The Long-term Consequences of Retrieval Demands During Working Memory

Vanessa M. Loaiza

Charlotte Doherty

Paul Howlett

This is a pre-print of an article accepted for publication at *Memory & Cognition* on 26 July 2020.

The final authenticated version is available online at https://doi.org/10.3758/s13421-020-01079-5

Word count: 8,496/8,500 (including main text and footnotes only)

Author Note

Vanessa M. Loaiza, Department of Psychology, University of Essex. Charlotte Doherty, Department of Psychology, University of Essex. Paul Howlett, Department of Psychology, University of Essex.

Correspondence concerning this article should be addressed to Vanessa M. Loaiza, Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. Email: v.loaiza@essex.ac.uk

Abstract

Word count: 246/250

Although it is well known that distraction impairs immediate retrieval of items maintained in working memory (WM, e.g., during complex span tasks), some evidence suggests that these items are more likely to be recalled from episodic memory (EM) compared to items that were studied without any distraction (e.g., during simple span tasks). One account for this delayed advantage of complex span over simple span, or the McCabe effect (McCabe, 2008), is that complex span affords covert retrieval opportunities that facilitate later retrieval from EM by cumulatively reactivating each successively presented item after distraction. This explanation is focused on the processing that occurs during presentation and maintenance of the items, but no work to date has explored whether the differential demands of immediate retrieval between simple and complex span may explain the effect. Accordingly, these experiments examined the impact of immediate retrieval demands on the McCabe effect by comparing typical immediate serial-recall instructions (i.e., recalling the words in their exact order of presentation) to immediate free-recall (Experiments 1-2) and no-recall (Experiments 2 and 3) instructions. The results suggested that the nature of retrieval may constrain the McCabe effect in some situations (Experiments 1 and 2), but its demands do not drive the McCabe effect given that it was observed in both serial-recall and no-recall conditions (Experiment 3). Instead, activities such as covert retrieval during the processing phase may underlie the McCabe effect, thus further evidencing the importance of processing in WM for the long-term retention of information.

Keywords: complex span, simple span, working memory, episodic memory, retrieval

The Long-term Consequences of Retrieval Demands During Working Memory

There is a long tradition of research investigating the factors that promote long-term retention in episodic memory (EM), the memory system widely agreed to reflect the permanent storage of personally experienced events and information (Atkinson & Shiffrin, 1968; Craik & Tulving, 1975; Johnson, 1992). Much recent work has considered the role of processes that support the online maintenance, manipulation, and updating of this information in working memory (WM), given that these processes may impact long-term retention as well (Bartsch et al., 2018; Camos & Portrat, 2015; Jarjat et al., 2018; Loaiza & Halse, 2019; Loaiza & McCabe, 2012; McCabe, 2008; Rose et al., 2014; Souza & Oberauer, 2017).

Much of this interest has centered on the processes that occur during the periods of presentation and maintenance of the to-be-remembered memoranda. For example, during complex span tasks, a typical measure of WM (Conway et al., 2005), several memoranda (e.g., words) are interspersed with distracting secondary processing components (e.g., arithmetic problems). WM capacity refers to the maximum number of items that can be accurately held in mind, which is often measured by serial recall of the items studied during complex span in their original order of presentation. Researchers are often interested in the underlying mechanisms that allow participants to hold these items in mind, despite the distraction (e.g., Barrouillet et al., 2011; Oberauer et al., 2012). Henceforth, we refer to the presentation and maintenance of memoranda during WM tasks as the *processing* phase versus the eventual *retrieval* phase wherein participants must attempt to recover the memoranda at the end of the trial.

An arguably disproportionate interest in the processing phase compared to the retrieval phase exists, as the retrieval phase is often considered to simply reflect the output of the operations taking place during the processing phase. Indeed, there are some theoretical views suggesting that WM capacity does not represent memory per se so much as the output of critical underlying processes that allow one to hold information in mind, such as the control of attention in the face of interference (Engle & Kane, 2004; Hasher et al., 2007). Other theories have focused on how attention keeps information active during the

processing phase, such as by reconstructing decaying memory traces (Barrouillet & Camos, 2015), by reactivating traces via searching the content of WM (Vergauwe et al., 2016; Vergauwe & Cowan, 2015), or by reinforcing bindings between memoranda and their contexts (Loaiza et al., 2015; Loaiza & McCabe, 2012; Loaiza & Souza, 2018, 2019). Regardless of the specific explanation, most theories focus predominantly on the underlying processes that support ongoing encoding and maintenance in WM, with retrieval often merely serving as an indication of their functioning rather than an interest in and of itself.

However, there is growing acknowledgement that retrieval from WM should not be taken for granted as simply a byproduct, but could moderate the impact of the purported underlying processes supporting encoding and maintenance in WM. For example, recent work has demonstrated that the method of retrieval (i.e., recall versus recognition) of visuospatial and auditory-verbal items can modify the extent to which cross-domain interference is evident in WM (Uittenhove et al., 2019). Such findings have profound implications for major theoretical debates, such as whether WM is more domain-specific or domain-general (Fougnie, Zughni, Godwin, & Marois, 2015; Logie, 2011; Morey, 2018; Rhodes et al., 2019). In a similar vein, Pratte (2020) demonstrated that retrieval limitations are more likely responsible for the detrimental effect of increasing the number of memoranda on precision of visual WM, rather than the more typical explanations concerning encoding or storage limitations. Analogous to the previous example, such results are pertinent to the theoretical debate regarding whether limitations in visual WM capacity are best understood as discrete slots or flexibly-allocated resources (Bays & Husain, 2008; Luck & Vogel, 1997; Zhang & Luck, 2008). Thus, more and more instances in the field suggest that retrieval from WM should be more frequently considered in researchers' theorizing.

Some theoretical views of WM have blurred the tacit boundary between the processing and retrieval phases by including retrieval from outside immediate awareness as a critical element to WM functioning. According to Unsworth and Engle's (2007) primary-secondary memory framework, WM capacity reflects a combination of active maintenance in primary memory and retrieval from secondary

memory. Primary memory keeps active and accessible about four distinct representations, and if this capacity limit is exceeded, whether by distraction (e.g., from the processing component of complex span tasks) or by presenting more items, then retrieval from secondary memory must occur to recover the displaced items. Thus, Unsworth and Engle's model advances the notion that retrieval from outside the central component of WM occurs during common measures of WM capacity, such as complex span tasks.

Following Unsworth and Engle's rationale and the conceptualization of WM as a central subset of active representations of long-term memory (Cowan, 1999; Oberauer, 2002), McCabe (2008) developed the covert retrieval model to specify how the processing phase may include retrieval from outside the central component of WM. Similar to previous work regarding maintenance operations during the processing phase of complex span tasks (e.g., Barrouillet et al., 2004, 2007), McCabe asserted that participants use the remaining free time after distraction to successively and cumulatively covertly retrieve the displaced memoranda back into the central component of WM. Given their structure, tasks like complex span afford these repeated internal retrieval practice opportunities, whereas simple span tasks (e.g., word span), presenting only a few memoranda without any distraction, do not require covert retrieval, as none of the items would have been displaced. Thus, although detrimental to immediate recall, McCabe asserted that these brief distracting tasks provide an opportunity to strengthen retrieval cues that could be later used during retrieval from EM. The evidence for this notion was demonstrated through a relative advantage in delayed free recall (DFR) of items processed during complex span over simple span. This long-term advantage of complex span over simple span, or the *McCabe effect*, also has been demonstrated in cued recall (Loaiza & McCabe, 2012) and recognition (Loaiza et al., 2015).

Much like aforementioned work, the covert retrieval model is focused on the processing phase during which covert retrieval purportedly occurs and less concerned with overt retrieval from WM. In the original study, McCabe conducted an experiment to ensure that the advantage of complex span over simple span was specific to the processing phase rather than the retrieval phase. In contrast to the covert

retrieval account, an alternative *immediate retrieval demands* account would suggest that the differential immediate recall demands between simple and complex span drive the McCabe effect. That is, serial recall of memoranda presented during complex span is much more challenging than simple span, and overcoming this relatively demanding overt retrieval may be what promotes long-term retention rather than any presumed covert retrieval. To adjudicate between these explanations, McCabe administered trials of simple and complex span that unpredictably ended with either a cue to immediately recall the items in serial order, as usual, or an unrelated task that precluded immediate recall. Although DFR was unsurprisingly lower overall for the no-recall versus serial-recall trials, owing to an overall effect of retrieval practice (Rowland, 2014), the McCabe effect was evident for both recall instructions. Thus, the actual act of overt retrieval had no impact on the long-term advantage of complex span over simple span.

Since this original paper and subsequent work (e.g., Abadie & Camos, 2018; Camos & Portrat, 2015; Jarjat et al., 2018; Souza & Oberauer, 2017), there has been little attention paid to whether differential retrieval demands may moderate long-term retention of information in WM. This is quite surprising given that complex span tasks typically require serial recall, whereas participants freely recall the items after a delay. This basic methodological mismatch in the retrieval instructions deserves attention in addition to the fact that retrieval methods and limitations are increasingly considered consequential for long-standing debates in WM (Pratte, 2020; Uittenhove et al., 2019). Besides McCabe's aforementioned experiment, there is some indication that the retrieval demands should not play a role in this context. For example, Loaiza and Borovanska (2018) replicated the finding that immediate recall did not moderate the McCabe effect in memory of different characteristics (phonological, semantic, temporal-contextual) of the studied items. Loaiza and colleagues (2011) have further shown that the improved immediate and delayed recall due to a deep, semantic level-of-processing was the same regardless of immediate serial-recall or free-recall instructions. Hartshorne and Makovski's (2019) meta-analysis also demonstrated that the impact of WM maintenance on EM was consistent regardless of

whether immediate recall from WM was involved. Further work has suggested that immediate serial and free recall are more similar in nature than not (e.g., Ward et al., 2010), and thus immediate-recall instructions may have little impact if they are supported by the same underlying mechanism.

On the other hand, the notion of desirable difficulties (Bjork, 1994) would suggest that effortful retrieval, such as during complex span, should only increase the likelihood of retaining that information over the long-term compared to easier tasks, such as simple span. Indeed, prior work has shown that the McCabe effect is larger for DFR correcting for accurate immediate recall compared to overall DFR (Loaiza & Halse, 2019; Souza & Oberauer, 2017). This may indicate that successfully engaging in difficult retrieval conditions is most important to long-term retention, over and above the impact of any processing phase activities. That is, WM may be important to EM not because of the operations underlying the processing phase, but rather the effortful operations to retrieve information from WM. If so, varying immediate overt retrieval demands should likewise vary the McCabe effect, regardless of any manipulation of the processing phase designed to vary the opportunity for covert retrieval (e.g., complex versus simple span).

In the current experiments, we investigated the relative contributions of covert retrieval during the processing phase versus effortful overt retrieval during the immediate recall phase to the long-term retention of information originally studied and maintained in WM. Like our previous work (Loaiza et al., 2015; Loaiza & McCabe, 2012; McCabe, 2008), participants studied four words presented in trials of simple span (i.e., word span) and complex span (i.e., operation span) for immediate and delayed recall. Novel to this work, during immediate recall, participants recalled the words from a set of eight possible choices: the four presented words and four never-presented lures (i.e., reconstruction; Bartsch et al., 2018; Oberauer, 2019). Rather than the more common method of self-generated recall, reconstruction provided a better opportunity for participants to comply with their immediate retrieval instructions and minimized the possibility that low levels of immediate recall could cause baseline differences between simple and complex span (Loaiza & Halse, 2019; Rose et al., 2014). In Experiment 1, participants were randomly

prompted to reconstruct the words either in their original order of presentation (henceforth, *serial recall*) or to freely recall the items in any order (henceforth, *free recall*) to vary the immediate retrieval demands of the task. Experiment 2 included a further *no-recall* condition, as in McCabe (2008; Experiment 3). Experiment 3 considered whether participants approach the task differently depending on the proportion of serial-recall versus no-recall trials they expect, while also matching the retrieval method (i.e., reconstruction) between the immediate and delayed tests.[1] Thus, the consistent manipulation of task type alongside the different manipulations of immediate retrieval demands allowed us to investigate whether the difficulty of overt retrieval moderates the McCabe effect.

According to the covert retrieval account, we should observe an advantage of complex span over simple span at delay (i.e., a McCabe effect) regardless of the difficulty of immediate recall. This is expected because what drives the long-term advantage should be the internal, cumulative retrieval practice participants engage in during the processing phase of complex span, with the actual act of overt retrieval having little role. Conversely, the immediate retrieval demands account would predict a McCabe effect only when participants are instructed to serially recall the items, and not during the free-recall or no-recall conditions. That is, if the differential demands of overt retrieval from WM promote long-term retention, rather than any covert retrieval activities during the processing phase, then reducing the demands through free-recall or no-recall conditions should likewise diminish retrieval from EM.

Experiment 1

Method

Participants. In Experiments 1 and 2, we aimed to collect data from at least 24 participants based on similar prior research using the same sample size (Loaiza & Borovanska, 2018; McCabe, 2008; Souza & Oberauer, 2017). Twenty-four participants (M_{age} = 19.38 years, *SD* = 1.47 years) were recruited to

[1] We conducted two additional experiments that, in hindsight, were not particularly effective for addressing our research question, and thus we have reported them in the Supplementary Materials to preserve transparency.

volunteer from the University of Essex subject pool in exchange for course credit. Participants in all experiments provided informed consent before beginning and were fully debriefed at the conclusion of the experiments. The University of Essex ethics committee approved the ethics application for the experiments. Participants in each experiment were unique and did not participate in any other experiment in the series.

Materials and Procedure. The memoranda for Experiment 1 were randomly sampled without replacement from a list of 154 concrete, high-frequency nouns (letters: *M* = 5.35, *SD* = 1.29, range = 4-8; syllables: *M* = 1.47, *SD* = 0.50, range = 1-2; log HAL frequency: *M* = 9.29, *SD* = 0.96, range = 8.00-12.42) acquired from the English Lexicon database (Balota et al., 2007). A similar list of 224 words was developed for Experiments 2 and 3 given the increased number of items required for the design (letters: *M* = 5.43, *SD* = 1.11, range = 4-8; syllables: *M* = 1.55, *SD* = 0.50, range = 1-2; log HAL frequency: *M* = 9.22, *SD* = 1.03, range = 7.45-12.42). The words were randomly arranged for each participant. Experiments 1 and 2 and were programmed in Matlab with the Psychtoolbox extensions (Brainard, 1997; Kleiner et al., 2007).

Participants completed the experiment individually in quiet testing booths with an experimenter present for the duration of the experiment to ensure understanding and compliance with instructions. Before beginning the critical portion of all the experiments, participants practiced 10 example arithmetic problems (e.g., three + five = nine?) that later served as the secondary processing component of the complex span task until they reached an 85% accuracy criterion. Participants also received several practice trials preceding the first block and summary instructions for the remaining blocks thereafter.

The critical phase consisted of two blocks, each comprising a WM phase where simple span and complex span trials were administered followed by a period of distraction and a DFR phase. During the WM phase, each trial began with a fixation cross at the center of the screen for 1 s. Thereafter, words were successively presented at the center of the screen for 1 s (with a 0.5 s interstimulus interval, ISI) during simple span, and during complex span, one arithmetic problem followed each presented word for

3.5 s (0.5 s ISI). Participants were instructed to read the words and arithmetic problems aloud, and to solve the arithmetic problems aloud by saying true or false and pressing a right- or left-hand key, respectively. At the end of each trial, the four presented words and four words that were new to the experiment were randomly arranged each within a 2 x 4 grid of frames on the screen. For half the trials of each task type, the boxes and words turned red with the word "SERIAL" presented above them, prompting participants to use the mouse to click on the presented words in their original order of presentation (i.e., serial recall). For the other half of the trials, the boxes and words turned green with the word "FREE" presented above them, prompting participants recall the words without regard to their original order of presentation (i.e., free recall). An intertrial interval of 2 s followed the selection of four items. Each block comprised eight trials, with the task type and recall instruction randomly and evenly implemented (i.e., two trials of each task/recall per block).

After completing WM phase, the participants silently completed an unrelated distraction task of multiplication problems (e.g., 7 x 6 = 42?) for 1 min. Finally, each block ended with DFR: Participants were instructed to freely recall as many of the words as they could from the previous block by typing them into the computer. Their responses were echoed back to them on screen. DFR was manually checked for spelling mistakes and corrected if not ambiguous (e.g., a common typo "reciept" was corrected to "receipt," but "horm" was not corrected because it could be corrected as "harm" or "horn").

Design. The independent variables of immediate recall instruction (serial and free recall) and task type (simple and complex span) were manipulated within-subjects. The dependent variables were immediate recall (serial and free scoring) and DFR. Serial scoring refers to recall scored as accurate in the correct serial position, whereas free scoring refers to recall scored as accurate regardless of original serial position. We had also planned to report DFR conditionalized on accurate immediate recall, but for the sake of brevity, these analyses can be found on the Open Science Framework (OSF). The use of reconstruction during immediate recall greatly reduced the typical advantage of simple span over complex

span, and so the pattern of results was consistent between the two types of DFR measures. We additionally report performance on the secondary processing component of the complex span task (accuracy and response times, RTs).

Data Analysis. The results of all the experiments were pre-processed and analyzed in R (R core team, 2017). Our initial analysis used the BayesFactor package (Morey & Rouder, 2015) with its default settings to conduct Bayesian analyses of variance (BANOVAs; Rouder et al., 2012) and Bayesian t-tests (Rouder et al., 2009) for specific comparisons (e.g., to follow-up predicted or observed interactions). Bayesian inferential statistics allow the comparison of the likelihood of the data under one model (e.g., an alternative model that assumes a difference between complex span and simple span, M_1) relative to that of another model (e.g., a null model that only includes a random effect of participant, M_0). The ratio of these likelihoods is the Bayes factor (BF), expressing the relative evidence for one model over the other (e.g., the strength of evidence for the alternative model over the null model, BF_{10}). BFs ranging from 1 to 3 indicate weak evidence in favor of the model in the numerator, whereas BFs between 10 and 100 indicate strong and decisive evidence. We also derived measures of effect size (with their 95% highest-density intervals, HDIs) using Bayesian Estimation Software (BEST; Kruschke, 2013), but for the sake of brevity we do not report these results and direct the interested reader to the OSF.

To complement these analyses, we used the brms package (Bürkner, 2018) to fit hierarchical Bayesian logistic mixed effects models to predict the likelihood of recalling an item (1 or 0) during the delayed test as a function of our fixed effects (i.e., task type and recall instruction) and including random effects and slopes of participant. Although not originally planned, this approach is analogous to using BANOVA, with the main benefit being that it allowed us to leverage the heterogeneity across participants and trials rather than aggregate across it. As will be clear later, this was particularly important in the cases where the results of the BANOVAs and/or Bayesian t-tests were ambiguous, potentially signaling

insufficient power to observe an effect. Rather than spend more time and resources collecting data from more participants, it seemed prudent to capitalize on the data we had already collected.

The brms package uses Stan (Stan Development Team, 2018) to estimate posterior distributions of parameter estimates (i.e., regression weights representing the effects of task type, recall instruction, and their interaction). We applied weakly informative Cauchy priors (with location 0 and scale 5) on the regression coefficients, intercept, and variance for all the models, following prior similar work (Bartsch et al., 2018). The posterior parameter estimates of all the models were sampled through four independent Markov chains, each comprising 2,000 iterations, with the first 1,000 warmup iterations excluded from analysis. We checked for convergence by visually inspecting the four chains and verifying that the $\hat{R}$ statistic was close to 1 for all parameters of all the fitted models. Posterior predictive checks also ensured appropriate model fit to the data. We inspected the 95% HDIs of the posterior estimates of each McCabe effect to draw inferences, with HDIs not overlapping with 0 considered credible. We applied this approach to only the delayed performance results as they were most important for our main hypotheses, but note that there were several other instances of ambiguity in other reported results. For the sake of brevity, we do not detail the results of these analyses hereafter except to report the estimates of crucial pairwise comparisons to the hypotheses, especially so that they may clarify any ambiguous results in the planned aggregate analyses. The interested reader can find the analyses and full results on the OSF.

Results and Discussion

We first assessed participants' performance on the secondary processing component of the complex span task (see Table 1). There was moderate evidence against a difference between serial and free recall instructions in terms of response accuracy (BF_{01} = 3.68) and RTs (BF_{01} = 3.81). Thus, participants responded similarly during the processing component regardless of the immediate recall instructions.

Next, we examined participants' likelihood to follow the recall instructions depending on the type of task using separate 2 (immediate recall instruction: serial, free) x 2 (task type: simple, complex) within-

subjects BANOVAs applied to immediate free and serial scoring (Table 2 and Figure 1). For serial scoring, there was clear evidence for the full model including an interaction between recall instruction and task type: Participants complied with the free recall instruction overall, leading to an ambiguous difference between task types (BF_{10} = 1.13), whereas serial recall still proved to be a challenge for complex span compared to simple span (BF_{10} = 4.36 x 10^5). The results of the free scoring indicated that participants were able to recall many of the items regardless of their order, although the best model including only an effect of task type suggests that there was still a disadvantage for complex span compared to simple span.

The most important results concerned DFR (Table 3 and Figure 1). We observed that the best model included only a main effect of task type, but this model was not substantially preferred to the next best full model including an interaction between recall instruction and task type. The specific comparisons revealed evidence for a McCabe effect in the serial-recall condition (BF_{10} = 7.86), but the effect was ambiguous in the free-recall condition (BF_{01} = 1.67). The pairwise comparisons of the posterior estimates from the hierarchical Bayesian logistic mixed effects model more firmly indicated a credible McCabe effect for the serial-recall condition (estimate = -0.92 [-1.56, -0.24]), but not for the free-recall condition (estimate = -0.42 [-1.08, 0.24]). These results conflict with the covert retrieval account and instead support the immediate retrieval demands account, such that the demands of serial recall may promote long-term retention of complex span items that lead to the McCabe effect.

An alternative explanation of these results is that the lack of a credible McCabe effect in the free-recall condition may have occurred because the act of free recall interferes with the cumulative covert retrieval that participants engage in during the processing phase. That is, it may not be that retrieval demands promote the McCabe effect so much as the free-recall instructions in Experiment 1 created a mismatch between the encoding processes (i.e., cumulative covert retrieval of the serially presented items during complex span) and the retrieval method (i.e., free recall; Morris et al., 1977).

To determine whether the null McCabe effect in free recall was due to reduced retrieval demands or to the mismatch between encoding and retrieval, we conducted Experiment 2 with an additional no-recall condition alongside the instructed serial-recall and free-recall conditions. Thus, participants were randomly prompted to either recall the memoranda in their original serial order, in a "free" order, or completed an unrelated task to preclude immediate recall. Including the no-recall condition allows for a more dramatic manipulation of immediate overt retrieval demands than the free-recall instruction that, as explained, may have introduced an encoding-retrieval mismatch. According to the covert retrieval account, a McCabe effect should be evident for the serial-recall and no-recall conditions. This would replicate McCabe (2008; Experiment 3) and provide clear evidence for the notion that, regardless of immediate retrieval, the same underlying process of cumulative covert retrieval supports the ongoing maintenance and consequent long-term retention of the memoranda. Additionally, there may be a null McCabe effect in the free-recall condition, consistent with the encoding-retrieval mismatch explanation of the results of Experiment 1. Conversely, the immediate retrieval demands account would predict a McCabe effect only in the serial-recall condition. This prediction follows the assumption that serial recall instills the overt retrieval demands that promote long-term retention, whereas the free- and no-recall conditions do not engender such demands and should therefore exhibit no McCabe effect.

Experiment 2

Method

Participants and Design. Twenty-nine participants (M_{age} = 19.28 years, *SD* = 0.75 years) were recruited in exchange for course credit. One additional participant was excluded from analysis due to experiment malfunction. The experiment followed a 3 (immediate recall instruction: serial, free, no recall) x 2 (task type: simple, complex) within-subjects design.

Materials and Procedure. The materials and procedure were similar to Experiment 1, except that we included a no-recall condition. There were four blocks each comprising six trials, one for each cell of

the design and randomly intermixed. The serial- and free-recall trials were the same as in Experiment 1. During the no-recall trials, the frames and double-digit numbers (e.g., 48, 63, 95) within them turned blue with the word "DIGITS" presented above them, prompting participants to select the four of eight possible numbers that were both even. Like the arithmetic problems, participants practiced this digit task for 10 trials to reach an 85% criterion prior to the critical phase of the experiment.

Results

We first checked that participants responded to the secondary processing component of the complex span task in a similar way, regardless of immediate-recall condition (see Table 1). The results indicated moderate evidence for a null effect of recall instructions on response accuracy (BF_{01} = 5.17) and RTs (BF_{01} = 8.64). Participants' performance on the no-recall digit task was also very high during both simple span trials (*M* = 1.00, *SD* = 0.00) and complex span trials (*M* = 1.00, *SD* = 0.01).

We next conducted separate two-way BANOVAs for the immediate recall (see Table 2 and Figure 1) and DFR measures (see Table 3 and Figure 1). For serial scoring, the best model included main effects of both recall instruction and task type, which was only ambiguously preferred to the next best model including an interaction term. Thus, participants still appeared to serially recall items more often during simple span than complex span, even when instructed to freely recall them. For free scoring, the results were more similar to those of Experiment 1, such that participants largely recalled many of the presented items, but were still disadvantaged for complex span compared to simple span.

We next turn to the DFR results that pertain to the critical hypotheses. The best model included only an effect of recall condition, signaling that recall was unsurprisingly worse in the no-recall condition compared to the serial-recall and free-recall conditions. This model was ambiguously preferred to the next best main effects model including an effect of task type. This ambiguity in the omnibus model comparisons may have been driven by an ambiguous McCabe effect in the free-recall condition (BF_{10} = 1.63), whereas there were no McCabe effects in the serial-recall (BF_{01} = 4.54) or no-recall (BF_{01} = 4.50) conditions. When

considering pairwise comparisons of the posterior estimates from the hierarchical Bayesian mixed effects model, there was no evidence for a McCabe effect in any of the recall conditions (no-recall estimate = -0.06 [-0.50, 0.41]; serial-recall estimate = -0.07 [-0.73, 0.59]; free-recall estimate = -0.40 [-0.82, 0.03]). The crucial lack of McCabe effects in the serial-recall and no-recall conditions conflict with the covert retrieval account. Note that we had pre-registered a further analysis of DFR across serial position, but given the lack of McCabe effects and in the interest of brevity, we have decided not to report these results. The interested reader can find the results for all the experiments on the OSF.

In summary, the results of Experiment 2 overall demonstrated that the variable retrieval demands nullified the McCabe effect in DFR, thereby negating the covert retrieval account that cumulative covert retrieval during processing phase promotes long-term retention of information studied in WM. However, the results do not perfectly align with the immediate retrieval demands account either given that no McCabe effect was observed in the serial-recall condition. As serial recall was required for only a third of the trials, it may be that participants took a reactive approach to the task: They may have been unlikely to engage in covert retrieval during the processing phase and instead simply respond to the retrieval demands when prompted, thereby nullifying the McCabe effect. Furthermore, delayed performance was very low overall, which could obfuscate any differences between the conditions. Relatedly, Experiments 1 and 2 did not address the aforementioned issue that the retrieval method is not consistent between immediate (i.e., reconstruction) and delayed tests (i.e., DFR), which could promote differences between retrieval of simple-span and complex-span items between the two times of test.

To address these issues, we designed Experiment 3 that was similar to the previous experiments, such that simple-span and complex-span trials ended unpredictably with serial or no recall. Importantly, however, participants were randomly assigned to one of three groups wherein they were informed before the task began whether there would be more serial-recall than no-recall trials (75%), fewer serial-recall than no-recall trials (25%), or an even split (50%). Furthermore, reconstruction was used to assess recall

during both the immediate and delayed tests. Given the relatively greater number of participants required for this mixed design and due to the suspension of in-lab testing during the coronavirus pandemic, we conducted this experiment online. We first conducted a control experiment with only serial recall from simple-span and complex-span trials to ensure that a McCabe effect can be demonstrated online and when matching the retrieval method (i.e., reconstruction) between the times of test.

We predicted a large McCabe effect for both serial- and no-recall trials when most of the trials of the block (i.e., 75%) require serial recall, a still sizable effect when the trials are evenly split (replicating McCabe, 2008, Experiment 3), and a smaller or null effect when there are fewer serial-recall than no-recall trials (i.e., 25%). That is, participants may change their approach to the task depending on the retrieval conditions they anticipate, such that they engage in covert retrieval more often during the processing phase when most of the trials (i.e., 75%) will inevitably require serial recall, thereby yielding a McCabe effect for both serial-recall and no-recall trials. However, if the task encourages a reactive approach because very few (i.e., 25%) of the trials require serial recall, as may have been the case in Experiment 2, then participants may be less likely to engage in covert retrieval, thereby mitigating the McCabe effect in both serial-recall and no-recall conditions. Conversely, the immediate retrieval demands account would expect a McCabe effect only when there is immediate serial recall and regardless of the ratio of serial-recall to no-recall trials, consistent with the notion that the demands of overt retrieval drive the advantage of complex span over simple span at delay. Using reconstruction to assess retrieval for both immediate and delayed tests would further reinforce that the pattern of results is not due to a mismatch in how the items are retrieved.

Experiment 3

Method

Participants and Design. We recruited participants to take part online via Prolific (www.prolific.co). In order to enhance the similarity to participants in the previous experiments, we

applied a pre-screening so that only native English speakers aged 18-35 years, with normal or corrected-to-normal vision, with no history of cognitive impairment, and who were using a desktop/laptop were able to sign up for the study. In total, 122 participants (M_{age} = 26.15 years, *SD* = 4.97 years) were randomly assigned to one of three groups that varied the proportion of trials requiring immediate serial recall (henceforth referred to as "serial-recall group"): 25% (*n* = 40), 50% (*n* = 40), and 75% (*n* = 42). The remaining factors of task type (simple or complex span) and immediate-recall condition (serial or no recall) were manipulated within-subjects as in the previous experiments. Given that reconstruction was used at both times of test, free and serial scoring for both the immediate and delayed tests were the principal dependent variables.

An additional 30 participants (M_{age} = 25.67 years, *SD* = 4.62 years) completed a control experiment that only varied task type within-subjects. One additional participant in the control experiment and five additional participants in the main experiment were excluded from analysis for quitting before finishing the experiment. One further participant in the main experiment completed the experiment twice for an unknown reason, and thus only their first dataset of was included in the analysis. The experiment lasted approximately 10-20 minutes for most participants, and they were compensated with £2.50.

Materials and Procedure. Experiment 3 was programmed in Inquisit (2018). The advertisement on Prolific advised participants that they should be prepared to do the experiment in one continuous sitting in a quiet, distraction-free environment. They were also informed of the general nature of the task of trying to remember information while performing distracting tasks and that they could view their overall performance at the end of the experiment to increase interest and motivation. After signing up, participants installed a plugin to allow the experiment to fill the screen, thereby preventing them from engaging in other tasks on their computers during the experiment.

Participants first completed the practice arithmetic and digit tasks that were identical to Experiment 2. They next received instructions for the critical task that entailed one block of 16 trials, 8

trials of each task type (simple and complex span), randomly intermixed. Like the previous experiments, participants were instructed to read each word out loud only one time as it appeared and try to remember them. They were also instructed to read aloud and respond to the arithmetic problems as quickly and accurately as possible when they were presented. The trials ended unpredictably with either serial recall or the no-recall digits task. Like in the previous experiments, during the serial-recall condition, each of the four presented words were randomly arranged among four never-presented lures in red font and red frames, with the instruction "SERIAL" and "Use the mouse to try to select the 4 presented words in their original order" above the frames. During the no-recall condition, eight double-digit numbers were randomly arranged in blue font and blue frames with the instruction "DIGITS" and "Use the mouse to try to select the 4 double-digit numbers that are both even" above them.[2]

Most importantly, before the block began, participants were told, according to their group assignment, the ratio of serial-recall to no-recall trials to expect. Participants in the 25% serial-recall group completed four serial-recall trials (two of each task type) and 12 no-recall trials (six of each task type); participants in the 50% serial-recall group completed eight trials of both serial and no recall (four of each task type); and participants in the 75% serial-recall group completed 12 serial-recall trials (six of each task type) and four no-recall trials (two of each task type). All the trials were randomly intermixed. The task also regularly emphasized the importance of following the instructions, and participants were warned that they would be sent back to the practice round if their responses were not registered. Twenty participants received a first warning during the critical task, and a further four participants returned to the arithmetic practice phase once during the block for continuing to not respond to the arithmetic problems after the first warning. Furthermore, at the conclusion of the experiment, participants filled in a survey regarding whether they read and answered the arithmetic problems aloud, read the words aloud

[2] Note that unlike the previous experiments, it was possible to select the same item more than once. This only occurred 1.04% and 0.28% of the time during the immediate and delayed tests, respectively. These instances were corrected so that a response was not marked as correct more than once.

only once, and completed the experiment in one sitting in a quiet, distraction-free environment. Most participants reported compliance, and the results were similar when excluding the 13 participants who reported not complying with one or more of these instructions.

After completing the block, participants completed a 1-min distraction phase identical to the previous experiments, followed by instructions for the delayed reconstruction test. Participants were presented with all the trials of the previous task in a new random order, with each displaying the four originally presented words that were randomly arranged among four never-presented lures within black frames and in black font. Participants were once again instructed to try to recall the four presented words in their original order. After completing the instruction compliance and basic demographics survey, participants were offered the chance to view their overall performance.

The control experiment was very similar to the main experiment, except that participants only practiced the arithmetic problems and completed one block of 100% serial-recall trials of simple and complex span. There were eight trials of each task type, randomly intermixed. Two participants received a first warning during the critical task and no participants repeated the practice phase. Only one participant reported not reading and responding to the arithmetic problems aloud. As was the case for the main experiment, the results were similar when excluding this participant.

Results and Discussion

We first report on the results of the control experiment. Participants were similarly accurate during the processing task as the previous experiments, albeit generally faster (Table 1). Figure 2 also suggests that, like the previous experiments, immediate reconstruction was greater for simple span than complex span in both serial scoring (BF_{10} = 180.10) and free scoring (BF_{10} = 10.59). Most importantly, a credible McCabe effect in delayed reconstruction was observed in free scoring, but not serial scoring (see Table 4). This demonstrates that the McCabe effect can be replicated using reconstruction at delay and

when administering the experiment online, although apparently just in the free-scoring measure. These results also provide a benchmark against which to compare the next results of the main experiment.

For the main experiment, we first ensured that participants were consistent in their processing task performance, regardless of the recall conditions or their assigned serial-recall group (see Table 1). The results of two 2 (immediate-recall condition: serial, no recall) x 3 (serial-recall group: 25%, 50%, 75%) mixed BANOVAs largely confirmed this (all BF_{01}s > 2.24). A 2 (task type: simple, complex) x 3 (serial group: 25%, 50%, 75%) mixed BANOVA also indicated that participants were also highly accurate and consistent in their performance on the no-recall digits task (all *M*s > 0.95, all BF_{01}s > 6.66).

We next conducted two 2 (task type: simple, complex) x 3 (serial-recall group: 25%, 50%, 75%) mixed BANOVAs to assess immediate performance in terms of free and serial scoring (see Table 2 and Figure 2). The best model included main effects of task type and serial-recall group, which was ambiguously preferred to the next best model including only a main effect of task type. Thus, like the previous experiments, participants were more likely to recall the simple-span items in order compared to the complex-span items, with an ambiguous indication that serial scoring improved overall as the number of serial-recall trials increased. For free scoring, the best model included a main effect of task type, which was ambiguously preferred to the next best model including effects of both task type and serial-recall group. Thus, as in the previous experiments, participants were still slightly disadvantaged to recall complex-span items at all compared to simple-span items.

Finally, the most important results concerned delayed performance (see Figure 3). We conducted two 2 (immediate-recall condition: serial, no recall) x 2 (task type: simple, complex) x 3 (serial-recall group: 25%, 50%, 75%) mixed BANOVAs to assess delayed free and serial scoring. For the sake of brevity, we report on the best models, which for both measures included main effects of recall condition and task type (free: $BF_{10} = 1.16 \times 10^{25}$; serial: $BF_{10} = 7.86 \times 10^{7}$). These models were substantially preferred (free: BF = 7.12; serial: BF = 5.36) to the next best models including a recall x task type interaction (free: BF_{10} = 1.63

x 10^{24}; serial: BF_{10} = 1.46 x 10^{7}). Thus, the results showed an unsurprising testing effect, such that having attempted serial recall during the immediate test improved long-term retention compared to the no-recall condition. Furthermore, the overall effect of task type indicated a McCabe effect regardless of the other factors.

Given our specific predictions, we more closely examined the McCabe effect of each comparison as in the previous experiments (see Table 4). At first glance, the BFs of these results seem to conflict with the omnibus BANOVA, such that a McCabe effect was only clearly evident in a few comparisons, but the pattern did not fit with either the covert retrieval or immediate retrieval demands accounts. We noticed that the cases where the effect was clearest happened to be the cells of the design that had the most trials. For example, a McCabe effect was evident in the no-recall condition of the 25% serial-recall group, but this level included eight trials per participant, whereas its serial-recall condition only had four trials per participant. This provided an unexpected opportunity to confirm that the mixed effects modeling included in the previous experiments may be better sensitive to test these effects. Indeed, as presented alongside the BFs, the posterior estimates indicated credible McCabe effects in the free scoring measure of all but two of the comparisons. For serial scoring, the results were more consistent between the aggregate and mixed effects analyses, perhaps indicating that performance on this measure is more variable.

Overall, these results conflict with the immediate retrieval demands account given that a McCabe effect was observed in delayed free scoring largely regardless of immediate-recall conditions. The results also suggest that the anticipation of the retrieval demands also do not moderate the McCabe effect, and thus the null McCabe effects in Experiment 2 may instead be due to low overall performance. Finally, the observation of a McCabe effect under no-recall conditions indicates that the null McCabe effect in the free-recall condition of Experiment 1 was likely due to the interference introduced by recalling the items in a free order. Thus, the results of Experiment 3 shed new albeit nuanced light on the previous

experiments: The nature of immediate retrieval may constrain the McCabe effect in some situations (e.g., by introducing a mismatch to encoding during free recall), but its demands do not drive the McCabe effect. As we further discuss, the results do not unequivocally support the covert retrieval model, but we can be more confident that the McCabe effect does not owe to the act of overcoming the difficulty of recovering items from complex span over simple span.

General Discussion

The aim of the current study was to adjudicate between two accounts of the long-term benefits of studying and recalling information from WM. Specifically, we investigated whether the delayed advantage for items originally presented during complex span over simple span, or the McCabe effect (McCabe, 2008), may be moderated by the immediate retrieval demands of the tasks. According to the covert retrieval account (Loaiza & Halse, 2019; Loaiza & McCabe, 2012; McCabe, 2008), the activities during the processing phase are most important to the McCabe effect. That is, we have argued that cumulative covert retrieval of the memoranda occurs during the intermittent pauses afforded by complex span tasks, in turn reinforcing the retrieval cues used to recall the information again from EM. Conversely, the immediate retrieval demands account would suggest that serial recall is much more challenging during complex span compared to simple span, and this asymmetric difficulty of overt retrieval is what drives the McCabe effect. Thus, the two accounts focus on different elements of complex span as the source for the long-recognized role of WM for long-term retention (Atkinson & Shiffrin, 1968; Hartshorne & Makovski, 2019; McCabe et al., 2010; Unsworth, 2007, 2016): The covert retrieval account emphasizes covert, internal retrieval practice during the processing phase, whereas the immediate retrieval demands account emphasizes the actual act of overt recall during the retrieval phase.

Taken together, the collective results conflict with the immediate retrieval demands account that a McCabe effect should only be observed under difficult overt retrieval conditions. This stems from two main findings: A McCabe effect was not observed in the serial-recall condition of Experiment 2, but

McCabe effects were observed in most of the no-recall conditions of Experiment 3, in stark contrast to the predictions of the immediate retrieval demands account. Furthermore, the results of Experiment 3 provide an alternative interpretation of two other results seemingly in line with the immediate retrieval demands account. First, there was a lack of a McCabe effect in the free-recall condition of Experiment 1, but Experiment 3 replicated the McCabe effect under even-easier no-recall conditions (McCabe, 2008, Experiment 3). Thus, the lack of McCabe effect in the free-recall condition in Experiment 1 may have occurred due to a mismatch between the cumulative covert retrieval to keep the memoranda active during the processing phase and the retrieval method. Second, the null McCabe effects in the free-recall and no-recall conditions of Experiment 2 may be due to relatively low performance during the delayed test. Using reconstruction during both immediate and delayed tests as in Experiment 3 increased overall delayed performance, thereby allowing greater sensitivity to detect a McCabe effect in both serial- and no-recall conditions. Thus, the combination of results suggest that immediate retrieval demands are not responsible for the McCabe effect.

The results are instead more consistent with the covert retrieval model's assertion that the activities during the processing phase are most important to the McCabe effect, although there is admittedly some room for interpretation regarding what those activities entail. The findings of Experiment 3 were particularly important in that they indicated that a McCabe effect can be demonstrated regardless of overt immediate retrieval, thereby better isolating the effect to the processing phase rather than the retrieval phase. However, warning participants about the proportion of serial-recall trials to expect did not moderate the McCabe effect as we had predicted. This may suggest that participants do not adapt their maintenance strategy accordingly and simply engage in covert retrieval regardless of the anticipated retrieval requirements, or it may suggest that another factor is at play during the processing phase. Furthermore, the McCabe effect was more consistently evident in free than serial scoring. In our previous work, we asserted that covert retrieval is particularly important to

reinforcing content-context bindings that yield significantly greater use of temporal-contextual cues to guide retrieval (Loaiza & McCabe, 2012) and greater subjective experiences of recollection (Loaiza et al., 2015) during EM. Accordingly, we should have observed greater long-term retention of serial order of the complex-span versus simple-span items, indicating that the content-context bindings were more durable. However, it is possible that the use of reconstruction introduced another context layer of spatial position on the retrieval screen. The items were randomly arranged during both tests, and thus it is possible that this introduced interference that caused variability in delayed serial scoring performance. Further research will be necessary to investigate these possibilities.

In sum, the results suggest a nuanced conclusion about the source of the McCabe effect: Although the effect is not attributable to the act of overcoming the relatively difficult immediate retrieval demands of complex span versus simple span, it is clear that retrieval conditions more generally can moderate the influence of the processing phase activities underlying the McCabe effect. First, immediately recalling complex-span items in any order may introduce interference from mismatched encoding-retrieval conditions that mitigates the McCabe effect (Experiment 1). Thus, immediate serial recall may reinforce any cumulative covert retrieval during the processing phase of complex span, but it is not necessary to engage in immediate serial recall to observe a McCabe effect (Experiment 3). Furthermore, the delayed retrieval conditions are also important to affording the opportunity to observe a McCabe effect: If overall retrieval is too low, there will necessarily be a reduced opportunity to observe a McCabe effect (Experiment 2). Using retrieval paradigms that allow for better overall recall (e.g., reconstruction) will enhance the possibility to observe a McCabe effect. Finally, it is important to note that a sufficient number of trials and a mixed effects analysis approach will help ensure that inevitable variability in performance is adequately accommodated, and so we advise this for future work.

Finally, the results of Experiment 3 are particularly interesting given the matched retrieval method of reconstruction between times of test, and thus bears on the broader theoretical discussion regarding

the boundary between WM and EM. First, the mismatch in retrieval methods between immediate and delayed tests of our previous work has been overdue to address to ensure that this methodological difference was not responsible for the McCabe effect. The fact that we observed a McCabe effect using delayed reconstruction provides more certainty that the effect is replicable for multiple tests of EM. Furthermore, although most researchers would agree that delayed tests measure retrieval from EM, there are some who would argue that immediate retrieval from complex and simple span does not necessitate an additional WM system, but simply reflects EM at a shorter time-scale (e.g., Crowder, 1982; Nairne, 2002). However, that the reverse pattern of recall from simple and complex span occurred between two times of test, using the same method of retrieval, greatly conflicts with this unitary view of memory. Instead, the results suggest that a factor like distraction within a task has completely different effects on the online maintenance and manipulation in WM and later retrieval of information from EM.

In conclusion, the results of the current experiment contradict the notion that overt retrieval demands of information from WM promote their long-term retention. Our results suggest that overcoming the disproportionate immediate retrieval demands do not explain the long-term advantage of complex span over simple span, i.e., the McCabe effect. Instead, the McCabe effect is more likely attributable to activities taking place during the processing phase, which may include covert retrieval.

Open Practices Statement

The materials, data, and analysis scripts for all the experiments are available at the Open Science Framework (OSF): https://osf.io/c9dsw/ Experiments 2 and 3 were preregistered via AsPredicted.org, with the preregistration document available on the OSF.

References

Abadie, M., & Camos, V. (2018). Attentional refreshing moderates the word frequency effect in immediate and delayed recall tasks. *Annals of the New York Academy of Sciences*, *0*(0). https://doi.org/10.1111/nyas.13847

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–105). Academic Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

Barrouillet, P., & Camos, V. (2015). *Working Memory: Loss and reconstruction* (1 edition). 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

Bays, P. M., & Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Bjork, R. A. (1994). Memory and metamemory consideration in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knwoing about knowing* (pp. 185–205). MIT Press.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897x00357

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic Bulletin & Review*, *22*(4), 1029–1034. https://doi.org/10.3758/s13423-014-0772-5

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Cowan, N. (1999). An embedded process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 62–101). Cambridge University Press.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Crowder, R. G. (1982). The demise of short-term-memory. *Acta Psychologica*, *50*(3), 291–323. https://doi.org/10.1016/0001-6918(82)90044-0

Engle, R. W., & Kane, M. J. (2004). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. In *The psychology of learning and motivation: Advances in research and theory, Vol. 44* (pp. 145–199). Elsevier Science.

Fougnie, D., Zughni, S., Godwin, D., & Marois, R. (2015). Working memory storage is intrinsically domain specific. *Journal of Experimental Psychology: General*, *144*(1), 30–47. https://doi.org/10.1037/a0038211

Hartshorne, J. K., & Makovski, T. (2019). The effect of working memory maintenance on long-term memory. *Memory & Cognition*, *47*(4), 749–763. https://doi.org/10.3758/s13421-019-00908-6

Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory mechanisms and the control of attention. In *Variation in working memory* (pp. 227–249). Oxford University Press.

Inquisit 5. (5.0.14.0). (2018). [Windows]. https://www.millisecond.com

Jarjat, G., Hoareau, V., Plancher, G., Hot, P., Lemaire, B., & Portrat, S. (2018). What makes working memory traces stable over time? *Annals of the New York Academy of Sciences*. https://doi.org/10.1111/nyas.13668

Johnson, M. K. (1992). MEM: Mechanisms of Recollection. *Journal of Cognitive Neuroscience*, *4*(3), 268–280. https://doi.org/10.1162/jocn.1992.4.3.268

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1–16.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Loaiza, V. M., & Borovanska, B. M. (2018). Covert retrieval in working memory impacts the phenomenological characteristics remembered during episodic memory. *Consciousness and Cognition*, *57*(Supplement C), 20–32. https://doi.org/10.1016/j.concog.2017.11.002

Loaiza, V. M., Duperreault, K. A., Rhodes, M. G., & McCabe, D. P. (2015). Long-term semantic representations moderate the effect of attentional refreshing on episodic memory. *Psychonomic Bulletin & Review*, *22*(1), 274–280. https://doi.org/10.3758/s13423-014-0673-7

Loaiza, V. M., & Halse, S. C. (2019). Where working memory meets long-term memory: The interplay of list length and distractors on memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(8), 1455–1472. https://doi.org/10.1037/xlm0000652

Loaiza, V. M., & McCabe, D. P. (2012). Temporal–contextual processing in working memory: Evidence from delayed cued recall and delayed free recall tests. *Memory & Cognition*, *40*(2), 191–203. https://doi.org/10.3758/s13421-011-0148-2

Loaiza, V. M., McCabe, D. P., Youngblood, J. L., Rose, N. S., & Myerson, J. (2011). The influence of levels of processing on recall from working memory and delayed recall tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1258–1263. https://doi.org/10.1037/a0023923

Loaiza, V. M., & Souza, A. S. (2018). Is refreshing in working memory impaired in older age? Evidence from the retro-cue paradigm. *Annals of the New York Academy of Sciences*, *1424*(1), 175–189. https://doi.org/10.1111/nyas.13623

Loaiza, V. M., & Souza, A. S. (2019). An Age-Related Deficit in Preserving the Benefits of Attention in Working Memory. *Psychology and Aging*, *34*, 268–281.

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

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language*, *58*(2), 480–494. https://doi.org/10.1016/j.jml.2007.04.004

McCabe, D. P., Roediger, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology*, *24*(2), 222–243. https://doi.org/10.1037/a0017619

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, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs.* (0.9.12-2) [Computer software]. http://CRAN.R-project.org/package=BayesFactor

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, *53*, 53–81. https://doi.org/10.1146/annurev.psych.53.100901.135131

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. (2019). Working Memory Capacity Limits Memory for Bindings. *Journal of Cognition*, *2*(1), 40. https://doi.org/10.5334/joc.86

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Pratte, M. S. (2020). Set size effects on working memory precision are not due to an averaging of slots. *Attention, Perception, & Psychophysics*, *82*(6), 1–13. https://doi.org/10.3758/s13414-019-01902-5

R core team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Rhodes, S., Jaroslawska, A. J., Doherty, J. M., Belletier, C., Naveh-Benjamin, M., Cowan, N., Camos, V., Barrouillet, P., & Logie, R. H. (2019). Storage and processing in working memory: Assessing dual-task performance and task prioritization across the adult lifespan. *Journal of Experimental Psychology: General*, *148*(7), 1204–1227. https://doi.org/10.1037/xge0000539

Rose, N. S., Buchsbaum, B. R., & Craik, F. I. M. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition*, *42*, 689–700. https://doi.org/10.3758/s13421-014-0398-x

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Souza, A. S., & Oberauer, K. (2017). Time to process information in working memory improves episodic memory. *Journal of Memory and Language*, *96*, 155–167. https://doi.org/10.1016/j.jml.2017.07.002

Stan Development Team. (2018). *Stan Modeling Language: User's guide and reference manual (Version 2.17.4).* http://mc-stan.org/users/documentation

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is working memory storage intrinsically domain-specific? *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000566

Unsworth, N. (2007). Individual differences in working memory capacity and episodic retrieval: Examining the dynamics of delayed and continuous distractor free recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(6), 1020–1034. https://doi.org/10.1037/0278-7393.33.6.1020

Unsworth, N. (2016). Working memory capacity and recall from long-term memory: Examining the influences of encoding strategies, study time allocation, search efficiency, and monitoring abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 50–61. https://doi.org/10.1037/xlm0000148

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038–1066. https://doi.org/10.1037/0033-2909.133.6.1038

Vergauwe, E., & Cowan, N. (2015). Attending to items in working memory: Evidence that refreshing and memory search are closely related. *Psychonomic Bulletin & Review*, *22*(4), 1001–1006. https://doi.org/10.3758/s13423-014-0755-6

Vergauwe, E., Hardman, K. O., Rouder, J. N., Roemer, E., McAllaster, S., & Cowan, N. (2016). Searching for serial refreshing in working memory: Using response times to track the content of the focus of attention over time. *Psychonomic Bulletin & Review*, *23*(6), 1818–1824. https://doi.org/10.3758/s13423-016-1038-1

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, *36*(5), 1207–1241. https://doi.org/10.1037/a0020122

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233. https://doi.org/10.1038/nature06860

Table 1. Mean proportion accuracy and response times (and standard deviations) during the complex span secondary processing task across experiments.

Exp.	Immediate Recall Instructions	Serial-Recall Group	Accuracy	Response time (s)
1	Serial recall	-	0.93 (0.19)	2.20 (0.31)
	Free recall	-	0.92 (0.18)	2.22 (0.26)
2	Serial recall	-	0.92 (0.19)	2.18 (0.45)
	Free recall	-	0.93 (0.19)	2.16 (0.41)
	No recall	-	0.93 (0.19)	2.18 (0.42)
3	Serial recall	100% (control)	0.96 (0.05)	1.80 (0.45)
	Serial recall	25%	0.95 (0.11)	1.76 (0.43)
	No recall	25%	0.95 (0.08)	1.72 (0.36)
	Serial recall	50%	0.95 (0.07)	1.70 (0.41)
	No recall	50%	0.95 (0.07)	1.73 (0.35)
	Serial recall	75%	0.95 (0.06)	1.62 (0.37)
	No recall	75%	0.95 (0.08)	1.62 (0.44)

Note. Exp. = experiment.

Table 2. Results of the BANOVAs for immediate recall measures (both serial and free scoring) for each experiment.

			Fixed effects						
Exp.	Measure	Model (M) Ratio	Immediate recall instruction	Task type	Serial-recall group	Recall + Task	Recall + Task + Recall x Task	Task + Group	Task + Group + Task x Group
1	Serial scoring	BF_{10}	6.11×10^{12}	178.28	-	3.46×10^{18}	**7.74×10^{20}**	-	-
		Best M/M	1.27×10^{8}	4.34×10^{18}	-	223.80	Best	-	-
	Free scoring	BF_{10}	0.22	**21819.20**	-	4677.89	1412.08	-	-
		Best M/M	1.01×10^{5}	Best	-	4.66	15.45	-	-
2	Serial scoring	BF_{10}	1.14×10^{8}	4.08×10^{7}	-	**2.35×10^{21}**	1.01×10^{21}	-	-
		Best M/M	2.06×10^{13}	5.75×10^{13}	-	Best	2.33	-	-
	Free scoring	BF_{10}	0.60	**1351.24**	-	1039.24	661.39	-	-
		Best M/M	2261.14	Best	-	1.30	2.04	-	-
3	Serial scoring	BF_{10}	-	3.92×10^{6}	1.56	-	-	**7.53×10^{6}**	7.82×10^{5}
		Best M/M	-	1.92	4.83×10^{6}	-	-	Best	9.63
	Free scoring	BF_{10}	-	**21.82**	0.32	-	-	7.32	0.56
		Best M/M	-	Best	67.75	-	-	2.98	38.91

Note. All models include participant as a random effect. The Bayes factor (BF) refers to the evidence for the alternative model (BF_{10}) for each effect (shown in different columns) relative to the null model (i.e., intercept-only model). The best model is shown in boldface in the first row for each measure, and the second row for each measure compares the best model in the numerator to each of the other models in the denominator.

Table 3. Results of the BANOVAs for overall delayed free recall for Experiments 1 and 2.

		Fixed effects			
Exp.	Model (M) Ratio	Immediate recall instruction	Task type	Recall + Task	Recall + Task + Recall x Task
1	BF_{10}	0.43	**72.35**	34.12	48.38
	Best M/M	167.13	Best	2.12	1.50
2	BF_{10}	**218.09**	0.60	149.32	24.93
	Best M/M	Best	362.75	1.46	8.75

Note. All models include participant as a random effect. The Bayes factor (BF) refers to the evidence for the alternative model (BF_{10}) for each effect (shown in different columns) relative to the null model (i.e., intercept-only model). The best model is shown in boldface in the first row for each measure, and the second row for each measure compares the best model in the numerator to each of the other models in the denominator.

Table 4. Evidence for the McCabe effect in delayed free and serial scoring for each cell of the design of Experiment 3.

			McCabe effect		
Measure	Serial-recall group	Immediate recall instruction	BF_{10}	Effect size	HDI
Free scoring	100%	Serial recall	**20.84**	**-0.42**	**[-0.70, -0.14]**
		No recall	-	-	-
	25%	Serial recall	1/1.88	-0.44	[-0.90, 0.00]
		No recall	**10.08**	**-0.41**	**[-0.62, -0.20]**
	50%	Serial recall	1.49	**-0.45**	**[-0.81, -0.10]**
		No recall	1.54	**-0.35**	**[-0.62, -0.07]**
	75%	Serial recall	**62.33**	**-0.59**	**[-0.88, -0.33]**
		No recall	1/1.62	-0.28	[-0.66, 0.08]
Serial scoring	100%	Serial recall	2.07	-0.32	[-0.69, 0.08]
		No recall	-	-	-
	25%	Serial recall	1/3.64	-0.20	[-0.61, 0.19]
		No recall	**19.51**	**-0.46**	**[-0.73, -0.21]**
	50%	Serial recall	1/3.68	-0.19	[-0.58, 0.22]
		No recall	1/1.72	-0.28	[-0.62, 0.04]
	75%	Serial recall	**32.10**	**-0.47**	**[-0.76, -0.18]**
		No recall	2.87	**-0.56**	**[-0.93, -0.18]**

Note. BF = Bayes factor; HDI = highest density interval. Credible effects are highlighted in bold. BFs in favor of the null are expressed as their inverse to enhance clarity and comparison.

Figure 1. Mean proportion of recalled items, scored as accurate according to their original serial order of presentation (top panel), in any order (middle panel), and at delay (bottom panel) in Experiments 1 and 2. Error bars reflect 95% within-subjects confidence intervals.

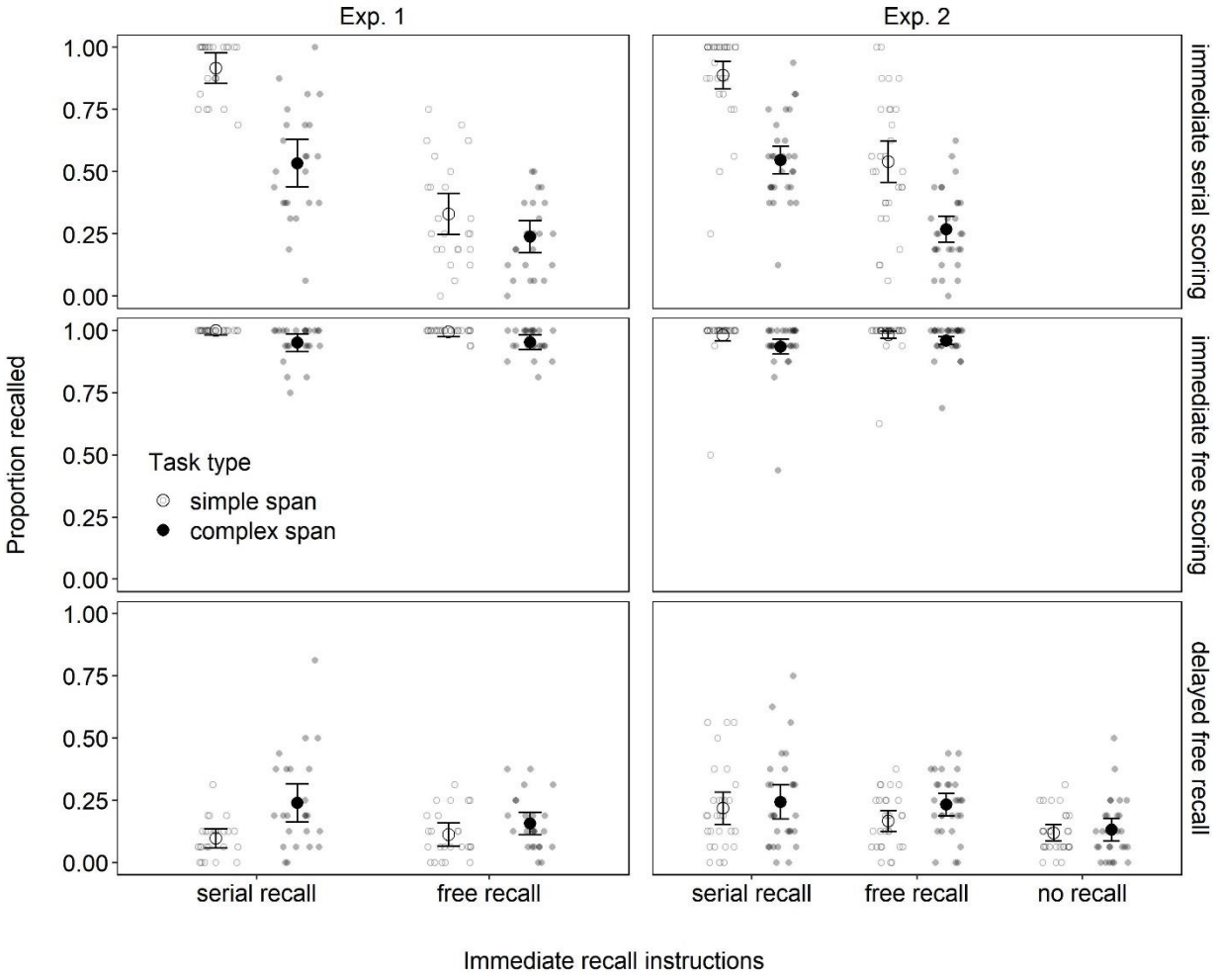

Figure 2. Mean proportion recalled at the immediate test in terms of serial scoring (top panel) and free scoring (bottom panel) in Experiment 3. Error bars reflect 95% within-subjects confidence intervals.

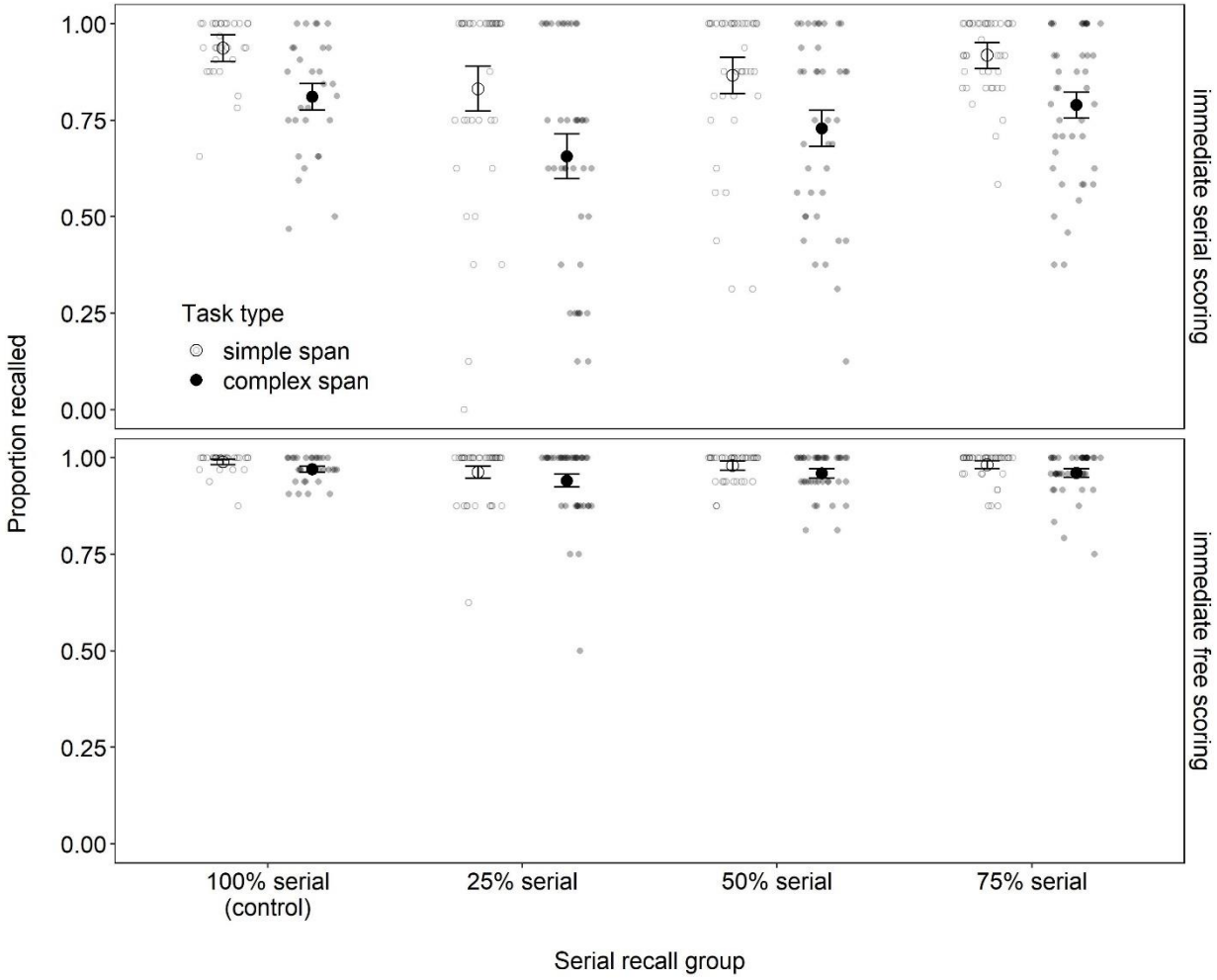

Figure 3. Mean proportion recalled at the delayed test in terms of serial scoring (top panel) and free scoring (bottom panel) in Experiment 3. Error bars reflect 95% within-subjects confidence intervals.