Author-final version of manuscript accepted for publication in the Journal of Experimental Psychology: Learning, Memory, and Cognition

Impact of memory load on processing diminishes rapidly during retention in a complex span paradigm

Tanya N. Joseph[1] & Candice C. Morey[1]

[1] Cardiff University

Author Note

We thank April Carne, Emma Chubb, Kelsey Jones, and Ralitsa Kostova for assistance with data collection and transcription, and Bill Macken and Dylan Jones for fruitful discussions. One of our research assistants (RK) was supported by a related grant from the British Academy (SRG18R1\180622) when she assisted with this project. TNJ and CCM are mutually responsible for designing the experiments, analyzing the data, and writing the manuscript. TNJ programmed the tasks with assistance from CCM and collected most of the data, and supervised data collection carried out by some of the research assistants. CCM drafted the analysis scripts, with assistance from TNJ, and supervised the remainder of the data collection. Data and materials are available at https://osf.io/dku3c/. Address correspondence to Dr. Candice C. Morey, School of Psychology, Cardiff University, 70 Park Place, CF103AT, Cardiff, United Kingdom. Email: MoreyC@cardiff.ac.uk.

Correspondence concerning this article should be addressed to Candice C. Morey, School of Psychology, Cardiff University, 70 Park Place, CF103AT, Cardiff, United Kingdom. E-mail: MoreyC@cardiff.ac.uk

Abstract

Previous work with complex memory span tasks, in which simple choice decisions are imposed between presentations of to-be-remembered items, shows that these secondary tasks reduce memory span. It is less clear how reconfiguring and maintaining various amounts of information affects decision speeds. We introduced preliminary "lead-in" decisions and post-encoding "lead-out" decisions to isolate potential influences of reconfiguration and maintenance on decision speeds. Compared with preliminary lead-in choice responses, the response associated with the first memory item slowed substantially. As the list accumulated, decision responses slowed even more. After presentation of the list was complete, decision responses sped rapidly: within a few seconds, decisions were at least as fast as when remembering a single item. These patterns appeared consistently regardless of differences in list length (4, 5, 6, or 7 to-be-remembered items) and response mode (spoken, selection via mouse). This pattern of findings is inconsistent with the idea that merely holding information in mind conflicts with attention-demanding decision tasks. Instead, it is likely that reconfiguring memory items for responding is the source of conflict between memory and processing in complex span tasks.

Keywords: working memory, complex working memory span, short-term memory, processing speed, response time

Word count: 11908

Impact of memory load on processing diminishes rapidly during retention in a complex span paradigm

Complex working memory span tasks are widely considered the gold-standard for measuring working memory span. In these tasks, participants are required to quickly make an undemanding decision in between presentations of items to remember in order. Something about placing this restriction on memory spans increases their utility: complex spans correlate more strongly than simple memory spans with many cognitive tasks, including measures of reading ability and general intelligence (Abreu, Conway, & Gathercole, 2010; Cowan, 2005; Daneman & Carpenter, 1980; Turner & Engle, 1989). These relationships suggest that understanding how the contents of this privileged mnemonic state are controlled is vital for understanding how we remember, why we forget, and variations in memory both within and across individuals.

A powerful feature of complex span tasks is that two measures, serial recall of a list of memoranda and a series of responses on a simple judgment task, are collected simultaneously. Examining the effects of each task on the other is one way to compare predictions from competing models of working memory. Models of working memory must explain why complex working memory spans are shorter than simple spans, while also predicting relationships between concurrent judgment and memory performance. Some models of working memory propose that multiple specialized modules for maintaining different kinds of memoranda and focusing attention act in concert (Baddeley, 2012; Logie, 2011). According to this logic, multiple resources may be applied to remembering a list, possibly by holding elements of a long list in different formats that load distinct modules, or by applying general attention to memoranda in addition to a specialized resource. During complex span tasks, if one module is devoted to the judgment task and another to the memory task, then little or no overt interference between judgment and memory is expected (Doherty et al., 2019; Doherty & Logie, 2016). Other models of working memory do not propose specialized modules for temporarily maintaining different kinds of information (Barrouillet, Portrat, & Camos, 2011; Cowan, 2005; Oberauer, 2013). Accordingly, these models assume that during complex span tasks, performing simple judgments and remembering serial lists both depend to some extent on some common attentional resource: memory spans measured via complex span should therefore be shorter than memory spans measured by simple span procedures because the interleaved decision task precludes devoting attention entirely to the memoranda. Dividing attention between remembering and processing may also slow processing judgments, relative to when no memoranda are presented. Barrouillet et al.'s time-based resource-sharing model (TBRS) further proposes a specific trade-off between the number of memoranda and processing time: as more memoranda are presented, more time is needed to iteratively refresh the memoranda so as to prevent their decay, and consequently processing judgments should become increasingly slower.

However, evidence of interference between memory recall and processing speed in complex span tasks is perplexingly mixed (Engle, Cantor, & Carullo, 1992; Friedman & Miyake, 2004; Jarrold, Tam, Baddeley, & Harvey, 2011; Maehara & Saito, 2007; C. C. Morey et al., 2018; Saito & Miyake, 2004; Towse, Hitch, & Hutton, 1998; Towse, Hitch, & Hutton, 2000, 2002). Because the decisions required in the judgment tasks are meant to be easy (at least if performed without time pressure), slowing of judgments with respect to some baseline is taken as evidence of conflict. However, our ability to make straightforward predictions about how judgment speed should change as the memory list accumulates is hindered because the effect of memory load on judgment speed appears non-monotonic. C. C. Morey et al. (2018) showed that in children, judgments made after presentation of the first memory item were substantially slower than judgments made after the second memory item, when participants would have been attempting to remember more information. In 8- to 10-year-old children, judgments seemed to slow again after presentation of the second item, revealing a sort of V-shaped pattern. Data from other sources with sufficient granularity suggests that comparable nonlinear fluctuations in processing speed during the complex span procedure are typical (Engle, Cantor, & Carullo, 1992; Friedman & Miyake, 2004). This pattern complicates theorizing: clearly, imposing memoranda influences judgment speed, but not in the linear manner predicted by unitary working memory models. Until we can clearly characterize effects of memory load on judgment speed, all interpretations of complex span performance remain viable, and little progress is made.

One explanation for the nonlinear judgment speeds commonly observed during complex span tasks is that the slow first response reflects the participant's transition into the task. Perhaps the first judgment is slow simply because the participant was recently doing something else (e.g., talking to the researcher, recalling the memoranda from the previous trial). If we consider the first processing judgment to be the first response in a task following a switch from doing something else, then the slowness of this response is not at all surprising, even though some seconds likely passed between the end of the previous trial and the response opportunity. It is well-known that responses immediately following a task switch are substantially slowed with respect to subsequent responses even when the participant is given time to prepare for the switch (e.g., Koch, Poljac, Müller, & Kiesel, 2018; Monsell, 2003; Vandierendonck, Liefooghe, & Verbruggen, 2010). To have any chance of observing a linear increase in processing task times during complex span as the memoranda accumulate, we would need to neutralize these consistent and robust consequences of a task switch. We chose to attempt this by imposing multiple processing judgments prior to the first memory item, so that the first of these judgments would bear the costs of a task switch. In Experiment 1, we introduced complex span trials with four "lead-in" judgments occurring before presentation of any memoranda. Without the lead-in judgments, we should observe the nonlinear pattern observed by C. C. Morey et al. (2018), which cannot be interpreted in terms of memory load, and is likely due to switching. However if the four lead-in judgments prior to the first memory item absorb the costs of the task switch, we may observe one of two interpretable outcomes: 1) little or no slowing with the introduction of the memory list, as expected based on the logic of multiple-component working memory models, or 2) incremental slowing as the memory list increases, as predicted by unitary working memory models. To foreshadow, when we included lead-in trials to absorb switch costs we observed a large response time cost with the introduction of the first memory item and consistent further slowing throughout the accumulation of the list, which contradicts the predictions derived from multiple-component working memory models.

The consistent slowing we observed is compatible with several models of working memory that differ in important ways, so we aimed to further characterize the reason for the slowing observed as memoranda accumulated with additional experiments. This is precisely the pattern one would predict if a single attentional resource were required both to perform the judgments and to "refresh" memoranda, preventing them from deteriorating in between their presentation and the opportunity to recall them some seconds later (Camos et al., 2018). However, the pattern is also consistent with the supposition that the conflict is not caused by maintenance per se, but rather reconfiguring the to-be-remembered information in preparation for responding (Myers, Stokes, & Nobre, 2017; Stokes, 2015), and functionally off-loading it to an appropriate effector system (D. M. Jones & Macken, 2018). Grapheme-to-speech response configuration is presumed to require executive attention (Siegel, 1994), and combined with the notion that output planning for short sequences starts at the start (e.g., Farrell, 2012; Ward & Tan, 2019), the demand of reconfiguration could scale with list length. Reconfiguration to a response-ready format is not the only possible transformation that participants might undertake to preserve information; they may use semantic elaboration, chunking, or evoke visual imagery, any of which might likewise require immediate attention. One important difference between transformation and refreshing is that refreshing is the only alternative that requires evidence of sustained "maintenance" activity. Because refreshing is meant to counteract decay, and memoranda may decay both during list presentation or during a retention interval, models that posit refreshing as necessary for preventing decay naturally predict that memoranda must be attended periodically until the response opportunity occurs. But response configuration or otherwise transforming the memoranda does not necessarily require evidence of sustained maintenance activity. Several lines of evidence that are perplexing if we assume that attention is needed persistently for sustaining memoranda become much clearer if we assume instead that reconfiguring information in preparation for responding provokes interference with concurrent attention-demanding tasks. For instance, the apparent neural "silence" associated with mere retention (Lewis-Peacock, Drysdale, Oberauer, & Postle, 2012) strongly suggests that recently-learned information does not need to be continuously and actively attended to be retrieved later. Likewise, the absence of sustained slowing of judgments when they occur after presentation of the entire to-be-remembered list (Vergauwe, Camos, & Barrouillet, 2014) suddenly makes sense if we assume that the resource-demanding process is reconfiguring the memoranda, rather than ongoing maintenance.

We therefore designed additional studies to test whether temporarily preserving the memory list until recall continues to slow processing judgments, as implied by most unitary working memory models and explicitly predicted by the time-based resource-sharing model (Barrouillet & Camos, 2015). In Experiments 2a and 2b we added trials in which participants completed four "lead-out" judgments after the final memory item, before the opportunity to recall; in Experiments 3a and 3b, we pushed this even further and imposed 8 judgments after the final memory item. If judgments slow during accumulation of the memoranda because attention is shared between making the judgments and refreshing the memoranda, then judgments should remain as slow as during retention of the list, while the participant awaits the opportunity to respond. There is no reason to assume that the memoranda are any less susceptible to decay during this period. However, if the slowing during list presentation reflects reconfiguring the memoranda for use in responding, then judgments after the final item should become quicker because there is no longer any conflict between processes of decision and reconfiguration once reconfiguration is complete. Judgments indeed grew faster after the final memory item, in some cases approaching their pre-list baseline speed. In Experiments 3a and 3b, we also confirmed that these patterns occurred regardless of whether baseline memory-only trials were intermixed with complex-span trials and that interpreting the verbal labels on our tone-judgment response buttons was not the sole source of interference between the memoranda and processing tasks. Over these 5 experiments we consistently found that judgments slowed with the addition of each memory item and speeded progressively after the memory list ended, which is most consistent with the idea that conflict between storage and processing during complex span reflects juggling the requirements of the processing task with reconfiguration of the memoranda in some manner, not continuous re-activation of the memoranda.

Experiment 1

Method

Participants. All of the experiments reported in this manuscript were authorized by our local research ethics committee. Adults aged 18-35 years old were recruited using the participant panel at Cardiff University and received course credit for their participation. We initially aimed to recruit 40 eligible participants, a somewhat larger sample than recently published investigations of effects of storage on processing (e.g., Jarrold, Tam, Baddeley, & Harvey, 2011; Vergauwe, Camos, & Barrouillet, 2014) because we were unsure how large differences between response times to the tone judgments following the first memory item with and without lead-in processing judgments would be. We planned to assess this after data from roughly 40 participants[1] were obtained using Bayes factor analyses, and possibly continue with data collection if results were unclear, as

[1] Supplemental analyses reproducing our tone judgment analyses on perfectly recalled trials only are available at https://osf.io/zw6mj/.

recommended by Schönbrodt and Wagenmakers (2017). Ultimately, we did not add to our initial sample after analyzing the data. Only those participants with normal hearing, normal or corrected-to-normal vision, and no diagnosis of learning disabilities were included in the study. Participants provided written consent before the study began. The sample included 42 participants aged 18 to 24 ($M = 19.55$ years, $SD = 1.23$) after excluding one participant who ignored the memory task during the complex span trials.

Materials. Participants completed letter memory only trials, tone judgment only trials, and complex span trials in which tone judgments were interleaved with presentation of the letter memoranda. All tasks were programmed in PsychoPy3 v3.0.0b7 (Peirce et al., 2019) and run on a desktop computer (14-inch screen set to 1680 $\times$1050 resolution). Participants exclusively used the mouse to respond in all tasks.

Memory only.

To-be-remembered items were randomly drawn from a set of nine consonants – D, F, K, M, Q, S, V, X, Z. We selected these consonants for variability of their places of articulation in the vocal tract with the constraint that no two consonants appeared consecutively in the alphabet (e.g., P and Q or S and T). Based on the place of articulation, F, M, and V are labial; D, S, and Z are coronal; and, K, Q, and X are dorsal consonants. This categorization is consistent with that provided by the International Phonetic Alphabet (International Phonetic Association, 1999).

List-length ranged from 4-7 items with six trials at each list length making for a total of 48 trials. Selection of consonants on each trial was random and controlled within Psychopy. A fixation cross appeared for 1 second at the beginning of each trial. Each memory item was presented for 1 second with a 500-ms inter-stimulus interval. After the final item in the list, a blank screen was shown for 500 ms and followed by the recall screen. Participants spent approximately 15 minutes (with one break occurring halfway through) completing this task.

Tone judgements only.

We chose tones rather than a verbal stimulus for the intervening judgment task so that these stimuli could not be mistaken for the to-be-recalled memoranda (Oberauer & Lewandowsky, 2016). Participants heard one of two tones – the note B (high tone; 308 Hz) or G (low tone; 245 Hz) – and had to decide as quickly as possible if the tone was high or low. After their response, another tone was presented, and so on. Tones were presented through Sennheiser HD280 Pro headphones. Before the task began, participants heard samples of the tones and were informed that only these two tones were used throughout the task. Tones were presented for 750 ms followed by a 250-ms blank screen. Subsequently, the words HIGH and LOW appeared on the right and left of the screen with the mouse pointer in the middle. This part of the task was self-paced and the program would only progress to next tone after the participant had clicked on one of the available choices. Trials were divided into sets of 4 to 11 tone judgments, with 6 trials at each length for a total of 48 trials. This task took approximately 15 minutes.

Complex span task.

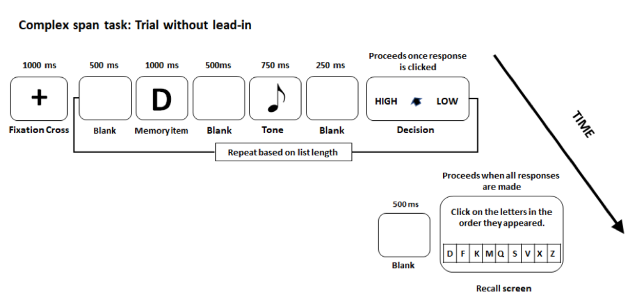

Figure 1. Complex span trial schematic. Lead-in trials were identical except that four tone judgments (with the same timings) preceded presentation of the first to-be-remembered letter.

The memory and judgment tasks described above were combined to create a complex span task. On half the trials, participants viewed a to-be-remembered item, then heard a tone and judged it, interleaving these tasks for 4, 5, 6, or 7 iterations. On the other half of the trials, a set of four 'lead-in' tone judgments were included prior to the introduction of the first to-be-remembered consonant. See Figure 1 for a procedural schematic including timings. List lengths varied from 4 to 7 judgment-letter alternations. There were always either 4 or 0 lead-in tone judgments.

The order of trials in the complex task was pseudo-random with the constraint that no more than two consecutive trials were of the same type (i.e., no more than two lead-in or non-lead-in trials were presented consecutively). This constraint applied to the list length as well – no more than two consecutive trials were of the same list length. There were 48 trials in total – 24 lead-in trials and 24 trials without a lead-in, with 6 trials at each list length per lead-in condition. Two trial orders - one beginning with a lead-in trial and one beginning with a non-lead-in trial - were created with the above constraints to control for order effects, with roughly half of the sample completing the tasks in each order. The order of the baseline memory and judgment tasks were counterbalanced across participants, and the complex span task was always conducted last.

Procedure. Participants were tested in a quiet laboratory with two sound-attenuated cubicles. An experimenter was present throughout the 60-minute testing session, sitting in a control area outside the cubicles after explaining task instructions to the participant personally. Inside a cubicle, participants were seated at a viewing distance of approximately 60 cm from the monitor. Headphones were worn for the duration of each task, but tones were presented only during the tone judgment and complex span task blocks. Data were anonymized by assigning a unique code number to each participant. On-screen instructions were provided for each task and two supervised practice trials were completed before the main trials began. Participants were offered opportunities to take a breaks at set points during each task. An on-screen message indicated that they could take a break and to click on the mouse to continue the task when they felt ready. At the end of each testing session, participants were debriefed.

Results

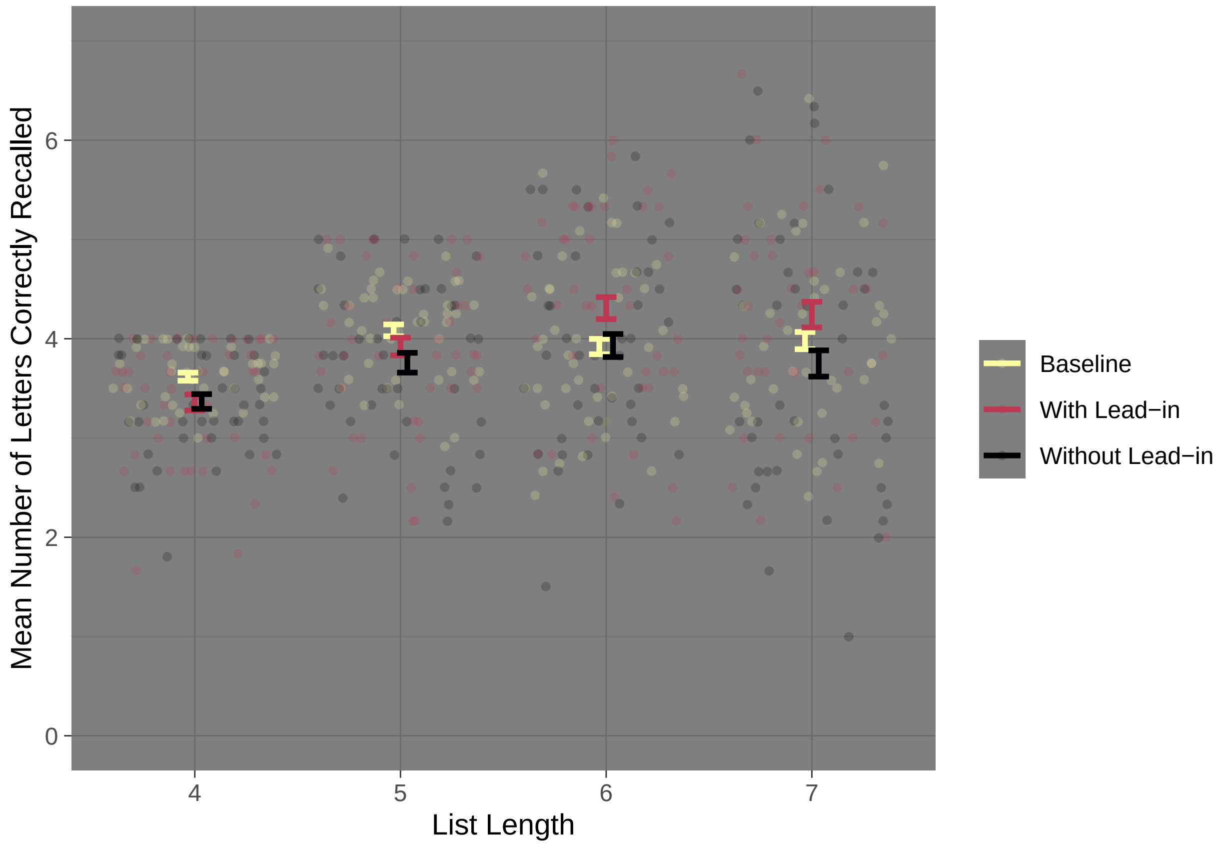

Figure 2. Mean number of items recalled per list length during baseline memory trials and during complex span trials with and without lead-in tone judgments, Experiment 1. Error bars mark within-participant standard errors around the mean calculated with the Cousineau-Morey method (R. D. Morey, 2008). Individual data points are participant-average number of letters recalled

All participants included in the analyses performed adequately on both the memory recall and tone judgment tasks. Figure 2 shows the participants' mean number of letters reported correctly per trial, provided separately for each list length in baseline and both complex span blocks. We present this description to show that the participants included in the analysis typically performed quite well on recall; at list lengths 4 and 5, many participants are performing at ceiling, and averages at list lengths 6 and 7 sit around 4

Table 1

Processing task performance as a function of complex span condition and memory recall, Experiment 1.

Condition	Listwise Recall	Mean Accuracy	SD	Mean RT	SD RT
With Lead-in	Incorrect	0.99	0.02	0.82	0.20
	Perfect	0.99	0.01	0.77	0.18
Without Lead-in	Incorrect	0.98	0.02	0.95	0.30
	Perfect	0.99	0.01	0.87	0.22

Note. N=42. RT = response time. SD = standard deviations.

items. These data also suggest that performance typically declined in the complex span task without lead-in relative to baseline (though note that the baseline trials were always performed first in this experiment).

Similarly, all participants performed well on the tone judgment task. Following typical conventions, we would have excluded a participant from all analyses if their average judgment accuracy fell below 85%. No participant required exclusion, and sample-average judgment accuracy was extremely high (M = 0.99, SD = 0.01 in both baseline and complex span conditions, minimum participants' accuracies 0.96 and 0.95 in the baseline and complex span conditions, respectively). Furthermore, tone judgment response times (given in seconds) were much faster in the baseline block (M = 0.59, SD = 0.14) than during the complex span task (M = 0.84, SD = 0.21), which is consistent with the presumption that remembering letters and making tone judgments conflict.

In complex span and similar dual-task designs, the analysis of judgment response times is often conditioned on correct recall of the whole memory list. The logic is that we can only know that the memory task loaded the judgment task when participants demonstrated perfect recall; under other conditions, participants may have strategically abandoned memoranda in order to excel on the judgment task. We entirely omitted one participant who clearly adopted this strategy by not recalling any memoranda during the complex span block. However, we included trials in our design that were expected to exceed a typical participant's span, so achieving less than perfect recall would not necessarily mean that the participant was not attempting to remember the list. Looking at recall and decision accuracies drawn from the complex span block, it appears that participants engaged with both tasks: processing task accuracies were nearly perfect, and participants demonstrated engagement with the memory task by correctly recalling multiple items from the list most of the time. Rather than remove all trials with any errors in recall a priori, we first looked at whether there was any evidence for a trade-off between letter recall and tone judgment accuracy or speed that would be consistent with the assumption that ignoring the memoranda would benefit tone judgment performance. In Table 1, mean accuracies and response times are given for the tone judgments as a function of complex span trial type and whether the memory list was recalled 100% correctly or not. If anything, tone judgments were *more* accurate and *faster* during memory trials with perfect recall than during trials with at least one memory error. This is opposite to what would be expected if participants strategically prioritized the tone component of the complex span task at the expense of the memory component. In light of this, we did not exclude trials in which memoranda were incorrectly reported from tone task analyses.[2]

Before proceeding with analysis of judgment response times during the complex span task, we used the R package trimr (Grange, 2015) to exclude errors, responses faster than 0.20 seconds and responses more than 2.5 SDs from each participant's mean (4.22% of otherwise valid responses were excluded using this trimming procedure). Figure 3 shows response times (in seconds) plotted by serial position of the processing episode with respect to the accumulating memory list for trials with and without four lead-in processing

[2] Supplemental analyses reproducing our tone judgment analyses on perfectly recalled trials only are available at https://osf.io/zw6mj/.

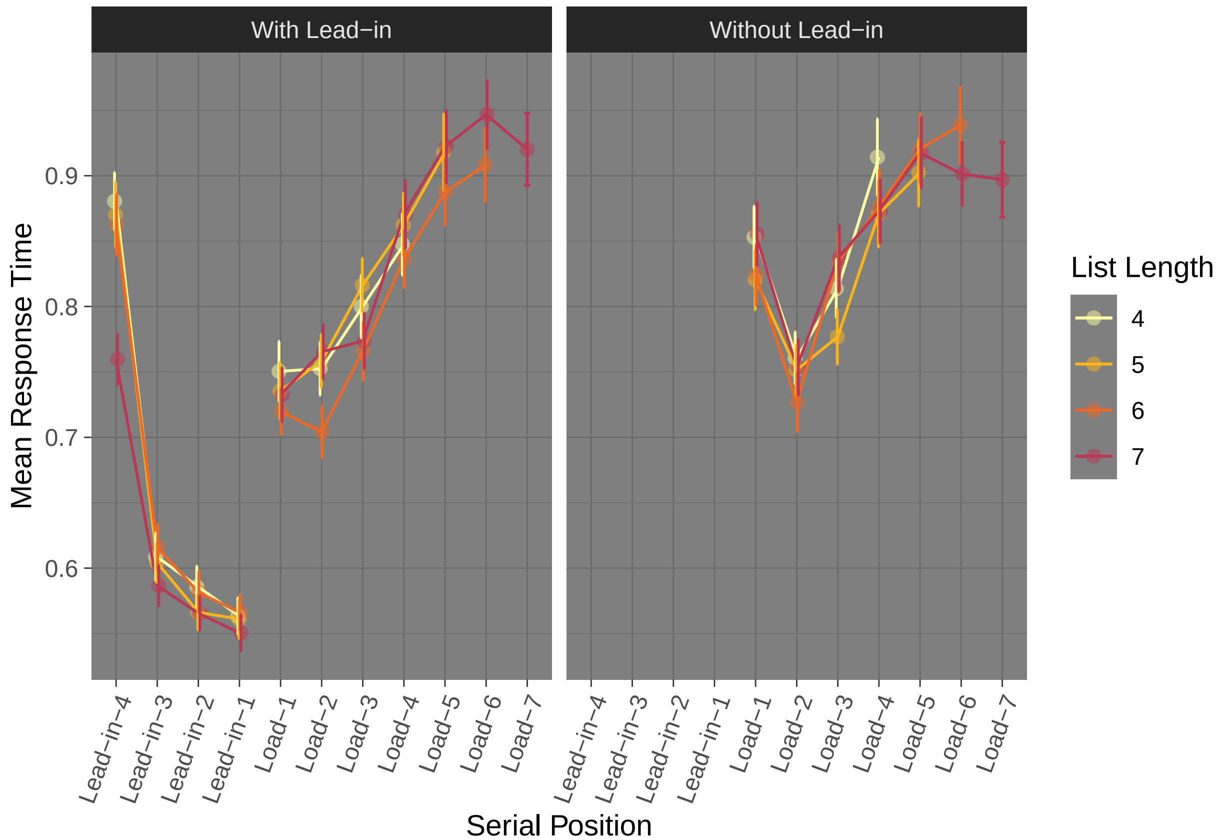

Figure 3. Mean tone judgment response times (in seconds) during complex span trials with and without lead-in tone judgments, by list length and serial position in the trial sequence, Experiment 1. Error bars are within-participant standard errors of the mean calculated with the Cousineau-Morey method (R. D. Morey, 2008).

judgments. Without the lead-in judgments (right panel), we observe a non-linear pattern similar to the pattern observed by C. C. Morey et al. (2018), with slow responses to the first memory item, speeding with the second item, then slowing through the remainder of the list. However, including lead-in judgments (left panel) appeared to change this pattern. While the first tone judgment (whether it was Lead-in-4 or Load-1) was always among the slowest, the other lead-in judgments were much faster, and comparable to the average response time on baseline trials. The best model revealed by a Bayesian ANOVA estimated using the BayesFactor R package (Richard D. Morey et al., 2018; Rouder, Morey, Speckman, & Province, 2012) with lead-in condition (with lead-in, without lead-in), serial position (restricted from Load-1 to Load-7), and list length (4, 5, 6, and 7) as within-participants factors performed on log-transformed trimmed response times included main effects of lead-in condition, serial position, and an interaction between lead-in condition and serial position, $BF = 5.5 \cdot 10^{70}$. Evidence favoring inclusion of the interaction (obtained by comparing the BF of the best model against the BF of the model without the interaction but including the same main effects) was decisive, $BF = 1{,}787.41$.

In line with our aims, we performed post-hoc comparisons to better interpret the effect of the lead-in on the judgments co-occurring with the first few memory items and to characterize the apparent slowing of judgments toward the end of the list. First, we confirmed that the interaction between lead-in condition and serial position was due to differences in response times between conditions at the start of the complex span sequences by considering only the processing judgments occurring after the first two memory items. This restriction allowed the possibility to check whether the change from Load-1 to Load-2 differed per lead-in condition. The best model indeed included main effects of lead-in condition, serial position, and an interaction between lead-in condition and serial position, $BF = 7 \cdot 10^{13}$. Including the interaction term was favored (compared with the model including the same effects but no interaction) decisively, $BF = 5{,}039{,}806.82$. Referring to the values in Figure 3, this interaction must reflect the relatively slow responses after the first memory item without the lead-in ($M = 0.84$, $SD = 0.24$) versus with the lead-in ($M = 0.73$, $SD = 0.22$); response times after the second (without lead-in $M = 0.75$, $SD = 0.22$; with lead-in $M = 0.75$, $SD = 0.22$) memory item did not seem to differ. It therefore appears that one effect of the lead-in judgments is to displace the awkward first judgment, which probably reflects re-engagement with the task after the previous trial, to a theoretically less interesting position with respect to the memoranda. This displacement has the desirable effect of rendering the response time serial position curve more linear, and thus simpler to interpret.

We also separately considered response speeds from the final three items from each list length to test whether, regardless of list length, comparable increases in tone judgment response times were observed as the list accumulated. Here, the best model included only a main effect of serial position, $BF = 9.5 \cdot 10^{21}$. Excluding the lead-in condition was favored by a factor of 31.45, which suggests that the lead-in is no longer influencing judgment times by the end of the list. It is clear from Figure 3 that the effect of serial position reflects slowing of judgment responses as the lists progressed.

We can assess whether encoding the first memory item incurred an unusual cost by considering the size of the increase in response times between the Lead-in-1 position, where there was not yet anything to simultaneously remember, and the Load-1 position, where there was a single letter to remember. A Bayesian ANOVA on processing response times including task block (baseline or complex span with lead-in), list length (note that baseline trials were organized in clusters consistent with complex span list lengths), and position of the judgment (referring to Figure 3, positions Lead-in-1 and Load-1) supported a model including main effects of task block and position, plus an interaction between them, $BF = 4.1 \cdot 10^{50}$. The interaction was favored by a factor of $1.3 \cdot 10^{22}$ over the model including only the same main effects. The interaction must be due to the slowing incurred after introduction of the first memory item in the complex span sequence ($M = 0.68$, $SD = 0.18$); the processing judgment prior to the first memory item ($M = 0.55$, $SD = 0.11$) was comparable to the baseline processing judgments (position corresponding to Lead-in-1 $M = 0.57$, $SD = 0.12$; position corresponding to Load-1 $M = 0.57$, $SD = 0.14$).

Discussion

Introducing four lead-in tone judgments clarified the cost of remembering an accumulating list of letters on judgment response times. Without lead-in judgments, we observed a non-linear pattern of speeding and slowing on the judgments as the memory list accumulated, similarly to C. C. Morey et al. (2018). With lead-in judgments, we observed substantial slowing to tone judgments with the introduction of the first memory item, followed by steadily increasing slowing as the memory list progressed. This pattern is consistent with the idea that a common resource is needed both to judge the tones and to remember the letters. The pattern of slowing we observed is predicted explicitly by the TBRS model of working memory (Barrouillet & Camos, 2015). According to TBRS, a common attentional resource is used to perform attention-demanding tasks and to serially refresh to-be-remembered information. TBRS therefore attributes the slowing to participants refreshing increasingly long lists of letters during each subsequent tone judgment. This account is generally consistent with unitary models of working memory as well. However, under this logic one would also expect that memory recall would decrease whenever intermittent judgments were imposed, regardless of whether lead-in judgments were included. We did not observe a clear cost to memory recall in complex span with lead-in judgments compared to baseline, nor did we observe evidence that successful retention of memoranda (that is, remembering a list perfectly) introduces steeper judgment costs than partial remembering. These findings suggest that the conflict between memory and judgment in complex span might occur for another reason.

Overall, Experiment 1 appeared to confirm that dual-task costs occur in complex working memory span paradigms. Participants recalled numerically fewer items on average in the typical complex span scenario (e.g., without lead-in judgments) than in the baseline condition. Participants also responded much more slowly to the tone stimuli during complex span administration than during baseline. These commonly-observed interference patterns are inconsistent with some assumptions of multiple component working memory, namely that resources used to direct attention are distinct from short-term storage. In Experiments 2a and 2b, we therefore shifted to testing whether these conflicts occur because stored memoranda must be refreshed to prevent loss, or because memoranda are translated into another form, possibly in preparation for responding. In Experiments 2a and 2b, we introduced conditions including four tone judgments after presentation of the final memory list item. If conflict between storage and processing occurs because attention is needed to prevent the memoranda from decay (Barrouillet & Camos, 2015), then judgment times should remain slow after presentation of the list, because the list must still be sustained via attentional refreshing. However, the conflict between storage and processing may also be attributed to reconfiguration of the memoranda (Myers, Stokes, & Nobre, 2017), perhaps into a representation transferable to its output form, in preparation for making a response. If so, tone judgments should speed again after this transformation has taken place, at some point in between presentation of the final list item and the opportunity for responding. We manipulated response modality, with responses made via mouse input in Experiment 2a (as in Experiment 1) and via speech in Experiment 2b, to descriptively assess consistency of effects of verbal memory load on processing performance.

Experiments 2a and 2b

Method

Participants. The sample for Experiment 2a included 31 adults aged 18 to 35 (M = 22.06 years, SD = 3.53) who had not taken part in Experiment 1. No participants were excluded based on recall or judgment performance. Experiment 2b included 16 new participants aged 20 to 54 (M = 25.88 years, SD = 8.79). Criteria for participation were the same as in Experiment 1. Three participants from Experiment 2b were excluded because they recalled an average of 2 or fewer items from 4-item lists, leaving N = 13. Because we observed such large effects of the key factors in Experiment 1, we stopped initial data collection with fewer participants. We did not need to implement sequential analysis in Experiment 2a; we found that the analyses were sufficiently convincing with the initial N = 31. In Experiment 2b, we examined the tone task data after acquiring eleven participants in order to provide a trainee researcher the chance to analyze data. We continued collecting data afterward until it was convenient to stop due to participant sample availability, without first examining the recall data.

Materials. In Experiment 2a, baseline tasks – the memory-only and processing-only tasks – were identical to those used in the previous experiment. The complex span task was modified such that on half the trials, in addition to the lead-in judgments, four processing items were added after presentation of the final list item. On these trials, after participants had been presented with list length alternating memory items and processing judgments, they completed four 'lead-out' processing judgments without a corresponding memory item before moving to the recall phase. Stimulus selection, presentation, and timings for all tasks were otherwise identical to Experiment 1.

Experiment 2b differed from Experiment 2a in that spoken recall was elicited. Instead of the mouse-driven response screen, at recall during both the memory-only and complex span tasks, participants heard an aural prompt (an artificial voice saying "Recall now") and spoke their response into a desk-top microphone. Responses were recorded for later transcription and verification. We reduced the number of trials by testing only list lengths 4 and 6.

Results

Experiment 2a. Figure 4 shows the average number of letters reported correctly across trials per participant and list length in both baseline and complex span trials. No participant's data needed to be omitted due to poor overall recall performance, and the overall patterns suggest that participants typically recalled multiple items from the list correctly. Similarly, no participant failed to achieve the 85% accuracy criterion on the processing task, and sample-average processing accuracy was extremely high ($M = 0.99$, $SD = 0.01$ in the baseline condition, and $M = 0.98$, $SD = 0.03$ in the complex span conditions; minimum participants' accuracies were 0.94 and 0.87 in the baseline and complex span conditions, respectively). Again, processing task response times (given in seconds) were much faster in the baseline block ($M = 0.60$, $SD = 0.11$) than during the

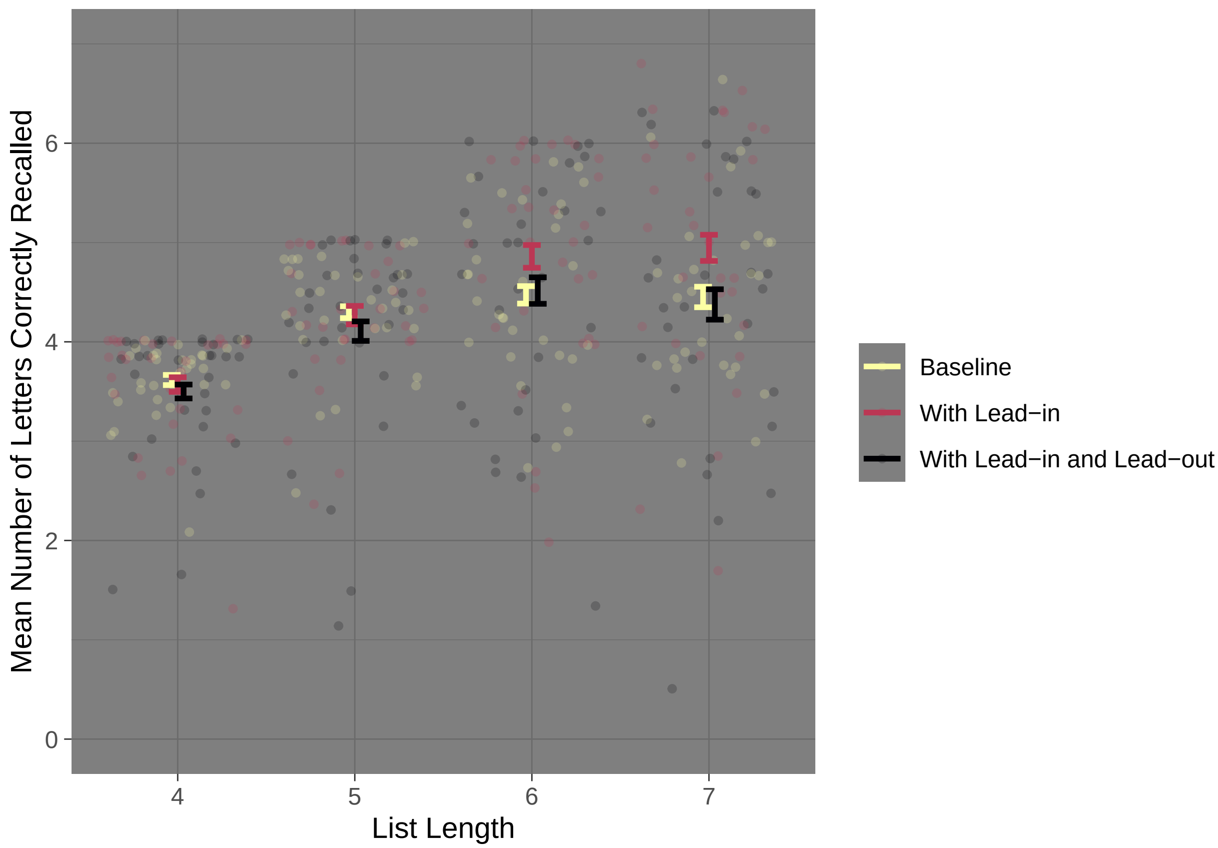

Figure 4. Mean number of items recalled per list length during baseline memory trials and during complex span trials with and without lead-in tone judgments, Experiment 2a. Error bars mark within-participant standard errors around the mean calculated with the Cousineau-Morey method (R. D. Morey, 2008). Individual data points are participant-average number of letters recalled

complex span task ($M = 0.83$, $SD = 0.28$), consistent with the assumption of some dual-task cost between storage and processing.

As in Experiment 1, we examined tone judgment performance as a function of accuracy in memory recall to check whether it was essential to condition analysis of the tone judgments on perfect recall. Table 2 provides these descriptive statistics, which show no reason to presume that participants traded-off accuracy on the storage task for accuracy or speed on the tone judgments.

Table 2

Processing task performance as a function of complex span condition and memory recall, Experiment 2a.

Condition	Listwise Recall	Mean Accuracy	SD	Mean RT	SD RT
With Lead-in	Incorrect	0.98	0.03	0.86	0.31
	Perfect	0.99	0.02	0.80	0.23
With Lead-in and Lead-out	Incorrect	0.97	0.03	0.87	0.31
	Perfect	0.98	0.03	0.77	0.19

Note. N=31. RT = response time. SD = standard deviations.

Before proceeding with analysis of processing response times during the complex span task, we trimmed responses as described in Experiment 1. We excluded 4.98% of otherwise valid responses based on this trimming procedure. Mean response times plotted as a function of lead-out condition, list length, and serial position are given in Figure 5. As we observed in Experiment 1, processing judgment response times decreased after the first lead-in judgment, but increased substantially when the first memory item was introduced, and continued increasing as the list accumulated. We ran a 3-way Bayesian ANOVA on log-transformed trimmed response times with processing cluster (lead-in responses, memory load responses, or lead-out responses), list length, and lead-in condition (lead-in only, or lead-in and lead-out) as factors. The best model included main effects of list length and processing cluster, $BF > 1$ million. The model including the processing cluster factor was preferred over a model including only list length by an overwhelming margin, $BF > 1$ million. We tested whether each of the three processing clusters differed from the other in average processing time by comparing output from models with simpler codings equating two levels of processing cluster (e.g., lead-in = lead-out, or memory load = lead-out) with

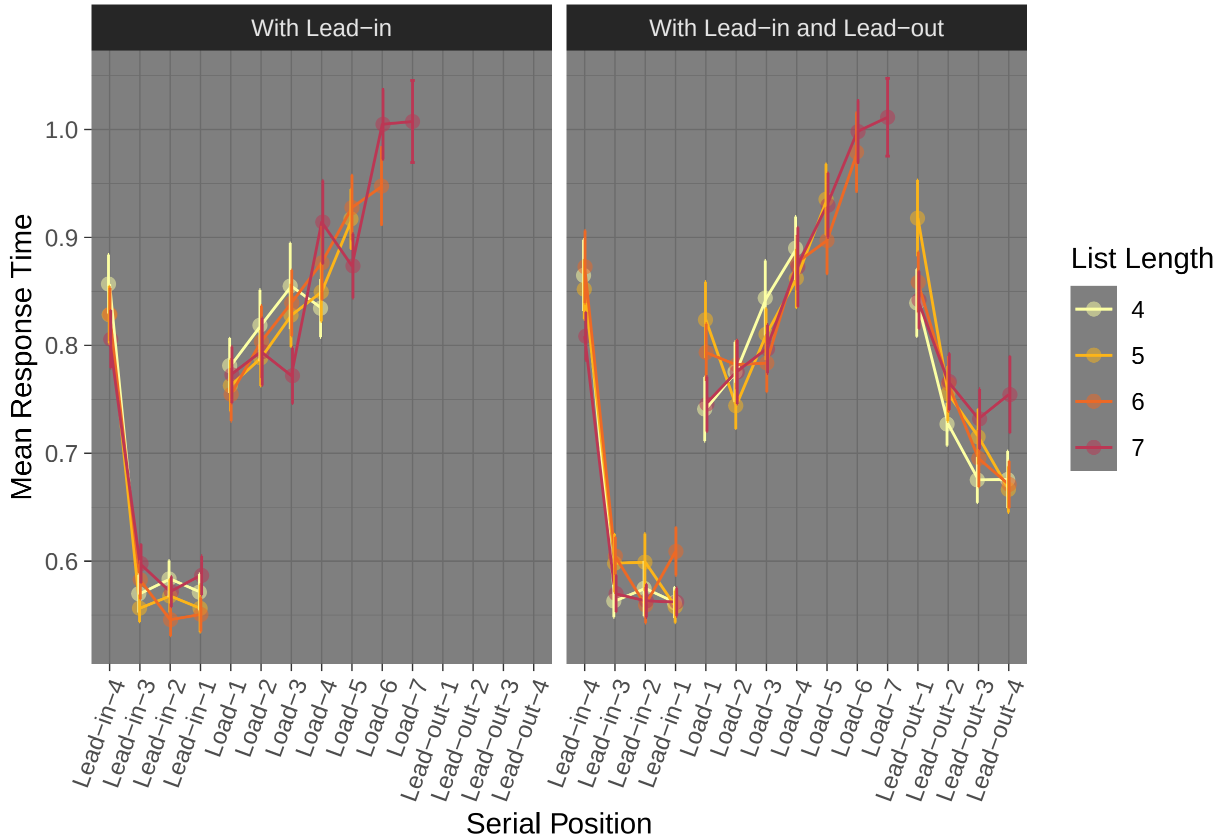

Figure 5. Mean processing response times (in seconds, means normalized) for complex span trials with only lead-in processing judgments (left) or both lead-in and lead-out judgments (right), by list length and serial position, Experiment 2a. Error bars are within-participant standard errors of the mean (R. D. Morey, 2008).

the output from the original 3-level coding. The best model (preferred by a factor of more than 1 million over the next-best simplification of the coding, in which the memory load and lead-out responses were assumed to be equivalent) distinguished all three levels of processing cluster, confirming differences between each level. Judgments were slowest when participants were simultaneously given a letter to remember ($M = 0.85$, $SD = 0.29$), fastest for lead-in judgments ($M = 0.64$, $SD = 0.15$), and intermediate for lead-out judgments ($M = 0.76$, $SD = 0.21$).

As we observed in Experiment 1, there was a stark difference between response times on the last lead-in decision and the decision following presentation of the first memory item. A Bayesian ANOVA on log-transformed response times for these two judgments only confirmed that the best model included this effect ($BF = 2.5 \cdot 10^{97}$) with no effects of list length or lead-out condition, nor any interactions (exclusions were favored by factors of at least 20.77). After observing the first to-be-remembered letter, tone judgments took an average of 0.77 seconds, compared to 0.57 prior to introduction of the letters. As in Experiment 1, this lead-in average RT is comparable to average RTs in the baseline block. This confirms that adding just the first item of an accumulating memory load slows judgment responses substantially.

Figure 5 shows speeding across the lead-in period and slowing during accumulation of the memory list similar to what we observed in Experiment 1. We carried out a fine-grained analysis of the lead-out judgments, plus the judgments at the start and end of the lists and the final lead-in judgment for comparison. The best model from a Bayes factor ANOVA focusing on the lead-out condition including terms for list length and processing position (including all of the lead-out judgments, the first and final memory load judgments, and the final lead-in judgment) included only the main effect of processing position, $BF = 3.5 \cdot 10^{138}$. Excluding effects of list length or an interaction was favored by a factor of 5.22. As shown in Figure 5, lead-out judgments became quicker (first lead-out $M = 0.87$, $SD = 0.31$; final lead-out $M = 0.70$, $SD = 0.20$; $BF = 2.8 \cdot 10^{17}$). By the third lead-out judgment ($M = 0.71$, $SD = 0.18$), decision speeds were faster than after the presentation of the first memory item ($M = 0.78$, $SD = 0.34$; $BF = 30.08$). The final lead-out judgment however did not become as fast as the final lead-in judgment ($M = 0.57$, $SD = 0.13$; $BF = 3.6 \cdot 10^{11}$).

Experiment 2b. We conducted Experiment 2b to confirm that similar patterns of tone judgments and recall were observed if participants were required to give spoken responses. Because transcribing spoken responses for analysis is so laborious, we acquired only a small sample of participants and provide a descriptive analysis to show that patterns

Table 3

Descriptive statistics of participants' mean number of items recalled correctly per list length, Experiment 2b.

Condition	List Length	Minimum	Maximum	Mean	SD
Baseline	4.00	2.50	4.00	3.49	0.47
	6.00	1.33	5.00	3.32	1.07
With Lead-in	4.00	1.83	4.00	3.47	0.61
	6.00	2.50	5.67	3.96	0.91
With Lead-in and Lead-out	4.00	2.50	4.00	3.53	0.50
	6.00	3.67	5.33	4.46	0.70

Note. N = 13.

observed are consistent with those seen with mouse-driven responding. No participant failed to achieve the 85% accuracy criterion on the processing task, and sample-average processing accuracy was again extremely high ($M = 0.99$, $SD = 0.01$ in the baseline condition, and $M = 0.98$, $SD = 0.02$ in the complex span conditions; minimum participants' accuracies were 0.96 and 0.93 in the baseline and complex span conditions, respectively). Again, processing task response times (given in seconds) were much faster in the baseline block ($M = 0.66$, $SD = 0.16$) than during the complex span task ($M = 0.93$, $SD = 0.20$). Furthermore, the data gave no reason to believe that participants abandoned the memory task to perform better in the judgment task (see Table 4).

Figure 6 shows now-familiar patterns of judgment response times as a function of processing cluster (lead-in, with accumulating memory load, or lead-out) and load. Descriptively, the outcomes are similar to those of Experiment 2a. Judgments appeared slowest when participants were simultaneously given a letter to remember ($M = 0.95$, $SD = 0.21$), fastest for lead-in judgments ($M = 0.71$, $SD = 0.13$)), and intermediate for

Table 4

Processing task performance as a function of complex span condition and memory recall, Experiment 2b.

Condition	Listwise Recall	Mean Accuracy	SD	Mean RT	SD RT
With Lead-in	Incorrect	0.98	0.02	0.98	0.26
	Perfect	0.99	0.02	0.88	0.16
With Lead-in and Lead-out	Incorrect	0.96	0.03	0.99	0.22
	Perfect	0.98	0.03	0.86	0.15

Note. N=13. RT = response time. SD = standard deviations.

lead-out judgments ($M = 0.86$, $SD = 0.18$)). Figure 6 also suggests that tone judgments slowed immediately after initiation of the list presentation, and continued to slow further as each item was presented, consistently with the outcomes of Experiments 1 and 2a.

Discussion

In two further experiments, we observed similar patterns of speeding on a tone judgment task during a lead-in period and slowing with the accumulation of a memory list to those documented in Experiment 1. We also learned that after presentation of the memory list ends, tone judgments speed up again. By the third lead-out tone judgment, decision speeds were faster than those associated with the first memory item. This speeding is difficult to explain under the assumptions of the TBRS model (Barrouillet & Camos, 2015), which holds that attentional refreshing of the to-be-remembered memoranda would compete for attentional resources with the tone judgment task, both as the memoranda accumulate and during any delay period. Our findings suggest that some other process that does not necessarily persist into a retention interval requires access to an

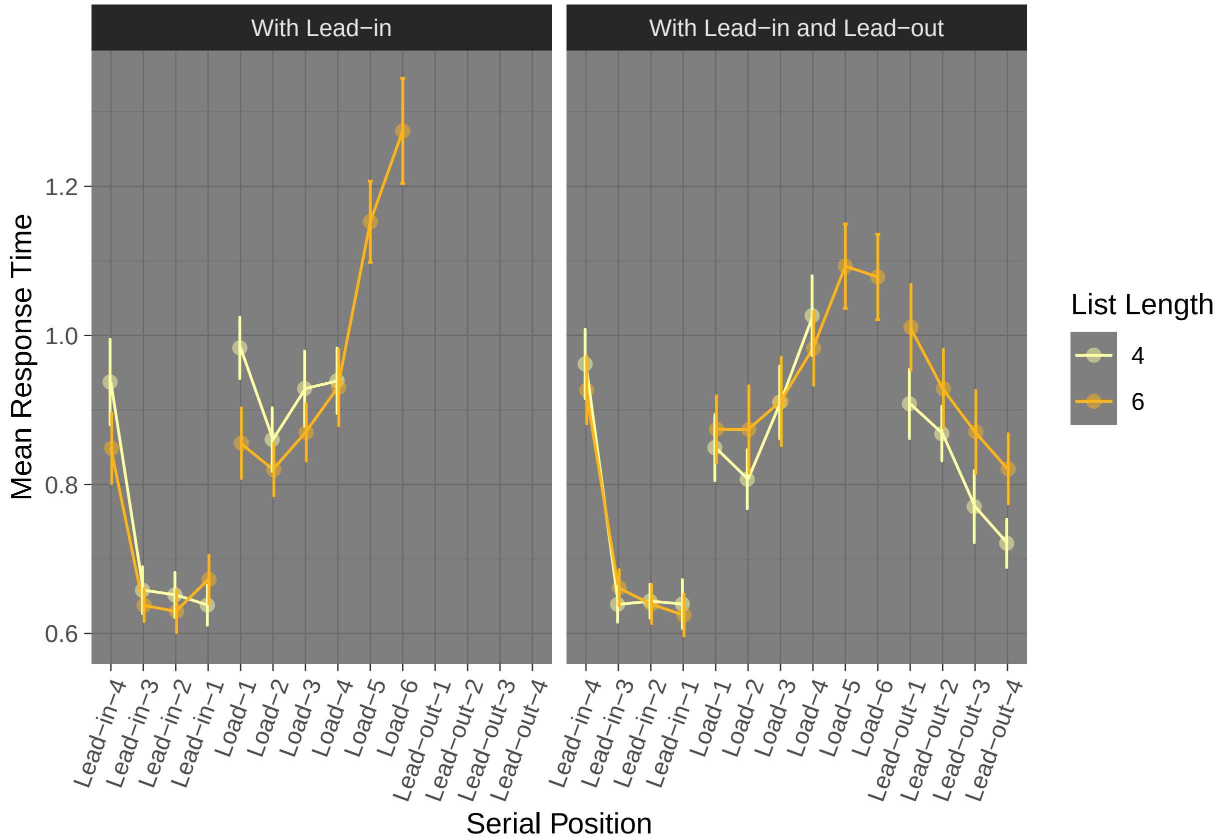

Figure 6. Mean processing response times (in seconds, means normalized) for complex span trials with only lead-in processing judgments (left) or both lead-in and lead-out judgments (right), by list length and serial position, Experiment 2b. Error bars are within-participant standard errors of the mean (R. D. Morey, 2008). Trimming procedure (same as described in Experiments 1 and 2a) resulted in exlcusion of 5.46% of data.

attentional resource that overlaps with the resources needed to judge the tone stimuli. This process could be reconfiguring the memoranda in preparation to respond (Myers, Stokes, & Nobre, 2017; Stokes, 2015), or perhaps otherwise transforming the memoranda,for instance via elaboration (Bartsch, Singmann, & Oberauer, 2018) or associative learning, (Cowan, 2019; G. Jones & Macken, 2015).

One outcome we have consistently observed that does not adhere well to either the predictions of TBRS, unitary working memory models, or the proposition that reconfiguration of memoranda during serial memory and discrete judgment tasks share the same resource is that imposing four lead-in judgments appears to abolish any cost to recalling the memoranda correctly in complex span compared with a single-task baseline. However, we are not yet convinced that lead-in judgments indeed render recall during complex span cost-free. In each experiment we have presented so far, baseline trials were administered at the start of the session, prior to complex span trials. Possibly, complex span recall benefited from practice across the session. We therefore carried out two additional experiments, randomly mixing baseline memory-only trials into the complex-span procedure, so that our baseline measures would be distributed across the session in the same manner as the complex span trials.

In these two experiments, we also took the opportunity to increase the number of lead-out trials, lengthening the lead-out period. It is plausible that after the memory list has been presented, the refreshing needed to prevent decay changes; perhaps sporadic refreshing is sufficient to prevent decay of the list, or possibly individual elements in the list are grouped and may be more efficiently refreshed than they were during list presentation. Entertaining any of these assumptions could lead to predicting that lead-out judgments would grow faster after list presentation is complete, but not faster than judgments associated with a memory load of one item. We added four additional lead-out judgments, so that the lead-out phase always included 8 judgments, to learn whether judgments with a memory load eventually become as fast as the lead-in judgments just before the introduction of the memory load. Experiment 3a closely replicated Experiment 2a, except that baseline serial recall trials were intermixed with complex span trials, and lead-out trials always included 8 tone judgments after the final list item was presented. Experiment 3b differed from Experiment 3a only in that we replaced the typed words “HIGH” and “LOW” during the tone processing judgments with visuo-spatial representations, in order to confirm that it was not merely reading these verbal representations that provoked interference between the letter recall and tone judgment tasks.

Experiments 3a and 3b

Method

Participants. Experiment 3a included 20 adults aged 18 to 22 (M = 21 years, SD = 1.15) who had not taken part in Experiment 1. One participant was excluded based on baseline recall of 2 or fewer out of 4 items, leaving N = 19. Experiment 3b included 33 new participants aged 18 to 27 (M = 21 years, SD = 2.13). Criteria for participation were the same as in Experiment 1. Four participants from Experiment 3b were excluded because they recalled an average of 2 or fewer items from 4-item lists or performed below 85% correct in the tone judgment task, leaving N = 29. In both experiments, we stopped initial data collection based on experimenter convenience. We did not collect additional data after stopping for analyses in either experiment.

Materials. In Experiment 3a, the baseline processing-only task was identical to those used in previous experiments. We added baseline memory trials to the Experiment 2a complex span task. Half the complex span trials included eight lead-out processing items after the final memory item. Participants completed trials with memory lists of 4, 5, 6, and 7 items. We also introduced variable delay periods to the baseline and lead-in only trials so that there were sometimes delay periods comparable to the duration of the lead-out period. Stimulus selection, presentation, and timings otherwise remained the same.

Experiment 3b differed from Experiment 3a only in that the verbally-labeled onscreen response options for the tone judgment task were replaced with spatially-arranged arrow buttons (an upwards-facing arrow near the top of the screen for the "high" response and a downwards-facing area near the bottom of the screen for the "low" response). We included this to confirm that interference between the tone judgment and letter memory tasks was not merely due to reading the verbal tone judgment response options.

Results

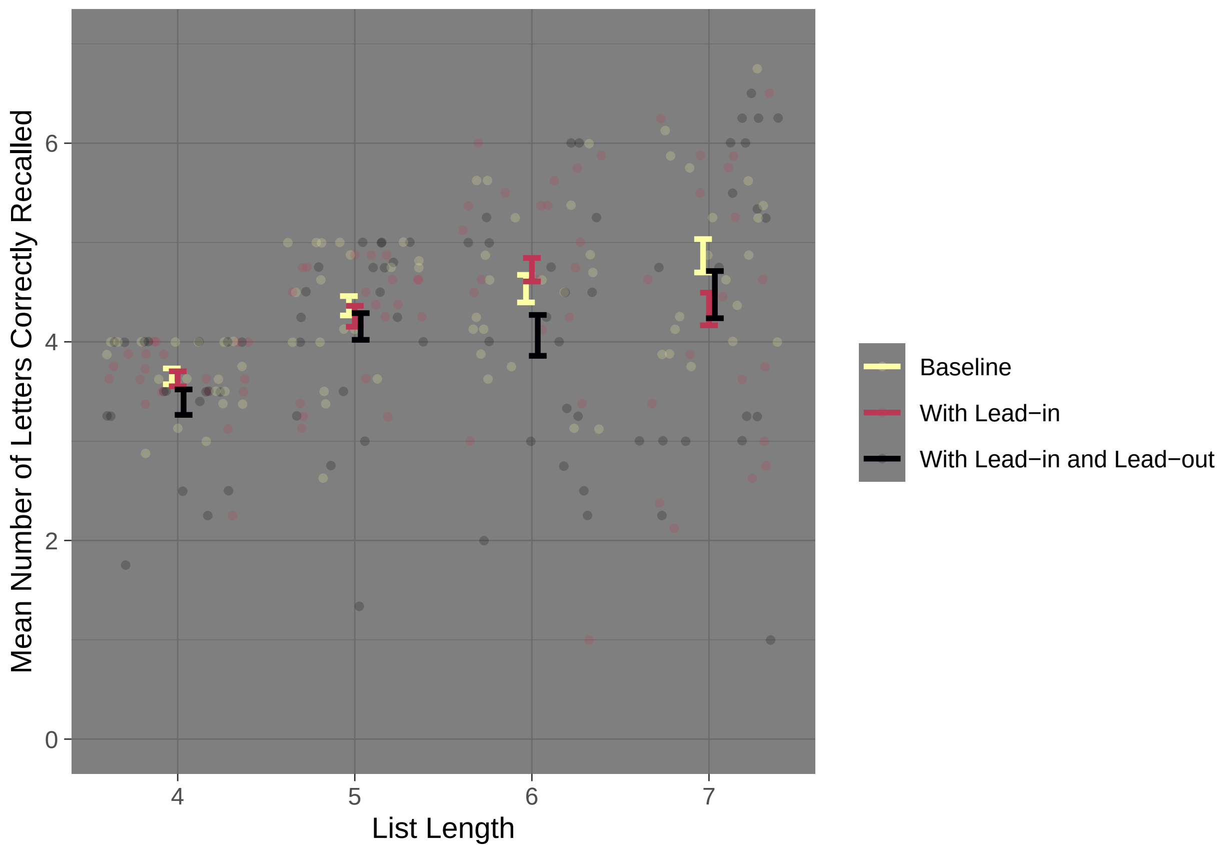

Figure 7. Mean number of items recalled per list length during baseline memory trials and during complex span trials with and without lead-in tone judgments, Experiment 3a. Error bars mark within-participant standard errors around the mean calculated with the Cousineau-Morey method (R. D. Morey, 2008). Individual data points are participant-average number of letters recalled

Experiment 3a. Figure 7 shows descriptive statistics for memory recall. The overall patterns suggest that participants typically recalled multiple items from the list correctly. Sample-average processing accuracy was extremely high ($M = 0.99$, $SD = 0.01$ in the baseline condition, and $M = 0.98$, $SD = 0.03$ in the complex span conditions; minimum participants' accuracies were 0.98 and 0.90 in the baseline and complex span conditions, respectively). Again, processing task response times (given in seconds) were faster in the baseline block ($M = 0.59$, $SD = 0.09$) than during the complex span task (M

Table 5

Processing task performance as a function of complex span condition and memory recall, Experiment 3a.

Condition	Listwise Recall	Mean Accuracy	SD	Mean RT	SD RT
With Lead-in	Incorrect	0.97	0.05	0.71	0.30
	Perfect	0.98	0.02	0.63	0.25
With Lead-in and Lead-out	Incorrect	0.98	0.03	0.65	0.23
	Perfect	0.98	0.04	0.59	0.19

Note. N=19. RT = response time. SD = standard deviations.

= 0.64, SD = 0.24).

As in previous experiments, we examined tone judgment performance as a function of accuracy in memory recall to check whether it was essential to condition analysis of the tone judgments on perfect recall. Table 5 provides these descriptive statistics, which again show no reason to presume that participants traded-off accuracy on the storage task for accuracy or speed on the tone judgments.

Tone judgment responses were trimmed by the same process described previously, resulting in the exclusion of 10.15% of otherwise valid responses. Mean response times plotted as a function of lead-out condition, list length, and serial position are given in Figure 8. Again, processing judgment response times decreased after the first lead-in judgment, but increased substantially when the first memory item was introduced, and continued increasing as the list accumulated. After list presentation was complete, judgment times decreased, apparently back to the lead-in baseline speed. We ran a 3-way Bayesian ANOVA on log-transformed trimmed response times with processing cluster (lead-in responses, memory load responses, or lead-out responses), list length, and lead-in

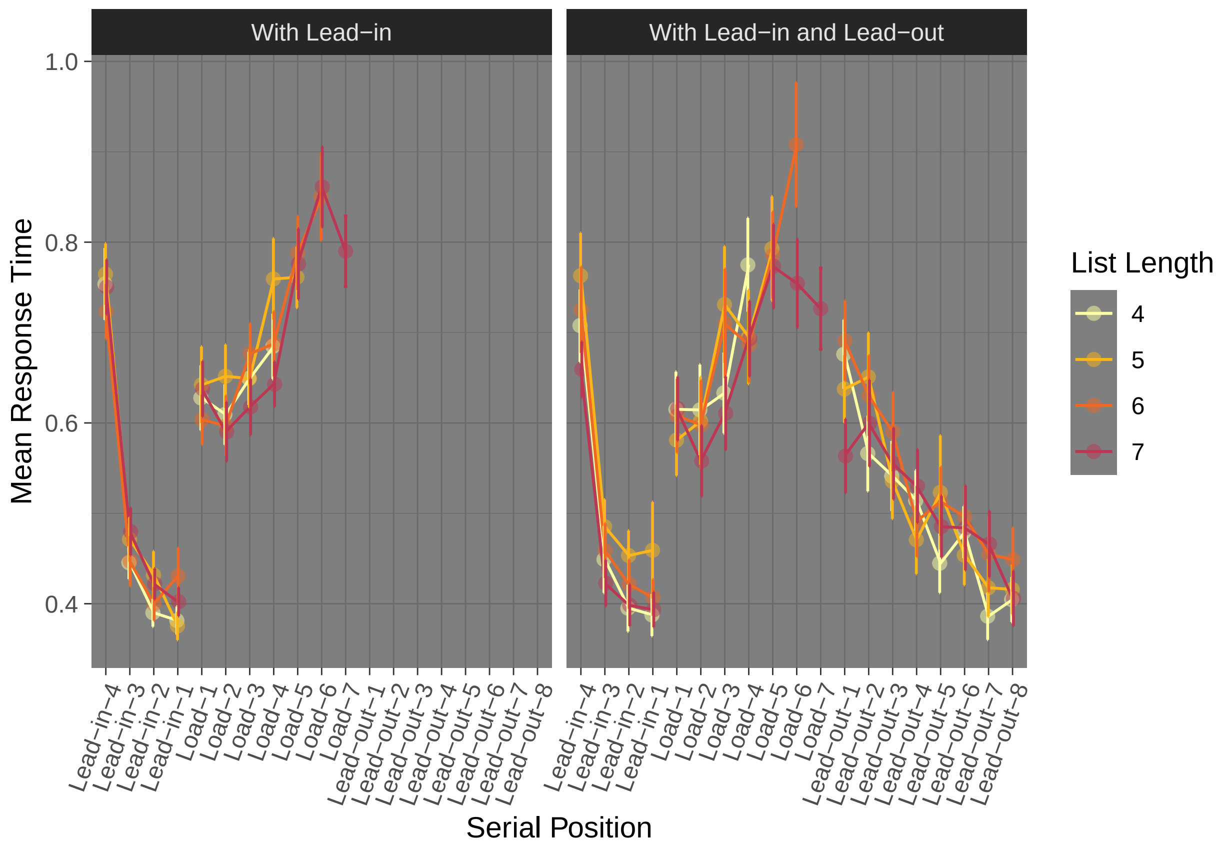

Figure 8. Mean processing response times (in seconds, means normalized) for complex span trials with only lead-in processing judgments (left) or both lead-in and lead-out judgments (right), by list length and serial position, Experiment 3a. Error bars are within-participant standard errors of the mean (R. D. Morey, 2008).

condition (lead-in only, or lead-in and lead-out) as factors. The best model included only a main effect of processing cluster, $BF = 1.5 \cdot 10^{156}$. Excluding other terms was strongly favored by factors of at least 37.05. We tested whether response times in each of the three processing clusters differed from the other by comparing models with simpler codings equating two levels of processing cluster (e.g., lead-in = lead-out, or memory load = lead-out) with the output from the original coding differentiating all three levels. The best model used a simplified coding of processing cluster, in which lead-in and lead-out judgments were considered equivalent, $BF = 5.6 \cdot 10^{156}$. This simpler model was favored over the full coding of processing cluster by a factor of 3.75. Judgments were slower in the memory load cluster ($M = 0.68$, $SD = 0.26$), than in the lead-in ($M = 0.50$, $SD = 0.14$)), and lead-out clusters ($M = 0.52$, $SD = 0.13$)). We followed this test with fine-grained t-tests comparing critical transitions across a trial. First, judgment times after the first memory item was introduced were convincingly slower than those immediately prior, $BF = 2.1 \cdot 10^{40}$; see Figure 8 for descriptive values. We compared average lead-out judgment RTs (from the 2nd onward) with the first value just after the list began. From the fourth lead-out judgment, there was evidence to support the position that lead-out judgments were faster than the judgment after one memory item, $BF = 96.83$.

We conducted a Bayesian ANOVA on lead-out responses only in order to check for influences of list length. Consistent with the absence of list length effects in the omnibus analysis, the best model included only a main effect of serial position, $BF = 1.5 \cdot 10^{23}$. Excluding list length (or its interaction with position) was favored by a factor of at least 46.94.

Experiment 3b. Figure 9 shows descriptive statistics for memory recall, which again show that participants typically recalled substantial portions of the lists correctly. Sample-average processing accuracy was extremely high ($M = 0.99$, $SD = 0.01$ in the baseline condition, and $M = 0.98$, $SD = 0.01$ in the complex span conditions; minimum participants' accuracies were 0.95 in the baseline and complex span conditions). Here, processing task response times (given in seconds) do not appear to differ much in the baseline block ($M = 0.63$, $SD = 0.14$) compared with the complex span task ($M = 0.64$, $SD = 0.19$) trials.

Table 6 provides mean accuracies and response times on the tone judgment task with and without perfect recall. Once again, it does not appear worthwhile to limit analysis of response times to conditions in which perfect recall occurred.

Tone judgment responses (see Figure 10) were trimmed by the same process

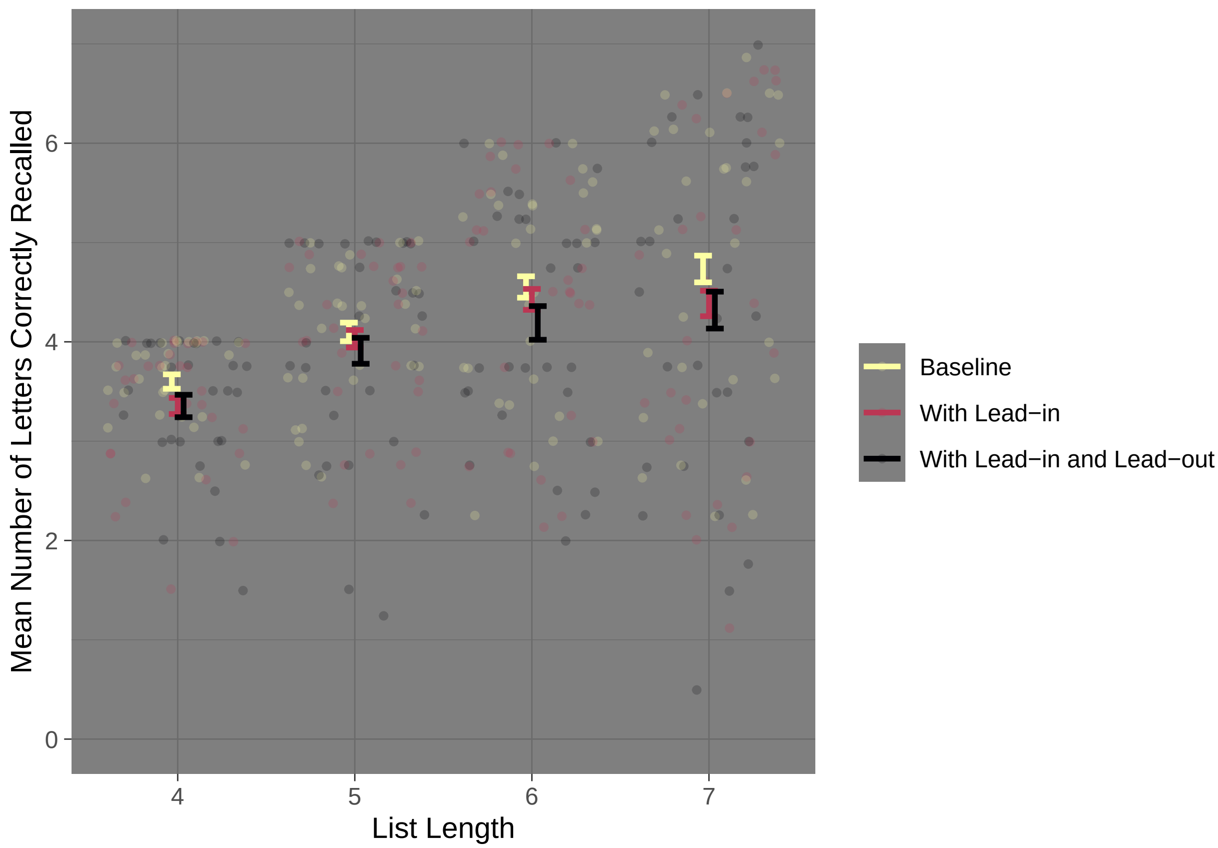

Figure 9. Mean number of items recalled per list length during baseline memory trials and during complex span trials with and without lead-in tone judgments, Experiment 3b. Error bars mark within-participant standard errors around the mean calculated with the Cousineau-Morey method (R. D. Morey, 2008). Individual data points are participant-average number of letters recalled

described previously, resulting in the exclusion of 10.23% of otherwise valid responses. Again, processing judgment response times decreased after the first lead-in judgment, increased substantially when the first memory item was introduced, and continued increasing as the list accumulated. After list presentation was complete, judgment times decreased precipitously, but not all the way back to the lead-in baseline speed. We ran the 3-way Bayesian ANOVA described in Experiment 3a on tone judgment responses from Experiment 3b. The best model included main effects of processing cluster and list length,

Table 6

Processing task performance as a function of complex span condition and memory recall, Experiment 3b.

Condition	Listwise Recall	Mean Accuracy	SD	Mean RT	SD RT
With Lead-in	Incorrect	0.98	0.02	0.70	0.21
	Perfect	0.99	0.02	0.63	0.19
With Lead-in and Lead-out	Incorrect	0.98	0.02	0.66	0.21
	Perfect	0.98	0.04	0.64	0.20

Note. N=29. RT = response time. SD = standard deviations.

$BF = 9.4 \cdot 10^{255}$. Including list length was favored by a factor of 9.09; including or excluding all other terms was favored by at least as much. The list length effect reflects a trend for slower tone judgment responses with longer list lengths (4 items: $M = 0.57$; 5 items: $M = 0.58$; 6 items: $M = 0.62$; 7 items: $M = 0.62$). Follow-up tests investigating differences between response times in each of the three processing clusters favored the model distinguishing all three processing clusters by a factor of at least $3.7 \cdot 10^{13}$. Judgments were slowest in the memory load cluster ($M = 0.69$, $SD = 0.20$), fastest in the lead-in cluster ($M = 0.49$, $SD = 0.13$), and intermediate in the lead-out cluster ($M = 0.57$, $SD = 0.15$). We followed this test with fine-grained t-tests comparing critical transitions across a trial. As in previous experiments, judgment times after the first memory item was introduced were convincingly slower than those immediately prior, $BF = 4.8 \cdot 10^{58}$; see Figure 10 for descriptive values. In this sample, lead-out judgments never reached the speed of the last lead-in judgment, $BF = 4.8 \cdot 10^{7}$. We compared lead-out judgment RTs with the last lead-out value to determine where the lead-out values plateaued. From the fourth lead-out judgment onwards, lead-out judgments appeared stable, $BF = 0.79$. Lead-out

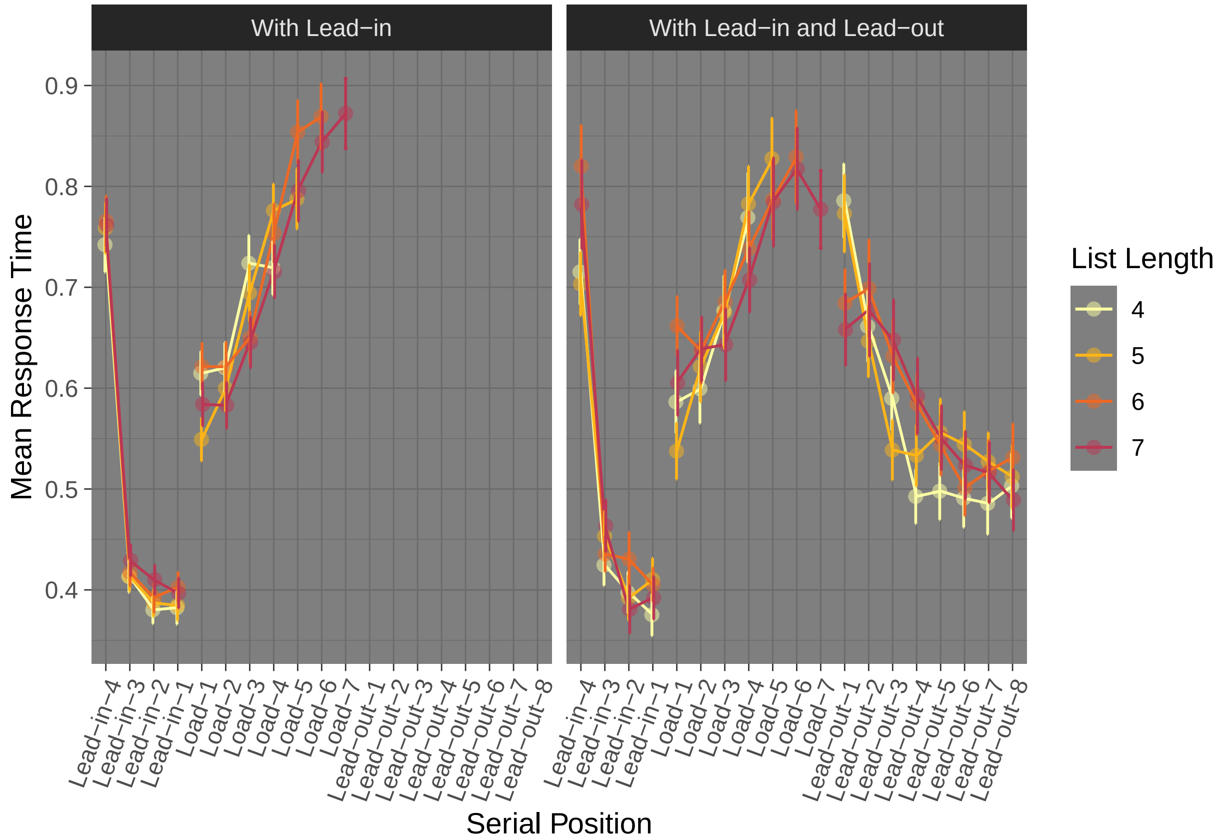

Figure 10. Mean processing response times (in seconds, means normalized) for complex span trials with only lead-in processing judgments (left) or both lead-in and lead-out judgments (right), by list length and serial position, Experiment 3b. Error bars are within-participant standard errors of the mean (R. D. Morey, 2008).

judgments did become faster than the first memory-loaded judgments, $BF = 1.2 \cdot 10^5$.

We conducted a Bayesian ANOVA on lead-out responses only in order to check for influences of list length. Here the best model included only a main effect of serial position, $BF = 1.2 \cdot 10^{28}$. Excluding list length (or its interaction with position) was favored by a factor of at least 1,190.78.

Recall dynamics: Experiments 3a and 3b. Recall per serial position for each lead-out condition and list length are shown in Figure 11. The best model according to a Bayes factor ANOVA ($BF = 2.7 \cdot 10^{243}$), included main effects of lead-in condition, list

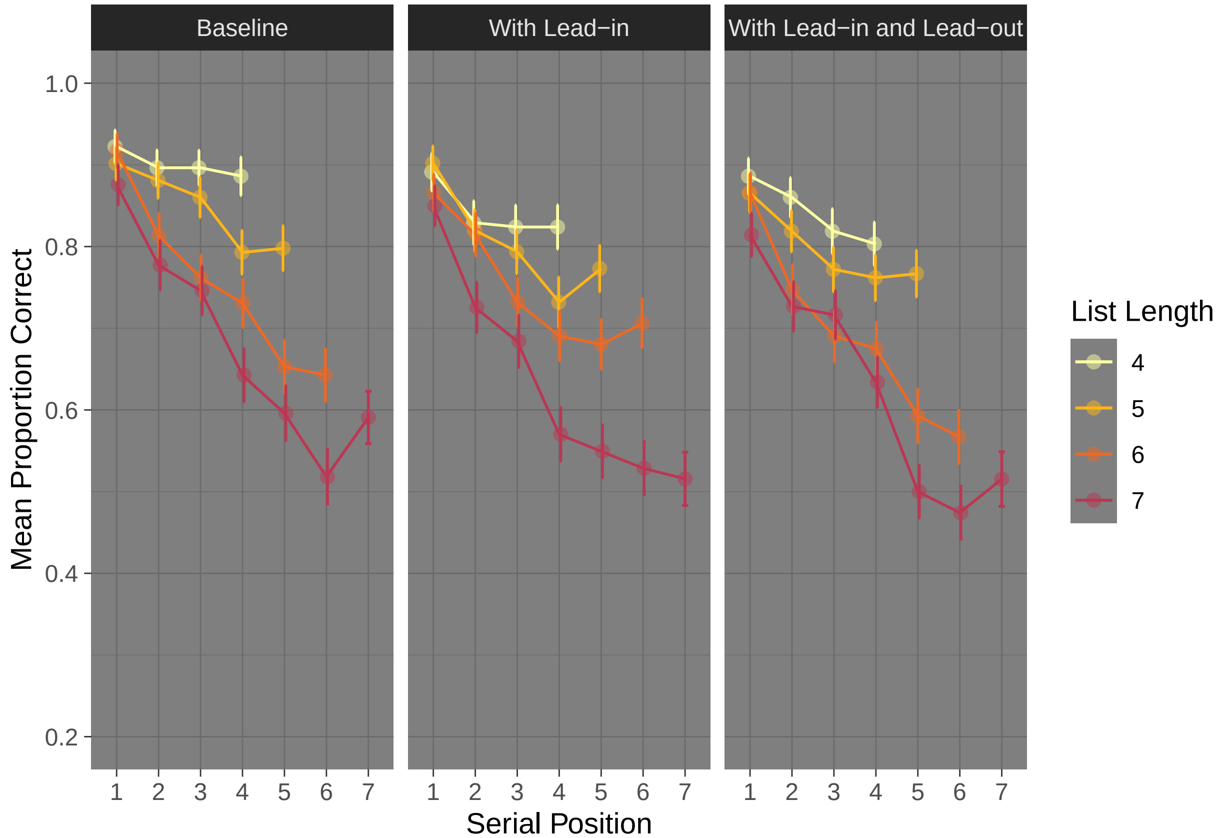

Figure 11. Mean recall accuracy for baseline and complex span trials in Experiments 3a and 3b with and without lead-in processing judgments, by list length and serial position. Error bars are within-participant standard errors of the mean (R. D. Morey, 2008).

length, and serial position, plus an interaction between list length and serial position. Inclusion of lead-in condition was favored by a factor of 1,530.13; inclusion of other terms was favored by even larger margins. Though the effect of condition was present and in the expected direction, it was rather small: recall is numerically superior in the baseline condition ($M = 0.78$, $SD = 0.26$) compared to the complex span conditions (With Lead-in: $M = 0.74$, $SD = 0.28$; With Lead-in and Lead-out: $M = 0.72$, $SD = 0.29$). We compared simplified models with 2-level codings of the condition variable (one in which baseline differed from the two complex span conditions, and one in which the lead-out condition differed from the other two conditions) to test whether all three levels of lead-in condition differed from each other. The simplified version in which the baseline trials were considered to differ from both types of complex span trial (which were not considered different from each other) was favored by a factor of 4.12 over the model distinguishing all three lead-in conditions. This confirms that recall in the complex span conditions was poorer than in the baseline condition. However, there is no reason to think that recall in the lead-out condition differed from the lead-in-only condition.

Discussion

With Experiments 3a and 3b, we again replicated the approximately linear slowing of tone judgments as a to-be-rememebered list of letters accumulates, along with a substantial cost to judgment response times with the introduction of the first item in the list. We removed verbal labels from the tone response options in Experiment 3b, and observed the same patterns as in each other experiment. This supports our contention that the conflict between letter memory and tone classification was not likely driven by any competing need to represent verbal labels in each task.

To gain clarity on how retaining a memory list and performing a judgment task may occur concurrently, in Experiments 3a and 3b we increased the number of tone judgments in the lead-out period from 4 to 8. With typical tone judgment times ranging from 0.4 - 0.8 seconds, each cycle of tone presentation and response lasted ~ 1.9 - 2.5 seconds, meaning that the time from the end of the list presentation in the lead-out conditions of Experiments 3a and 3b doubled to ~16 - 20 seconds in comparison with Experiments 2a and 2b. This increase in the lead-out period allowed us to observe tone judgment speeds clearly falling back below the response speeds incurred at the start of the memory list.

The main reason for conducting Experiments 3a and 3b was to test whether we would observe a clearer cost to memory recall during complex span trials compared with memory-only baseline trials if we randomly spread baseline trials across the session.

Indeed, when the baseline and complex span trials were interspersed, we found that recall during baseline exceeded recall during complex span, indicating that performing the tone judgments incurred a cost to memory (on top of the clear slowing observed to tone judgments with versus without memory load). However, the cost to recall was rather small, and we did not observe any difference in recall between the complex span trials with lead-in judgments and the trials with both lead-in and lead-out judgments. Under the assumptions of the TBRS model, these results are surprising. The speeding tone judgment responses during the lead-out period could be taken to indicate that participants are not continuously refreshing the memory list once presentation is complete. If so, one might also have expected that recall performance during the lead-out condition would be impaired, but there was no evidence for any difference between trials with versus without the lead-out judgments. This pattern is consistent with the assumption that reconfiguring the memoranda somehow, possibly in preparation for responding, is the process provoking general conflict, rather than maintenance per se.

General Discussion

Across five experiments, we observed that simple judgments slowed when intermixed with presentation of to-be-remembered letters in a complex working memory span paradigm. When a series of lead-in judgments was presented before any memoranda, we consistently observed substantial slowing after introduction of the first memory item, and further slowing with the introduction of each new memory item. Without these lead-in judgments, we observed a non-monotonic pattern of judgment speeds similar to those observed by C. C. Morey et al. (2018). The lead-in judgments evidently smooth this pattern, making it clear that response times increase as the memory load accumulates. We think that this finding resolves the mixed evidence in previous literature. Much of the prior evidence was presented as differences between judgments associated with the first and final list items; this evidence may have missed the non-monotonicity that makes the difference score uninterpretable. Going forward, researchers using complex span should expect to see the nonlinear pattern in judgment response times that we have documented. If researchers want to examine judgment responses as a function of serial position, they could introduce lead-in judgments to ensure that any apparent effects of memory load on judgments are interpretable.

We consistently observed a substantial start-of-list response time cost to judgments. This cost, apparent in all of our response time figures (compare Lead-in-1 with Load-1 values; typically 200-300 ms), was much larger than adding subsequent single items onto a list (see slopes from Load-1 to Load-N); the cost of adding one more item was usually tens of milliseconds, not hundreds of milliseconds. This suggests that the "load" effect on judgment response times is not merely mnemonic. If it were, then the slowing associated with the introduction of the first memory item after the lead-in judgments would be more in line with the slowing observed for adding another item to the list. While this difference in cost for the starting item compared to subsequent items is difficult to explain by appealing to an item-based memory load, it can be more readily explained if we suppose that from the introduction of the first memory item, two task sets must be juggled (Altmann, 2002). With the onset of the first memory item, participants must re-activate the serial recall task set and begin configuring their eventual response while also anticipating and executing their responses to the tone judgment task. We interpret the increased slowing as list length accumulates to configuration of an increasingly long response. This interpretation is consistent with the proposition that a response bottleneck (Pashler, 1992) shared between judging the tone stimuli and retaining the memory list must be negotiated, but this implies that "retaining" the memory list is synonymous with planning the intended response. We will consider additional possibilities for what could be taking place during presentation of memoranda below. We think the patterns that we have documented are important because they demonstrate that considering complex span in terms of task switches and the re-activation of task sets can perhaps explain how the components of the complex working memory span task interact better than assumptions about per-item memory load.

From Experiment 2, we introduced a series of lead-out judgments between presentation of the final memory item and the prompt to recall. If the memoranda must be refreshed to prevent decay, and if refreshing and judgment both require the focus of attention, then judgments occurring during a post-list delay should remain slow. At minimum, they should remain slower than the judgments occurring with the first list item, but possibly, they could remain as slow as judgments occurring with the final memory items. However, we observed rapid speeding of lead-out judgments, which agrees with some previous findings (e.g., Klapp, Marshburn, & Lester, 1983; Oberauer, 2002). After presentation of the last memory item judgments became faster until they were faster than the judgments occurring during presentation of the list. This speeding is not predicted by the TBRS model. It is more consistent with the idea that conflict during list accumulation is due to response configuration, specifically switching between the recall and discrimination task response sets where the recall response is gradually becoming more complex. Because response configuration would be complete at some point shortly after list presentation finished, it would presumably incur little further incremental task mixing cost (Poljac, Koch, & Bekkering, 2009), which could explain the speeding of responses during the lead-out judgments. However, the TBRS model could potentially account for this pattern by supposing that participants opted to verbally rehearse the lists after presentation was complete. Unlike attentional refreshing, verbal rehearsal would not necessarily provoke any cost to a non-verbal secondary task; verbal rehearsal is believed to operate independently of attentional refreshing (Camos, 2015; Camos, Lagner, & Barrouillet, 2009). It is believed that after rehearsal has been initiated, it can take place without much cost to a concurrent task (Naveh-Benjamin & Jonides, 1984). Given our results, namely that concurrent judgment response times were consistent with refreshing during but not after list presentation, one might surmise that participants used attentional refreshing to maintain the letters during list presentation, but switched to rehearsal after the list was complete. While it is plausible to suggest that the speeding of judgment RTs after the list ended reflects a switch of maintenance processes, the result nonetheless remains awkward to interpret if we depend on memory load rather than task switching phenomena. According to Naveh-Benjamin and Jonides' work, it is necessary to rehearse the entire list a few times before the slowing on a concurrent task diminishes (but see also Thalmann, Souza, and Oberauer (2019), who cast doubt on whether rehearsal ever becomes cost-free). Our results suggest that any load on the judgment task started diminishing in a much shorter period than would be needed to rehearse the memory list a few times. Moreover, we never observed an interaction involving list length in the tone judgment response time analyses. Rehearsal of short lists could be completed faster than rehearsal of long lists, so one might expect that lead-out judgments would speed faster for shorter lists after the participant switched from refreshing to rehearsal. However, we saw no evidence supporting this contention. Overall, we do not think that assuming a shift from refreshing to rehearsal provides a satisfactory explanation of these results, but additional research may be required to persuade the most skeptical readers. However, designing a conclusive experiment to test this hypothesis would be difficult without consensus on *when* rehearsal may take place without any concurrent cost.

One might also suppose that refreshing is needed frequently while the memory list is accumulating, but that after presentation has finished, participants may refresh differently; perhaps some memory items are grouped together and may be refreshed simultaneously (perhaps even list-wise). If so, then one might expect that refreshing and judgments might co-occur with more ease and less interference during a post-list retention period. However, while we think this explanation could account for why judgments did not remain as slow during lead-out period as they were during presentation of the final list items, we do not think that judgments should have become faster than they were when participants maintained only 1-2 items, as at the beginning of the list presentation. While the speeding of lead-out judgments are difficult, though not impossible, to square with the notion that attention is needed to prevent decay of memoranda, these patterns are perfectly consistent with the idea that the conflict with a secondary task during list accumulation occurs due to the establishment of two task sets and intermittent switching between reconfiguring the to-be-recalled response and performing the tone discrimination. Once the recall task set no longer requires updating (i.e., during the lead-out phase), these intermittent switches no longer occur and discrimination responses speed accordingly.

One puzzling finding in Experiments 1 and 2 was that we observed quite small impairments to recall in complex span compared with baseline recall in the conditions with lead-in judgments. This outcome is surprising if we assume that interference occurs *because* of a need to actively maintain information while performing the judgments. Whether or not extra judgments took place prior to introduction of the memoranda should not influence how much the concurrent judgments disrupt memory. In Experiments 3a and 3b, we provided a stronger test of whether recall was impaired on complex span trials compared to an uninterrupted baseline by mixing baseline serial recall trials with the complex span trials and closely matching the timings of retention intervals with and without lead-out judgments. We confirmed that recall was impaired in complex span compared to baseline, but also found no difference between trials with and without lead-out judgments. Overall, we think this is more consistent with the notion that interference between the memory and judgment task reflects task switching that occurs while the recall response is configured, rather than attentional refreshing during the delay period. One might have assumed that if refreshing does not occur (or occurs less frequently) during the delay, the memory list would be at risk of decay, particularly during the long lead-out periods of Experiments 3a and 3b. However, results show that participants recalled about as much with lead-out judgments (which presumably would have disrupted refreshing to some degree) as without them. This finding poses no difficulty for the assumption that the tone judgment task conflicted with reconfiguring the memoranda for responding. Assuming that judgments and memory are in conflict only while the response is being prepared would also explain why Vergauwe, Camos, and Barrouillet (2014) observed large effects of cognitive load only on the first response (i.e., the one nearest to the end of list presentation, during which response preparation might still have been underway) of a series of judgments imposed during the delay between list presentation and recall in a Brown-Peterson paradigm.

We did not explicitly manipulate any factor that allows us to conclude with certainty that it is translation of the memoranda for *responding* that provokes concurrent slowing in the tone judgment task, as opposed to transformation of the memory items in some other way. There are other transformative processes that could have taken place as our memory lists accumulated that might account for this pattern, for instance consolidation (e.g., Ricker, Nieuwenstein, Bayliss, & Barrouillet, 2018), or strategic decisions to elaborate (e.g., Bartsch, Singmann, & Oberauer, 2018). Each of these suggested transformations could plausibly explain the patterns we observed: each should require more attention when applied to longer series of items, and neither should require ongoing attention after the initial transformation finishes. However, while we cannot certainly rule out these possibilities, neither do they explain the patterns we observed better than supposing they occurred due to response reconfiguration. While any or all of these processes may have occurred during our task, we only know that the task required serial reconstruction of the lists. We therefore know that all participants must have accumulated responses, whereas we have no evidence about what else they might have done to boost memory that might also have contributed to the response time patterns we observed. Comparing these potential explanations offers a potential focus for future research.

As we concluded after Experiment 1, the multiple-component model (Baddeley, 2012; Logie, 2011) cannot adequately account for the conflicts we observed between retaining letters and judging tones. There is no obvious reason based on the multiple component model to predict the slowing that occurs in the processing task as the memory list accumulates; indeed, previous researchers have taken apparent absences of slowing as evidence in favor of multiple components (e.g., Maehara & Saito, 2007). The multiple component model might accommodate the start-of-list cost to judgment response time by appealing to a central executive. If we had observed judgment responses becoming faster during the lead-in period then slowing a constant amount during the memory list, this would have been consistent with the notion that a general attentional module such as the central executive coordinates switching between tone judgments and letter encoding. Of course, the multiple component model could likewise handle the speeding of judgments during the lead-out period. But the multiple-component model cannot clearly explain the linear slowing as memory load increases. According to the model, these letters would be loaded into a separate memory buffer, and should not themselves compete with the concurrent task (though there might be an overall slowing due to coordinating two tasks). A skeptical adherent to the multiple component model might suggest that tones required representation in the phonological loop, and this increased the interference we observed. However, the suggestion that non-verbal auditory material accesses the phonological loop and store would be in conflict with recent work proposing that nonverbal information is represented in a specialized tonal working memory module (Jordan, 2018; Schulze & Tillmann, 2013). We think it is clearly worth considering which functions might take place independently of others, but we maintain that it would be best to assume that domain-specific phenomena arise from specialized sensory and motor systems, rather than specialized short-term stores (C. C. Morey, Rhodes, & Cowan, 2019).

Though the clear conflicts we observed between a letter serial recall and tone judgment task could perhaps be explained with a unitary working memory model or with the TBRS model, we do not find the explanations arising from these fully compelling. We have already summarized the pitfalls for TBRS in explaining all of the patterns we observed. Any unitary working memory models assuming that attention is needed to hold information fare similarly. In our assessment, it remains possible to invoke TBRS, but we do not think the entire pattern of results strongly supports any model that assumes holding items in mind necessarily requires attention. We find the idea that conflict arises during the proactive reconfiguration of the memory response compelling, and fairly compatible in important respects with existing models of WM. This fairly new and influential idea (Myers, Stokes, & Nobre, 2017; Stokes, 2015) could potentially help to bridge the apparent need for both domain-specificity and generality in working memory. Further consideration of how the dynamics of preparing responses apply to memory tasks could explain perplexing findings that are inconsistent with both modular and unitary models of working memory. One such puzzle is why verbal memory interferes with visual memory tasks, but not the reverse (C. C. Morey, 2018; C. C. Morey & Mall, 2012; C. C. Morey & Miron, 2016; C. C. Morey, Morey, Reijden, & Holweg, 2013). Here, invoking reconfiguration as the source of conflict rather than a generalized storage resource is potentially a solution. Verbal responses may be proactively prepared via reconfiguration to speech for output. This reconfiguration might conflict with a simultaneous task, as we have apparently observed in the five new experiments reported here, even as this reconfiguration preserves the verbal information, possibly by co-opting an effector system (D. M. Jones & Macken, 2018) that is not implicated in the non-verbal task. In contrast, visual materials tend to be less directly convertible to motor output. While reconfiguration of visuo-spatial imagery in preparation for responding might likewise provoke general conflict, the response plan may not preserve visuo-spatial information as faithfully as articulation planning preserves verbal information, leading to the observed asymmetry.

In conclusion, the consistent patterns we report in these complex working memory span tasks accord well with the assumption that reconfiguring memoranda for eventual recall conflicts with execution of a simple decision task. This explanation of working memory processes can account for why judgments become slower when the memory stimuli commence, why judgments slow incrementally with the addition of subsequent memoranda, and why judgments become faster rapidly after the list presentation is finished. Because reconfiguration occurs regardless of whether memoranda are retained correctly, it can explain why memory and processing conflicts are observed regardless of eventual recall accuracy. Furthermore, the reconfiguration hypothesis does not require that we observe worse recall after a filled delay period compared to the same amount of free time, which suggests that persistent re-activation of memoranda across a several-second delay is not essential for preservation of the information. We think that incorporating reconfiguration, including likely differences between preparing verbal and non-verbal responses, into models of working memory offers a promising direction for balancing tensions between modular and unitary conceptions.

References

Abreu, P. M. J. E. de, Conway, A. R. A., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence*, *38*(6), 552–561. https://doi.org/10.1016/j.intell.2010.07.003

Altmann, E. M. (2002). Functional decay of memory for tasks. *Psychological Research*, *66*(4), 287–297. https://doi.org/10.1007/s00426-002-0102-9

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Barrouillet, P., & Camos, V. (2015). *Working memory: Loss and reconstruction.* Hove, U.K.: Psychology Press.

Barrouillet, P., Portrat, S., & Camos, V. (2011). On the law relating processing to storage in working memory. *Psychological Review*, *118*(2), 175–192. https://doi.org/10.1037/a0022324

Bartsch, L. M., Singmann, H., & Oberauer, K. (2018). The effects of refreshing and elaboration on working memory performance, and their contributions to long-term memory formation. *Memory & Cognition*, *46*(5), 796–808. https://doi.org/10.3758/s13421-018-0805-9

Camos, V. (2015). Storing verbal information in working memory. *Current Directions in Psychological Science*, *24*(6), 440–445. https://doi.org/10.1177/0963721415606630

Camos, V., Johnson, M., Loaiza, V., Portrat, S., Souza, A., & Vergauwe, E. (2018). What is attentional refreshing in working memory? *Annals of the New York Academy of Sciences*, *1424*(1), 19–32. https://doi.org/10.1111/nyas.13616

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, *61*(3), 457–469. https://doi.org/10.1016/j.jml.2009.06.002

Cowan, N. (2005). *Working memory capacity.* New York, NY: Psychology Press.

Cowan, N. (2019). Short-term memory based on activated long-term memory: A review in response to Norris (2017). *Psychological Bulletin*, *145*(8), 822–847. https://doi.org/10.1037/bul0000199

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Doherty, J. M., Belletier, C., Rhodes, S., Jaroslawska, A., Barrouillet, P., Camos, V., … Logie, R. H. (2019). Dual-task costs in working memory: An adversarial collaboration. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(9), 1529–1551. https://doi.org/10.1037/xlm0000668

Doherty, J. M., & Logie, R. H. (2016). Resource-sharing in multiple-component working memory. *Memory & Cognition*, *44*(8), 1157–1167. https://doi.org/10.3758/s13421-016-0626-7

Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 972–992. https://doi.org/10.1037/0278-7393.18.5.972

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271. https://doi.org/10.1037/a0027371

Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*(1), 136–158. https://doi.org/10.1016/j.jml.2004.03.008

Grange, J. (2015). Trimr: An Implementation of Common Response Time Trimming Methods. Retrieved from https://cran.r-project.org/web/packages/trimr/index.html

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge, UK: Cambridge University Press.

Jarrold, C., Tam, H., Baddeley, A. D., & Harvey, C. E. (2011). How does processing affect storage in working memory tasks? Evidence for both domain-general and domain-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 688–705. https://doi.org/10.1037/a0022527

Jones, D. M., & Macken, B. (2018). In the beginning was the deed: Verbal short-term memory as object-oriented action. *Current Directions in Psychological Science*, *27*(5), 351–356. https://doi.org/10.1177/0963721418765796

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition*, *144*, 1–13. https://doi.org/10.1016/j.cognition.2015.07.009

Jordan, C. (2018). Exploring a possible tonal loop in musicians and non-musicians and the relationship between musical expertise and cognitive ageing. Retrieved from https://era.ed.ac.uk/handle/1842/31077

Klapp, S. T., Marshburn, E. A., & Lester, P. T. (1983). Short-term memory does not involve the "working memory" of information processing: The demise of a common assumption. *Journal of Experimental Psychology*, *112*(2), 240–264. https://doi.org/10.1037/0096-3445.112.2.240

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking-An integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557–583. https://doi.org/10.1037/bul0000144

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79. https://doi.org/10.1162/jocn_a_00140

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Current Directions in Psychological Science*, *20*(4), 240–245. https://doi.org/10.1177/0963721411415340

Maehara, Y., & Saito, S. (2007). The relationship between processing and storage in working memory span: Not two sides of the same coin. *Journal of Memory and Language*, *56*(2), 212–228. https://doi.org/10.1016/j.jml.2006.07.009

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Morey, C. C. (2018). The case against specialized visual-spatial short-term memory. *Psychological Bulletin*, *144*(8), 849–883. https://doi.org/10.1037/bul0000155

Morey, C. C., Hadley, L. V., Buttelmann, F., Könen, T., Meaney, J.-A., Auyeung, B., … Chevalier, N. (2018). The effects of verbal and spatial memory load on children's processing speed. *Annals of the New York Academy of Sciences*, *1424*(1), 161–174. https://doi.org/10.1111/nyas.13653

Morey, C. C., & Mall, J. T. (2012). Cross-domain interference costs during concurrent verbal and spatial serial memory tasks are asymmetric. *Quarterly Journal of Experimental Psychology*, *65*(9), 1777–1797. https://doi.org/10.1080/17470218.2012.668555

Morey, C. C., & Miron, M. D. (2016). Spatial sequences, but not verbal sequences, are vulnerable to general interference during retention in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(12), 1907–1918. https://doi.org/10.1037/xlm0000280

Morey, C. C., Morey, R. D., Reijden, M. van der, & Holweg, M. (2013). Asymmetric cross-domain interference between two working memory tasks: Implications for models of working memory. *Journal of Memory and Language*, *69*(3), 324–348. https://doi.org/http://dx.doi.org/10.1016/j.jml.2013.04.004

Morey, C. C., Rhodes, S., & Cowan, N. (2019). Sensory-motor integration and brain lesions: Progress toward explaining domain-specific phenomena within domain-general working memory. *Cortex*, *112*, 149–161. https://doi.org/10.1016/j.cortex.2018.11.030

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial for Quantitative Methods in Psychology*, *4*, 61–64.

Morey, Richard D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor: Computation of Bayes Factors for Common Designs. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in Cognitive Sciences*, *21*(6), 449–461. https://doi.org/10.1016/j.tics.2017.03.010

Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 369–385. https://doi.org/10.1037/0278-7393.10.3.369

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2013). The focus of attention in working memory-from metaphors to mechanisms. *Frontiers in Human Neuroscience*, *7*, 673. https://doi.org/10.3389/fnhum.2013.00673

Oberauer, K., & Lewandowsky, S. (2016). Control of information in working memory: Encoding and removal of distractors in the complex-span paradigm. *Cognition*, *156*, 106–128. https://doi.org/10.1016/j.cognition.2016.08.007

Pashler, H. (1992). Attentional limitations in doing two tasks at the same time. *Current Directions in Psychological Science*, *1*, 44–48. https://doi.org/10.1111/1467-8721.ep11509734

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-01193-y

Poljac, E., Koch, I., & Bekkering, H. (2009). Dissociating restart cost and mixing cost in task switching. *Psychological Research PRPF*, *73*(3), 407–416. https://doi.org/10.1007/s00426-008-0151-9

Ricker, T. J., Nieuwenstein, M. R., Bayliss, D. M., & Barrouillet, P. (2018). Working memory consolidation: Insights from studies on attention and working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 8–18. https://doi.org/10.1111/nyas.13633

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Saito, S., & Miyake, A. (2004). On the nature of forgetting and the processing-storage relationship in reading span performance. *Journal of Memory* *and Language*, *50*(4), 425–443. https://doi.org/10.1016/j.jml.2003.12.003

Schönbrodt, F. D., & Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 1–15. https://doi.org/10.3758/s13423-017-1230-y

Schulze, K., & Tillmann, B. (2013). Working memory for pitch, timbre, and words. *Memory (Hove, England)*, *21*(3), 377–395. https://doi.org/10.1080/09658211.2012.731070

Siegel, L. S. (1994). Working memory and reading: A life-span perspective. *Journal of Behavioral Development*, *17*(1), 109–124. https://doi.org/10.1177/016502549401700107

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405. https://doi.org/10.1016/j.tics.2015.05.004

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). Revisiting the attentional demands of rehearsal in working-memory tasks. *Journal of Memory and Language*, *105*, 1–18. https://doi.org/10.1016/j.jml.2018.10.005

Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, *39*(2), 195–217. https://doi.org/10.1006/jmla.1998.2574

Towse, J. N., Hitch, G. J., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory & Cognition*, *28*(3), 341–348. https://doi.org/10.3758/BF03198549

Towse, J. N., Hitch, G. J., & Hutton, U. (2002). On the nature of the relationship between processing activity and item retention in children. *Journal of Experimental Child Psychology*, *82*(2), 156–184.

https://doi.org/10.1016/S0022-0965(02)00003-6

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626.

Vergauwe, E., Camos, V., & Barrouillet, P. (2014). The impact of storage on processing: How is information maintained in working memory? *Journal of Experimental Psychology-Learning Memory and Cognition*, *40*(4), 1072–1095. https://doi.org/10.1037/a0035779

Ward, G., & Tan, L. (2019). Control processes in short-term storage: Retrieval strategies in immediate recall depend upon the number of words to be recalled. *Memory & Cognition*, *47*(4), 658–682. https://doi.org/10.3758/s13421-018-0891-8