exp. Psychol. Hum. Percept. Perform.* **18:** 849–860.
9. Chun, M.M. & M.C. Potter. 1995. A two-stage model for multiple target detection in rapid serial visual presentation. *J. Exp. Psychol. Hum. Percept. Perform.* **21:** 109–127.
10. Potter, M.C. 2012. Conceptual short term memory in perception and thought. *Front. Psychol.* **3:** 113.
11. Potter, M.C., B. Wyble, C.E. Hagmann, *et al.* 2014. Detecting meaning in RSVP at 13 ms per picture. *Atten. Percept. Psychophys.* **76:** 270–279.
12. Potter, M.C. 1975. Meaning in visual search. *Science* **187:** 965–966.
13. Potter, M.C. 1993. Very short-term conceptual memory. *Mem. Cogn.* **21:** 156–161.
14. Luck, S.J., E.K. Vogel & K.L. Shapiro. 1996. Word meanings can be accessed but not reported during the attentional blink. *Nature* **383:** 616–618.
15. Vogel, E.K., S.J. Luck & K.L. Shapiro. 1998. Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **24:** 1656–1674.
16. Donchin, E. 1981. Surprise! . . . surprise? *Psychophysiology* **18:** 493–513.
17. Craston, P., B. Wyble, S. Chennu, *et al.* 2009. The attentional blink reveals serial working memory encoding: evidence from virtual and human event-related potentials. *J. Cogn. Neurosci.* **21:** 550–566.
18. Polich, J. 2007. Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* **118:** 2128–2148.
19. Vogel, E.K. & S.J. Luck. 2002. Delayed working memory consolidation during the attentional blink. *Psychon. Bull. Rev.* **9:** 739–743.
20. Arnell, K.M. 2006. Visual, auditory, and cross-modality dual-task costs: electrophysiological evidence for an amodal bottleneck on working memory consolidation. *Atten. Percept. Psychophys.* **68:** 447–457.
21. Jolicœur, P. & R. Dell'Acqua. 1998. The demonstration of short-term consolidation. *Cogn. Psychol.* **36:** 138–202.
22. Stevanovski, B. & P. Jolicœur. 2007. Visual short-term memory: central capacity limitations in short-term consolidation. *Vis. Cogn.* **15:** 532–563.
23. Nieuwenstein, M.R. & 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.
24. Bayliss, D.M., J. Bogdanovs & 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.
25. De Schrijver, S. & P. Barrouillet. 2017. Consolidation and restoration of memory traces in working memory. *Psychon. Bull. Rev.* **24:** 1651–1657.
26. Barrouillet, P. & V. Camos. 2015. *Working Memory: Loss and Reconstruction.* New York, NY: Psychology Press.
27. Souza, A.S., L. Rerko & K. Oberauer. 2015. Refreshing memory traces: thinking of an item improves retrieval from visual working memory. *Ann. N.Y. Acad. Sci.* **1339:** 20–31.
28. Vergauwe, E. & N. Cowan. 2015. Attending to items in working memory: evidence that refreshing and memory search are closely related. *Psychon. Bull. Rev.* **22:** 1001–1006.
29. 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.
30. Barrouillet, P., G. Plancher, A. Guida, *et al.* 2013. Forgetting at short term: when do event-based interference and temporal factors have an effect? *Acta Psychol.* **142:** 155–167.
31. Ricker, T.J. & K.O. Hardman. 2017. The nature of short-term consolidation in visual working memory. *J. Exp. Psychol. Gen.* **146:**1551–1573.
32. Sun, H., H.D. Zimmer & X. Fu. 2011. The influence of expertise and of physical complexity on visual short-term memory consolidation. *Q. J. Exp. Psychol.* **64:** 707–729.
33. Chen, H. & B. Wyble. 2015. Amnesia for object attributes: failure to report attended information that had just reached conscious awareness. *Psychol. Sci.* **26:** 203–210.
34. Chen, H. & B. Wyble. 2016. Attribute amnesia reflects a lack of memory consolidation for attended information. *J. Exp. Psychol. Hum. Percept. Perform.* **42:** 225–234.
35. Wyble, B., H. Bowman & M. Nieuwenstein. 2009. The attentional blink provides episodic distinctiveness: sparing at a cost. *J. Exp. Psychol. Hum. Percept. Perform.* **35:** 787–807.
36. Wyble, B., M.C. Potter, H. Bowman, *et al.* 2011. Attentional episodes in visual perception. *J. Exp. Psychol. Gen.* **140:** 488–505.

37. Oberauer, K. & L. Hein. 2012. Attention to information in working memory. *Curr. Dir. Psychol. Sci.* **21:** 164–169.
38. Blalock, L.D. 2015. Stimulus familiarity improves consolidation of visual working memory representations. *Atten. Percept. Psychophys.* **77:** 1143–1158.
39. Vogel, E.K., G.F. Woodman & S.J. Luck. 2006. The time course of consolidation in visual working memory. *J. Exp. Psychol. Hum. Percept. Perform.* **32:** 1436–1451.
40. Blalock, L.D. 2013. Mask similarity impacts short-term consolidation in visual working memory. *Psychon. Bull. Rev.* **20:** 1290–1295.
41. Woodman, G.F. & E.K. Vogel. 2005. Fractionating working memory: consolidation and maintenance are independent processes. *Psychol. Sci.* **16:** 106–113.
42. Woodman, G.F. & E.K. Vogel. 2008. Selective storage and maintenance of an object's features in visual working memory. *Psychon. Bull. Rev.* **15:** 223–229.
43. Nieuwenstein, M.R., M.C. Potter & J. Theeuwes. 2009. Unmasking the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **35:** 159–169.
44. Ricker, T.J. 2015. The role of short-term consolidation in memory persistence. *AIMS Neurosci.* **2:** 259–279.
45. Dudai, Y. 2004. The neurobiology of consolidations, or, how stable is the engram? *Annu. Rev. Psychol.* **55:** 51–86.
46. Poo, M.-M., M. Pignatelli, T.J. Ryan, *et al.* 2016. What is memory? The present state of the engram. *BMC Biol.* **14:** 40.
47. Souza, A.S. & K. Oberauer. 2017. Time to process information in working memory improves episodic memory. *J. Mem. Lang.* **96:** 155–167.
48. Tombu, M.N., C.L. Asplund, P.E. Dux, *et al.* 2011. A unified attentional bottleneck in the human brain. *Proc. Natl. Acad. Sci. USA* **108:** 13426–13431.
49. 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.
50. Tombu, M. & P. Jolicœur. 2003. A central capacity sharing model of dual-task performance. *J. Exp. Psychol. Hum. Percept. Perform.* **29:** 3–18.
51. Lehle, C. & R. Hübner. 2009. Strategic capacity sharing between two tasks: evidence from tasks with the same and with different task sets. *Psychol. Res.* **73:** 707–726.
52. Fischer, R. & F. Plessow. 2015. Efficient multitasking: parallel versus serial processing of multiple tasks. *Front. Psychol.* **6:** 1366.
53. Ruthruff, E., J.C. Johnston & R.W. Remington. 2009. How strategic is the central bottleneck: can it be overcome by trying harder? *J. Exp. Psychol. Hum. Percept. Perform.* **35:** 1368–1384.
54. Kane, M.J., M.K. Bleckley, A.R. Conway, *et al.* 2001. A controlled-attention view of working-memory capacity. *J. Exp. Psychol. Gen.* **130:** 169–183.
55. Lavie, N. 2005. Distracted and confused? Selective attention under load. *Trends Cogn. Sci.* **9:** 75–82.
56. Corbetta, M. & G.L. Shulman. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3:** 201–215.
57. Jolicoeur, P. 1998. Modulation of the attentional blink by on-line response selection: evidence from speeded and unspeeded Task 1 decisions. *Mem. Cogn.* **26:** 1014–1032.
58. Most, S.B., M.M. Chun, D.M. Widders, *et al.* 2005. Attentional rubbernecking: cognitive control and personality in emotion-induced blindness. *Psychon. Bull. Rev.* **12:** 654–661.
59. Olivers, C.N., S. Van Der Stigchel & J. Hulleman. 2007. Spreading the sparing: against a limited-capacity account of the attentional blink. *Psychol. Res.* **71:** 126–139.
60. Nieuwenstein, M.R., M.M. Chun, R.H. van der Lubbe, *et al.* 2005. Delayed attentional engagement in the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **31:** 1463–1475.
61. Nieuwenstein, M.R. & M.C. Potter. 2006. Temporal limits of selection and memory encoding: a comparison of whole versus partial report in rapid serial visual presentation. *Psychol Sci.* **17:** 471–475.
62. Bowman, H. & B. Wyble. 2007. The simultaneous type, serial token model of temporal attention and working memory. *Psychol. Rev.* **114:** 38–70.
63. Nieuwenstein, M.R. 2006. Top-down controlled, delayed selection in the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **32:** 973–985.
64. Di Lollo, V., J.-I. Kawahara, S.S. Ghorashi, *et al.* 2005. The attentional blink: resource depletion or temporary loss of control? *Psychol. Res.* **69:** 191–200.