Elaborative Strategies Contribute to the Long-Term Benefits of Time in Working Memory

Vanessa M. Loaiza

Emilio Tomas Lavilla

University of Essex

This is a pre-print of an article accepted for publication at *Journal of Memory and Language* on 6 December 2020.

Author Note

Vanessa M. Loaiza, Department of Psychology, University of Essex. Emilio Tomas Lavilla, 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

The authors acknowledge Naveed Khan, Priyasha Khurana, and Maria Sanz Taberner for their assistance with data collection.

The pre-registrations, materials, data, and analysis scripts for the experiments are available at the Open Science Framework: https://osf.io/3rqgf/

Highlights

- Episodic memory (EM) is greater for complex span and slow span versus simple span.
- We predicted that slow span disproportionately affords opportunity for elaboration.
- Spontaneous elaboration contributed to both complex and slow span benefits.
- Irrespective of covert retrieval, time in WM allows elaboration that promotes EM.

Abstract

Word count: 150/150

The current experiments investigated the long-term advantage for words studied during complex span versus simple span, i.e., the McCabe effect. According to the covert retrieval account, the McCabe effect occurs because complex span affords covert retrieval opportunities that facilitate episodic memory (EM). Conversely, the time-in-working-memory (WM) hypothesis asserts that the time items spend in WM predicts EM, irrespective of any opportunity for covert retrieval. We investigated whether time specifically allows for elaboration in WM by considering the influence of reported and instructed strategies during simple span, complex span, and slow span, where for the latter, a pause of equivalent duration to the distraction in complex span interleaved the memoranda. The results indicated that (i) elaboration is just as frequent during complex span as slow span and (ii) spontaneous elaboration accentuates the advantage of complex span and slow span over simple span, commensurate with the elaboration account of the time-in-WM hypothesis.

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

Elaborative Strategies Contribute to the Long-Term Benefits of Time in Working Memory

Memory is a critical component of human cognition. In daily life, we often must efficiently maintain and update information in service to an ongoing task, such as following a conversation or instructions for arrival at a destination. The memory system thought to support this ongoing cognition is working memory (WM). Moreover, this information may become later relevant after it has left our immediate awareness in WM, such as thinking back to the conversation a few minutes, hours, or days later. In this case, the information must be retrieved from long-term episodic memory (EM). Much work has investigated the interaction and overlap between these two memory systems, given that investigating their relationship provides a means of understanding the architecture of and underlying mechanisms supporting human memory.

Research concerning the WM-EM[1] relationship can be traced back over 100 years to the beginnings of psychology (Ebbinghaus, 1885; James, 1890), a history that has been fraught with debate and no clear resolution in sight (Cowan, 2019; Norris, 2017). Much of this research concerns how presently experienced events and information are retained over the long term. This work has often centered on the importance of WM given that it is thought to provide the mental workspace to hold and manipulate task-relevant information in mind (Baddeley & Hitch, 1974; Cowan, 2017) and because its capacity limitations are known to severely constrain higher-order cognition like EM (McCabe et al., 2010; Unsworth & Spillers, 2010; Unsworth et al., 2013; Wilhelm et al., 2013). More recently, the topic has found renewed interest in a growing area of research concerning the downstream consequences of mechanisms and processes presumed to support maintenance in WM (Bartsch et al., 2018, 2019; Camos & Portrat, 2015; Jarjat et al., 2018; Loaiza & McCabe, 2012; McCabe, 2008; Rose et al., 2014; Souza & Oberauer, 2017). In the current work, we considered three theoretical

[1] It should be noted that previous terms used in the literature have respectively referred to the short-term store/long-term store, short-term/long-term memory, and primary/secondary memory rather than WM/EM, or sometimes WM/long-term memory. We do not wish to contribute to the glut of terms that likely are meant to reflect the same respective constructs. Our reasoning for using the terms WM and EM are as follows: We use "WM" given that it better reflects a more dynamic immediate memory system described previously, and "EM" because "long-term memory" is technically ambiguous as it is often broadly assumed to comprise both EM and semantic memory (Tulving, 1972, 2002). Thus, we will use WM and EM throughout the current work.

accounts for why maintaining information in WM may impact later retrieval from EM: the covert retrieval model, the time-in-WM hypothesis, and an elaboration version of the time-in-WM hypothesis.

The Covert Retrieval Model

The main premise of the covert retrieval model is that long-term retention is at least partly driven by the relatively durable retrieval cues that are established as a consequence of reviving memory traces after attention has been distracted. McCabe (2008) demonstrated evidence for this account with two common measures in the WM literature: complex span and simple span. Complex span tasks (e.g., operation span; Turner & Engle, 1989) interleave several successively presented memoranda (e.g., words) and short distracting processing components (e.g., arithmetic problems to read aloud and solve), with a cue to recall the words at the end of the trial in their original serial order. McCabe (2008) argued that as complex span trials progress, participants must successively and cumulatively covertly retrieve the memoranda back into the central component of WM after their attention has been distracted from maintaining the items. This repeated, internal retrieval practice of recovering information from outside immediate awareness was asserted to instantiate the cues that would later facilitate retrieval from EM (McCabe, 2008). Conversely, simple span tasks (e.g., word span) that present only a few items without any distraction would not require covert retrieval given that attention is never distracted from their maintenance. In line with this reasoning, McCabe observed that delayed recall of memoranda that originally had been studied during complex span was greater than that of simple span, also known as the *McCabe effect*.

The McCabe effect provides an intriguing means by which to investigate the impact of WM processing on long-term EM. Given that complex span tasks require flexible maintenance of items while coordinating distraction, they are often considered a canonical measure (Conway et al., 2005) that many researchers have used to investigate the underlying mechanisms supporting WM (e.g., Barrouillet et al., 2004; Camos et al., 2009; Engle et al., 1999). Thus, the McCabe effect demonstrates direct evidence that whatever processes that allow maintenance and manipulation in WM have long-

term consequences and has inspired much subsequent work (e.g., Abadie & Camos, 2018; Bartsch et al., 2018, 2019; Camos & Portrat, 2015; Jarjat et al., 2018; Souza & Oberauer, 2017). In the next sections, we detail two alternative accounts of the McCabe effect that speak to the overarching issue of why underlying WM mechanisms may improve long-term retention.

The Time-in-WM Hypothesis

An alternative account of the McCabe effect suggests that the overall time that items spend in WM, rather than any covert retrieval per se, moderates the likelihood of their recall from EM (Jarjat et al., 2018; Souza & Oberauer, 2017). For example, Souza and Oberauer have asserted that the memoranda presented during complex span are not displaced from WM as the covert retrieval model holds, but instead their relatively greater likelihood to be recalled owes to the fact that the distracting processing components prolong their encoding and maintenance into EM compared to simple span. To investigate this possibility, Souza and Oberauer presented participants with trials of simple and complex span as in McCabe (2008), but additionally presented *slow span* trials wherein the memoranda were interleaved with brief unfilled retention intervals that equaled the length of the duration of the distracter-filled retention interval of complex span. The results indicated that delayed recall from slow span was greater than that of complex span and simple span, and delayed recall from complex span was greater than simple span only when conditionalizing on accurate initial recall (Souza & Oberauer, 2017). Such findings present a strong challenge for the covert retrieval model: If cues established by covertly retrieving displaced information from WM are crucial for long-term retention, then, strictly speaking, EM should be greatest for complex span that affords covert retrieval opportunities via distraction compared to simple and slow span wherein attention was never distracted from maintenance. The results of Souza and Oberauer stand in stark contrast to this prediction and instead suggest that time spent in WM predicts EM.

Other recent work has similarly demonstrated the importance of time in WM for retrieval from EM, harking back to early dual-store models of memory (Atkinson & Shiffrin, 1968; Waugh & Norman, 1965). For example, Jarjat and colleagues (2018) demonstrated a logarithmic relationship

between accumulated free time during the presentation of complex span memoranda and their delayed recall: The impact of time on delayed recall was strongest for shorter accumulated free time, and became less influential as accumulated free time increased. Hartshorne and Makovski (2019) also showed via a meta-analysis of previous work and 13 novel experiments that items were more likely to be recalled from EM the longer they had been maintained in WM.

Of course, these results beg the question of what precisely "time" in WM allows or promotes in terms of the activities that facilitate retrieval from EM. Jarjat and colleagues (2018) suggested that prolonged time in WM may allow for greater opportunity to engage in refreshing, thereby leading to EM benefits, as has been observed in previous work (Camos & Portrat, 2015; Johnson et al., 2002; but see Bartsch et al., 2018, 2019). Refreshing is considered a domain-general, attention-based mechanism that briefly reactivates just recently available information to keep it active for ongoing processing in WM (see Camos et al., 2018 for a recent review). Souza and Oberauer (2017) further suggested that greater opportunity to consolidate just-encoded memoranda as they are successively presented may be afforded by the longer time that items spend in WM. The goal of the current work was to examine Souza and Oberauer's second suggestion that time may afford greater opportunity to engage in elaborative strategies that consequently benefit EM, and importantly, this opportunity for and benefit of elaboration should occur regardless of whether distraction, and perhaps concomitantly covert retrieval, is present during the task.

Time-in-WM-for-Elaboration Hypothesis

A great deal of work has demonstrated the beneficial impact of elaborative strategies, such that retrieval from EM is more likely when information is meaningfully processed with regard to its deeper, semantic rather than shallower, phonological/orthographic characteristics (Craik & Tulving, 1975; Hyde & Jenkins, 1969, 1973; Loaiza & Camos, 2016; Loaiza et al., 2011; Rogers et al., 1977; Rose et al., 2014, 2010; Rose & Craik, 2012). Although much of this work has been specific to the EM literature, a growing literature has investigated the impact of elaborative strategies on WM performance as well (Bailey et al., 2011; Bartsch et al., 2018; Dunlosky & Kane, 2007; Loaiza et al.,

2011; Loaiza & Camos, 2016, 2018; Mazuryk & Lockhart, 1974; Nishiyama, 2014, 2018; Rose et al., 2010, 2014, 2015; Rose & Craik, 2012; Shivde & Anderson, 2011). This literature can be divided into two main methods of ascertaining the effects of elaboration on memory performance: (i) reported spontaneous strategy use and (ii) instructing participants to use certain strategies.

First, some work has investigated how participants spontaneously adopt and implement elaborative strategies, especially in the context of WM paradigms (e.g., Bailey et al., 2008; Bailey et al., 2009, 2011; Dunlosky & Kane, 2007; Friedman & Miyake, 2004). Typically, this work has administered *strategy reports* either concurrently (i.e., at the same time as or immediately after the presentation of the memoranda/trials) or retrospectively (i.e., after the memory task is complete). These strategy reports often ask participants to select from a discrete set of possible strategies to examine how often participants report using normatively ineffective versus elaborative strategies, and how memory performance differs as a function of these reported strategies. Although elaborative strategy use during WM tasks is not frequently reported, a positive correlation between reported elaboration and performance is often demonstrated in the WM literature (Bailey et al., 2008; Bailey et al., 2009, 2011; Dunlosky & Kane, 2007) and the EM literature (Dunlosky & Hertzog, 1998, 2001; Richardson, 1998).

Furthermore, participants can be instructed to adopt different processing strategies (e.g., instructing semantic, elaborative strategies versus shallow, rehearsal-based strategies) that they implement themselves, in turn yielding an advantage of elaboration for EM performance (Bartsch et al., 2018; Blumenfeld & Ranganath, 2006; Naveh-Benjamin & Jonides, 1984; Thalmann et al., 2019). Some work has also suggested that instructing or training elaborative strategies can improve WM performance (Bailey et al., 2014; McNamara & Scott, 2001), although there is evidence to the contrary, such that instructing participants to use elaboration only impacted EM and not WM (Bartsch et al., 2018, 2019). Notwithstanding, the literature overall suggests that elaborative strategies, whether spontaneously adopted or instructed, are beneficial for retrieval from EM. Thus, it may be the case that greater time allows the opportunity for greater elaboration.

The goal of the current work was to draw upon this literature concerning elaboration in WM to better understand the underlying factors that promote long-term retention in EM. As we have reviewed, there is evidence that elaboration (Bartsch et al., 2019, 2018; Craik & Tulving, 1975) and the total time items spend in WM (Hartshorne & Makovski, 2019; Jarjat et al., 2018; Souza & Oberauer, 2017) improve retrieval from EM. We aimed to determine whether spontaneous and instructed elaborative strategy use during WM is the source of the beneficial effect of time for EM. Most importantly, this elaboration version of the time-in-WM hypothesis and covert retrieval model yield different predictions regarding how elaboration and task type (simple span, complex span, and slow span) may interact: If time allows the opportunity to implement elaborative strategies in WM that improves EM, then elaborative strategies should be similarly reported and beneficial for complex span and slow span given the fixed time between the two task types. Thus, an advantage of complex span and slow span over simple span should be similarly and exclusively evident when participants report using or are instructed to use elaborative strategies, thereby accounting for the effect of time in WM on EM. Conversely, according to the covert retrieval model, elaborative strategies should be more frequent and beneficial to long-term retention of slow span items, specifically, given that attention is never distracted from their maintenance. The presence of distraction during complex span may reduce the opportunity for elaboration, thereby requiring participants to rely on covert retrieval instead. This would imply that the slow span effect is specific to when participants use elaborative strategies, whereas the McCabe effect should occur regardless of strategy use.

In summary, for the covert retrieval model, the relatively greater delayed recall of slow span and complex span items over simple span may occur for different reasons, such that elaboration explains the slow span effect, whereas covert retrieval better accounts for the McCabe effect. The time-in-WM-for-elaboration hypothesis is arguably more parsimonious in its prediction that elaboration should underlie both the slow span and McCabe effects. Besides helping to elucidate the underlying WM processes that facilitate long-term retention, such results would also speak to the overarching question of the overlap between WM and EM.

Current Experiments

We conducted two experiments to address whether spontaneous and instructed strategy use while maintaining information in WM may impact its long-term retention. The experiments replicated the design and procedure of Souza and Oberauer (2017, Experiment 3): Participants studied and immediately recalled words that were presented during simple span, complex span, and slow span trials, followed by delayed recall of the words. The substantive addition to this design was that participants retrospectively reported their strategy use after finishing the entire memory phase (Experiment 1) or concurrently after each trial (Experiment 2). In Experiment 1A, no specific strategies were instructed to maintain the words in WM, whereas in Experiment 1B, participants were instructed and trained to use either ineffective strategies (i.e., silently and repetitively rehearse the words) or elaborative strategies (i.e., generating meaningful sentences, groupings, and imagery of the words). In Experiment 2, participants were randomly assigned to one of these three groups (i.e., no instruction, instructed ineffective strategies, instructed elaborative strategies).

Our pre-registered predictions centered on the use and impact of elaborative strategies on the delayed recall advantages that have previously been observed for complex span and slow span over simple span. Given the covert retrieval account, we predicted that participants would report using elaborative strategies more often during slow span than complex or simple span both spontaneously and when instructed to do so. In turn, this disproportionate use of elaborative strategies should account for any slow span advantage in delayed recall over complex span and simple span, such that the slow span advantage should only be evident when participants report using or are instructed to use elaborative strategies. Conversely, no slow span advantage should be exhibited when participants report or are instructed to use ineffective strategies. Such results would indicate that it is not time in WM per se that improves long-term retention, but rather that increased time allows relatively greater opportunity to engage in elaboration during slow span. Furthermore, the predicted specific benefit of elaboration during slow span and not complex span would suggest that different mechanisms may underlie their beneficial effects for long-term retention.

Experiment 1

Experiment 1 was designed to address whether elaborative strategies are more often implemented during slow span compared to complex span and simple span, thereby leading to the slow span advantage in delayed recall performance. In Experiment 1A, participants were not instructed to use any particular strategies, whereas participants in Experiment 1B were instructed to use either ineffective or elaborative strategies.[2] We predicted that participants would spontaneously report elaborative strategies more often during slow span compared to simple and complex span, and that the slow span advantage over the other task types should be most evident for elaborative strategies (Experiments 1A and 1B). Similarly, we expected that instructing elaborative strategies should be most beneficial for delayed recall from slow span compared to simple and complex span, whereas there should be no slow span advantage for participants who are instructed to use ineffective strategies (Experiment 1B). These were the most crucial hypotheses for our experiment; the full pre-registration of research questions and predictions can be found on the OSF.

Method

Participants. Twenty-four participants (M_{age} = 19.38, *SD* = 1.01) were recruited from the Department of Psychology subject pool at the University of Essex and assigned to complete the experiment with no strategy instructions (Experiment 1A). An additional 48 participants (M_{age} = 21.04, *SD* = 2.80) were evenly and randomly assigned to either the ineffective or elaborative strategy instruction groups (Experiment 1B). We aimed to collect at least 24 participants per instruction group based on similar prior research (Loaiza & Borovanska, 2018; McCabe, 2008; Souza & Oberauer, 2017). Seven additional participants were excluded from the analysis due to failure to complete the study (i.e., leaving during the arithmetic practice phase, *n* = 5), failure to follow instructions (*n* = 1), or experiment malfunction (*n* = 1). Participants in both experiments were native English speakers and had normal or corrected-to-normal vision. All participants provided written informed consent before

[2] Note that Experiment 1B was conceived after conducting Experiment 1A. Thus, we distinguish between Experiments 1A and 1B only because participants were not randomly assigned to the three instruction groups as they were in Experiment 2.

beginning their experiments, and at the conclusion of the experiments, were debriefed and compensated with either partial course credit or £7.50 per hour of participation. The ethics committee at the University of Essex approved the ethics application for all the experiments.

Materials and Procedure. The memoranda were randomly sampled without replacement from a set 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). The words were randomly arranged for each participant. Experiment 1 was programmed in Matlab with the Psychtoolbox extensions (Brainard, 1997; Kleiner et al., 2007).

Participants completed the experiment individually in a quiet testing booth, and the experimenter remained in the testing booth for the duration of the experiment in order to ensure understanding and compliance with the instructions. All participants completed three phases: a practice arithmetic phase, a memory phase, and a strategy report phase. Participants in Experiment 1B additionally completed a strategy instruction and training phase following the practice arithmetic phase but before the memory and strategy report phases, which were otherwise identical to Experiment 1A.

First, the practice arithmetic phase served to familiarize the participants with the processing component that would later appear during the complex span trials. Specifically, 20 multiplication problems (e.g., 7 x 4 = 28?) successively appeared on the screen, and participants were instructed to read each problem aloud and decide whether the provided answer was true or false aloud while pressing a corresponding right- or left-hand key on the keyboard. Half of the problems were true, and half were false. Each problem was presented at a fixed pace for 3.5 s with an interstimulus interval (ISI) of 0.1 s. Participants' response times (RTs) and accuracy were recorded, and participants were required to reach an 85% accuracy criterion to progress to the next phase of the experiment.

Participants in Experiment 1B were then instructed and trained in their assigned strategy; those in Experiment 1A proceeded to the memory phase, described next. Specifically, participants

were informed that the remainder of the experiment would comprise a memory task of a series of random words, and that they should implement a "classic strategy"[3] (i.e., the ineffective strategy) or an "elaborative strategy" to try to remember the words. Participants assigned to use the "classic strategy" were instructed, "try your best to read the words and repeat them in your mind over and over again like you would if trying to remember a phone number." Participants assigned to use the "elaborative strategy" were instructed, "try your best to elaborate on the words by creating mental images, group the words meaningfully, and/or creating sentences in your mind." The next page of instructions displayed a visual example of a series of four words successively appearing on a screen (i.e., window, desk, apple, king). For the ineffective group, a thought bubble alongside each of the words showed the words repeating as each was presented, whereas for the elaborative group the thought bubble showed images of the words as they appeared, grouped meaningfully (i.e., an image of a king standing next to a desk with an apple on it next to an open window). Participants then practiced their assigned strategy out loud for 10 words, each successively presented for 3.5 s (0.5 s ISI) so that the experimenter could monitor the participants and ensure understanding of the assigned strategy. The practice phase repeated, with the instructions re-explained, if the participants did not understand or comply with the instructions. The participants were instructed that they should execute their assigned strategy just as they had practiced, but silently in their mind instead during the memory phase. They were also informed that the conditions would be different than the practice, but to still try their best to execute the strategy consistently as they had practiced.[4] The participants received reminders about their assigned strategy before beginning each block of the critical memory task.

[3] Note that the phrase "classic strategy" was used so that the participants assigned to use an ineffective strategy would not think that it was disadvantageous for their performance and therefore try to implement a different strategy.

[4] Due to an unforeseen copy/paste error, one page of instructions for the elaborative group stated that they should repeat the words silently in their mind. However, this only appeared once at the end of the strategy practice phase and the correct instructions were reiterated throughout the rest of the instructions. Given that it was discovered at the conclusion of the study and no participants had appeared to spot the inconsistency, it is unlikely that this error impacted their strategy execution.

Next, all participants completed six blocks of the memory phase of the experiment, with each block comprising one trial of four to-be-remembered words for each task type (i.e., simple span, complex span, and slow span). The order of the three trials was fully counterbalanced across the blocks, with the counterbalance order randomly presented for each participant. Detailed instructions and a practice phase of three trials (one of each task type) preceded the first block, and thereafter a summary of the instructions preceded the subsequent blocks. The practice trials were repeated with the instructions verbally reiterated if the experimenter did not think that the participant had understood the instructions or completed the practice trials incorrectly. To start the trials of all task type, a fixation of "******" appeared on the screen for 1 s. During the simple span trials, the to-be-remembered words were presented successively for 0.9 s (0.1 s ISI). During the complex span trials, each to-be-remembered word was followed by an arithmetic problem (e.g., 7 x 4 = 28?) presented at a fixed rate of 3.5 s (0.1 s ISI) to be read and solved aloud as during the arithmetic practice phase. During the slow span trials, the to-be-remembered words were followed by a blank screen presented for the same duration as the arithmetic problem from the complex span trials (i.e., 3.6 s). In all the trials, participants were instructed to read the words aloud and try to recall them aloud in their original order of presentation at the end of the trial when prompted by the cue to recall (i.e., "????"). Participants' recall was audio-recorded and later transcribed after the session was finished.

After all three trials were presented, a distraction task followed for approximately 3 min wherein participants completed a spatial WM task of trying to remember six of 30 randomly presented dots on the screen.[5] As the task was not relevant to the current study and only served to replicate the procedure of Souza and Oberauer (2017) precisely, the data were only recorded but not analyzed, and thus will not be discussed further. The data can be found on the OSF. Next, the participants completed a delayed recall task, in which they were instructed to freely recall as many of the words as possible that they had seen in the last phase, with no regard to order. Participants typed their recall into the computer, and their responses were echoed back to them on a 3 x 4 grid (i.e., 12 total words to recall

[5] We thank Alessandra Souza for sharing her code for the distraction task.

per block). Spelling mistakes and pluralizing in both experiments were later manually corrected if they were 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").[6] Participants were then offered the opportunity to take a short break before beginning the next block. This sequence of immediate recall, distraction, and delayed recall elements of the memory phase repeated for all six blocks.

Finally, the experiment concluded with a retrospective strategy report phase, wherein the words of each of the memory phase trials were re-presented, and participants were asked to choose one of eight options to indicate how they had tried to remember the words. Following prior work (Dunlosky & Kane, 2007; Loaiza & McCabe, 2013), all the words of each trial were simultaneously presented on the screen in the order they had been presented originally, with the following prompt below them:

How did you originally try to remember the word from the series above?

1 = read each word as it appeared

2 = repeated the words as much as possible

3 = thought back to/directed my attention back to the words

4 = used a sentence to link the words together

5 = developed mental images of the words

6 = grouped the words in a meaningful way

7 = did something else

8 = don't remember

Options 1-2 and 4-7 correspond respectively to passive reading, rote rehearsal, sentence generation, imagery, meaningful grouping, and other, and are the same reported strategies that have been investigated in prior work (e.g., Bailey et al., 2011; Dunlosky & Kane, 2007; Loaiza & McCabe, 2013). Option 3 was meant to represent refreshing and was included as a novel addition to the typical

[6] Spelling mistakes occurred in 0.91% and 0.92% of the recalled items in Experiments 1A and 1B, respectively. The results were similar regardless of whether the mistakes were corrected or not. The typos were corrected in the reported results.

strategy report to examine how often participants report using refreshing as a strategy. Option 8 was included to ensure that participants did not simply guess a strategy if they did not remember having seen the trial at all. All 18 trials of the experiment were re-presented in their original order of presentation.

Data Analysis. All pre-processing steps and analyses were conducted in R using R studio (R core team, 2017). Following prior work (Dunlosky & Kane, 2007; Loaiza & McCabe, 2013), the reported strategies were qualified as ineffective strategies (read and rehearse) or elaborative strategies (sentence, imagery, and grouping). Given the novelty of enquiring about refreshing as a strategy, it was not combined into ineffective or elaborative strategies. The last two possibilities of "did something else" or "don't remember" were tabulated but excluded from analyses. Across both experiments, 3.3% of the strategies were either missing or ambiguous (e.g., "13" suggesting that the participant tried to enter two strategies despite instructions to report only one strategy). These missing and ambiguous values were also excluded from analysis.

We relied on Bayesian inference to assess the evidence for our predictions. Bayesian inference involves updating one's prior beliefs about some parameters of interest in light of the observed data. For example, a couple of parameters of interest in the current work is the effect of task type, such as an advantage of slow and complex span over simple span, and its potential interaction with elaborative strategies. The updated beliefs are the posterior distributions of each of these parameters, which are typically reported with the mean of the parameter estimate and a credibility interval (CI) that gives a sense how certain (often 95%) we can be that the true value of the parameter lies in this range. Furthermore, the CI also allows us to draw inferences about how credibly different from 0 or not that this parameter is: CIs that overlap with 0 are considered null, whereas we can be more confident that there is a true, credible effect when its CI is narrow and far from 0. Accordingly, to draw inferences about our hypotheses, we inspected and report the 95% CIs of the posterior estimates of each effect (e.g., task type, instruction group, reported strategy, and their interactions) and relevant pairwise comparisons, with CIs that do not overlap with 0 being considered credible.

Although we had planned to use Bayesian analysis of variance to address our primary research questions, we opted to instead take an analogous mixed effects approach. The main benefit of this approach is that it allowed us to leverage the heterogeneity across participants and trials rather than aggregate across it, thereby optimizing the power to detect credible effects and providing more certainty regarding observed null effects. The original planned analyses are reported on the OSF for the sake of brevity while still preserving transparency. However, we emphasize that this mixed effects approach is analogous to the aggregate analysis and thus is not an extreme departure from our pre-registered analysis. Furthermore, rather than an aggregate analysis, a mixed effects analysis was better suited to allow the flexibility that participants used different strategies across trials, resulting in uneven cells. Accordingly, we used the brms package (Bürkner, 2018) to fit Bayesian logistic mixed effects models to predict the likelihood of recalling a given item (1 or 0) as a function of our fixed effects (i.e., task type, reported strategy type, and instruction group) and including random effects and slopes of participant. The brms package uses Stan (Stan Development Team, 2018) to estimate posterior distributions of parameters using Monte-Carlo algorithms (Bürkner, 2018). Following prior work (Bartsch et al., 2018; Gelman et al., 2013), we applied weakly informative Cauchy priors (with location 0 and scale 5) on the regression coefficients, intercept, and variance for all the models in the current work. 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 of the four chains by visually inspecting the chains as well as verifying that the $\hat{R}$ statistic was close to 1 for all parameters of all the fitted models. Posterior predictive checks were also conducted to ensure appropriate model fit to the data.

Results and Discussion

Immediate Performance. We first report on immediate recall performance in terms of serial scoring (recall scored as accurate in the correct serial position) and free scoring (recall scored as accurate regardless of original serial position) as a function of task type (Experiments 1A and 1B) and instruction group (Experiment 1B). The results of each model showed credible effects of task type,

with no effect of or interaction with instruction group in Experiment 1B. As evident in Table 1, in all cases, immediate recall from complex span was substantially worse than simple span and slow span. Furthermore, simple span and slow span showed similar immediate recall, except in the case of serial scoring in Experiment 1A where simple span showed credibly greater recall than slow span.

Reported Strategy Use. Table 2 reports the descriptive statistics for each of the specific eight possible options, as well as for the combined strategy types (i.e., overall ineffective and elaborative), the latter of which were used for analysis.[7] Note that we did not analyze reported use or delayed recall as a function of these individual strategies given that this would yield cell sizes with too few observations. Instead, we only focus on the combined ineffective versus elaborative strategies, in line with prior work (e.g., Bailey et al., 2011; Dunlosky & Kane, 2007). We considered the likelihood of reporting ineffective and elaborative strategies as a function of task type (Experiments 1A and 1B) and instruction group (Experiment 1B; see Figure 1). In Experiment 1A, participants were not more likely to report using elaborative strategies during slow span compared to complex span (estimate = -0.41 [-1.12, 0.32]) or simple span (estimate = -0.66 [-1.42, 0.11]), conflicting with our prediction. However, participants did report using ineffective strategies more often during simple span compared to slow span (estimate = 0.59 [0.01, 1.17]). There was no such difference in reported ineffective strategies between slow span and complex span (estimate = 0.21 [-0.37, 0.79]). Participants also did not differ in their reported strategy use between complex span compared to simple span for elaborative (estimate = -0.26 [-0.93, 0.39]) or ineffective strategies (estimate = 0.39 [-0.16, 0.96]).

In Experiment 1B, participants reported that they did not follow their instructed strategy during 46.5% of the trials (ineffective group: 39.4%; elaborative group: 53.7%). However, participants who were instructed to use elaborative strategies were overall more likely to report using them (estimate = 1.44 [0.27, 2.57]) and less likely to report ineffective strategies (estimate = -1.27 [-2.19, -0.34]) compared to participants who were instructed to use ineffective strategies. Moreover, this

[7] Note that we had originally pre-registered that we would include refreshing as a reported strategy type in our analyses. However, given the very low incidence of its reported use, we were advised to exclude it from analysis during the review process. The analyses including refreshing can be found on the OSF.

overall effect of instruction group interacted with the slow span task type (ineffective estimate = -1.21 [-2.54, -0.14]; elaborative estimate = 1.36 [0.17, 2.67]), but not the complex span task type (ineffective estimate = 0.15 [-0.73, 1.00]; elaborative estimate = 0.03 [-0.96, 1.08]). As is evident in Figure 1, the elaborative instruction group reported using elaborative strategies more often during slow span compared to simple span (estimate = 2.15 [1.34, 3.08]) and complex span (estimate = 1.64 [0.83, 2.55]), whereas there was no difference in reported elaborative strategies between simple and complex span (estimate = -0.51 [-1.17, 0.17]). The elaborative instruction group was also less likely to report using ineffective strategies during slow span compared to simple span (estimate = -1.46 [-2.56, -0.58]) and complex span (estimate = -1.45 [-2.57, -0.53]), whereas there was no difference in reported ineffective strategies between simple and complex span (estimate = 0.14 [-0.59, 0.61]). Conversely, the ineffective instruction group showed no credible differences between the tasks for reported elaborative or ineffective strategies.

Taken together, the results regarding reported strategy use suggest that participant do not spontaneously report using elaboration more often during slow span versus simple and complex span when receiving no instructions in Experiment 1A, conflicting with our prediction. However, when participants were instructed to elaboration in Experiment 1B, they did report using it more often during slow span compared to simple and complex span. This divergence in results between those with and without instruction may suggest that participants may be reminded of the different strategies they could use and consequently use elaborative strategies when they are instructed to do so. On the other hand, elaborative strategies may simply have been retrospectively reported for slow span items because they were more memorable overall. We return to this issue later on. It is also noteworthy that participants did not consistently comply with their strategy instructions in Experiment 1B, especially those who were instructed to use elaborative strategies. This poor compliance with instructed strategies means that this variable cannot be taken at face value as a clear characterization of what participants do to remember the items. As we discuss further on, this is important for considering the impacts of instruction group and reported strategies on delayed recall.

Delayed Performance. We report the results of delayed free recall conditionalized on correct immediate recall, henceforth referred to as corrected delayed recall. The results based on overall delayed free recall, regardless of correct immediate recall, can be found on the OSF. In brief, the results were generally similar, with the main exception being that the reported McCabe effects were sometimes smaller or null when using overall delayed recall, as has been observed in previous work (Loaiza & Halse, 2019; Souza & Oberauer, 2017). As Rose and colleagues (2015) have reasoned, long-term differences between conditions could arise simply from baseline differences in WM. In other words, given that immediate recall was greater overall for items studied during simple and slow span, delayed recall of these items may be disproportionately advantaged compared to complex span items due to an overall beneficial influence of testing (Rowland, 2014). Thus, using corrected delayed recall may be a more appropriate measure for drawing inferences about EM given that it corrects for initial baseline differences between the task types. Notwithstanding, we will flag any results that were not consistent between the two measures.

We tested a series of four models: The first model included only task type (simple, complex, slow span) as a predictor. The second and third models further included an effect and interaction with reported strategies (ineffective, elaborative) and instruction group (none, ineffective, elaborative), respectively. The final model included all three predictors (i.e., task type, reported strategies, instruction group). Accordingly, the overall effect of task type in Model 1 is broken down according to reported and instructed strategies in the subsequently tested models. This allowed us to address our primary research question regarding whether reported strategies (Model 2) and instructed strategies (Model 3), uniquely or in combination (Model 4), contribute to or moderate the effects of task type (Model 1) on delayed recall. We also conducted the first two models and all four models separately for Experiments 1A and 1B, respectively, given that the experiments were conducted in succession rather than simultaneously, as explained previously. However, these results are reported on the OSF for the sake of brevity and given their consistency with what is reported here.

Of the four models, the best-fitting was Model 2 that included only effects of task type and reported strategies, compared to the next-best Model 4 that included all three predictors, ΔLOO = 7.1 (SE = 3.9), ΔWAIC = 6.4 (SE = 3.9). The results of the four models are presented in Figure 2, and the pairwise comparisons from Models 2 and 3 are presented in Table 3. The results of Model 1, i.e., the model including only task type, showed that delayed recall of simple span was credibly worse than that of complex span (estimate = -0.88 [-1.13, -0.63]) and slow span (estimate = -0.83 [-1.09, -0.59]), as we predicted and consistent with prior work (McCabe, 2008; Souza & Oberauer, 2017). However, there was no credible difference between complex span and slow span (estimate = 0.05 [-0.20, 0.30]). Including reported strategies in Model 2 yielded a substantial improvement to model fit, ΔLOO = 76.2 (SE = 12.7), ΔWAIC = 76.6 (SE = 12.7). There was an overall benefit of reported elaborative over ineffective strategies (estimate = 0.69 [0.32, 1.08]). Furthermore, the McCabe effect remained (estimate = 0.69 [0.41, 0.98])[8] and did not interact with reported strategies (estimate = 0.32 [-0.15, 0.78]), such that the effect was observed regardless of whether participants reported using ineffective (estimate = 0.69 [0.41, 0.98]) or elaborative (estimate = 1.01 [0.62, 1.39]) strategies. However, delayed recall from complex span was substantially greater when participants reported using elaborative versus ineffective strategies (estimate = 1.01 [0.63, 1.41]). The slow span effect also remained (estimate = 0.35 [0.06, 0.64]), but unlike the McCabe effect, it did interact with reported strategies (estimate = 0.69 [0.26, 1.13]): The slow span effect was stronger when participants reported using elaborative (estimate = 1.05 [0.68, 1.41]) versus ineffective strategies (estimate = 0.35 [0.06, 0.64]). It should be noted that this represents an ordinal interaction, and thus we refrain from interpreting it too much (see Loftus, 1978; Wagenmakers et al., 2012). Furthermore, like complex span, recall from slow span was greater when participants reported using elaborative versus ineffective strategies (estimate = 1.39 [1.01, 1.77]). There was a credible advantage for complex span over slow span when

[8] Note that the McCabe effect in Model 2 was null when using overall delayed recall, consistent with the aforementioned work that baseline differences in immediate recall mitigate the advantage of complex span over simple span (Loaiza & Halse, 2019; Souza & Oberauer, 2017).

participants reported using ineffective strategies (estimate = 0.34 [0.03, 0.66]), whereas performance was similar between tasks for reported elaborative strategies (estimate = -0.04 [-0.43, 0.34]).

Model 3 including effects of task type and instruction group provided the worst fit compared to Model 2, ΔLOO = 76.5 (SE = 12.8), ΔWAIC = 76.8 (SE = 12.8). First, there was no overall effect of instructed elaborative (estimate = 0.22 [-0.27, 0.68]) or ineffective strategies (estimate = 0.36 [-0.09, 0.80]) relative to receiving no strategy instructions (i.e., Experiment 1A). Furthermore, there was a credible McCabe effect (estimate = 1.16 [0.79, 1.53]) that did not interact with instructed elaborative strategies (estimate = -0.17 [-0.73, 0.40]), but it did interact with instructed ineffective strategies (estimate = -0.70 [-1.27, -0.12]). This again appears to be an ordinal interaction: The pairwise comparisons revealed that the McCabe effect was strongest for participants who did not receive any strategy instructions (i.e., Experiment 1A; estimate = 1.16 [0.79, 1.53]) and for the instructed elaborative strategy group (estimate = 0.99 [0.57, 1.43]), but slightly smaller for the instructed ineffective strategy group (estimate = 0.46 [0.02, 0.92]). It is worth noting that there were no credible differences between the three instruction groups in their recall from complex span. There was also a credible slow span effect (estimate = 0.82 [0.42, 1.22]) that did not interact with instructed elaborative (estimate = 0.53 [-0.04, 1.10])[9] or ineffective (estimate = -0.50 [-1.06, 0.05]) strategies. The pairwise comparisons, however, revealed that the slow span effect was credible for participants with no strategy instructions (i.e., Experiment 1A; estimate = 0.82 [0.42, 1.22]) and those instructed to use elaborative strategies (estimate = 1.35 [0.93, 1.80]), but not for those instructed to use ineffective strategies (estimate = 0.32 [-0.10, 0.72]). There were no credible differences between complex and slow span for any of the instruction groups. The final fourth model showed similar results to the previous models, such that there was an overall benefit of reported elaborative strategies (estimate = 0.65 [0.06, 1.26])[10] and a credible McCabe effect (estimate = 1.03 [0.60, 1.47]) that interacted with

[9] When using overall delayed recall, the interaction between the slow span effect and elaborative strategies was credible (estimate = 0.60 [0.04, 1.21]). This inconsistency likely reflects issues of low power, as we later discuss.
[10] The benefit of reported elaborative strategies was not credible when using overall delayed recall (estimate = 0.54 [-0.02, 1.14]).

instructed ineffective strategies (estimate = -0.74 [-1.40, -0.09]). However, the overall slow span effect was no longer credible (estimate = 0.32 [-0.15, 0.78]), but did interact with reported elaborative strategies as in Model 2 (estimate = 0.81 [0.06, 1.54]). No other main effects or interactions were credible. As is clear in the last panel of Figure 2, there was much greater variability in the results of Model 4 given the fewer observations, and given that Model 2 was superior to Model 4, we refrain from interpreting these results too much.

Results Summary. These results suggest that elaborative strategy use was not more common during slow span compared to the other tasks (Experiment 1A), against our prediction, unless participants were explicitly instructed to use elaborative strategies (Experiment 1B). Furthermore, participants in Experiment 1B often reported strategies that were not consistent with their group assignment. This issue of compliance perhaps led to Model 2 providing the best fit particularly compared to Model 3, suggesting that spontaneous strategies are more influential than instructed strategies, although the pattern of results was similar between the two models. There was an overall benefit of spontaneous elaboration to EM, but the predicted benefit of instructed elaboration was not credible. Most importantly, the McCabe and slow span effects were generally consistent regardless of spontaneous and instructed strategies, although there was an overall benefit of elaboration for both tasks. Interestingly, there was no advantage of slow span over complex span for reported elaborative strategies, and an advantage of complex span over slow span for ineffective strategies. Thus, the results so far suggest that elaboration contributes to both the slow span and McCabe effects, consistent with the time-in-WM-for-elaboration hypothesis. However, that these effects are largely evident even during reported or instructed ineffective strategies suggests that elaboration may not exclusively contribute to the long-term advantages of the time items spend in WM.

There were several important limitations of Experiment 1 that preclude definitive interpretation of these results, however. First, there were clearly issues of power, such that there were fewer observations per cell of the design as the number of predictors in the model increased (e.g., Figure 2, Model 4), thereby limiting strong interpretation of the reported null effects. This issue

of power is exacerbated by the fact that there were fewer possible observations due to low delayed recall, and fewer still when using corrected delayed recall given that items not immediately recalled (most often from complex span) were omitted from analysis. Finally, participants reported their strategies retrospectively, which could introduce bias that prevents strong interpretation of the influence of reported strategies. Although previous research has validated the use of retrospective reports (Dunlosky & Kane, 2007), it is possible that participants may have simply reported more elaborative strategies for items that they remember better over the long-term. These limitations thus preclude strong interpretation of the aforementioned point that reported strategies more strongly contributed to EM than instructed strategies, and, most importantly, present issues for interpreting the differential influence of elaboration on the McCabe and slow span effects. Experiment 2 was conducted to address each of these issues.

Experiment 2

In order to address the aforementioned limitations of Experiment 1, we conducted Experiment 2 with a very similar design and rationale, except for the following: First, we randomly assigned the participants to one of the three instruction groups (none, instructed elaborative, instructed ineffective) and increased the number of participants in each group per experiment to achieve greater power. We also removed “refreshing” and “don’t remember” as possible strategies to report given their low use and to increase the number of possible useable observations. Thus, the strategy report in Experiment 2 matched prior research (e.g., Bailey et al., 2011; Dunlosky & Kane, 2007). Second, rather than overt, self-generated immediate and delayed recall, we used reconstruction, wherein participants recalled the four presented items in each trial by selecting them from eight possible options presented on the screen. This reconstruction paradigm has several important advantages: Its use increases overall recall, thereby increasing observations and consequently power while mitigating baseline differences in immediate recall between task types, as well as allows for a consistent retrieval method to be used between immediate and delayed recall. We have also successfully used this reconstruction paradigm in recent prior work, demonstrating a

McCabe effect (Loaiza et al., 2020), thus justifying its use in the current experiment. Finally, we considered whether the use of retrospective strategy reports impacted our results by using concurrent strategy reports. We conducted the experiment online given the greater number of participants required as well as due to the suspension of in-lab testing during the coronavirus pandemic.

As in Experiment 1, we conducted Experiment 2 to determine whether the McCabe and slow span effects in delayed recall are moderated by spontaneous and instructed elaboration in WM. Our pre-registered predictions were very similar to Experiment 1: Regardless of concurrent or retrospective reporting, we expected that participants would be more likely to report using elaboration during slow span compared to simple and complex span trials, and that this relatively greater use of elaboration (whether spontaneous or instructed) underlies the slow span effect in delayed recall. That is, the advantage of slow span over simple span should only be evident when participants use elaboration, whereas it should be reduced or null when participants use ineffective strategies. Conversely, the McCabe effect should be less impacted by reported or instructed strategy use if covert retrieval, and not elaboration, underlies the effect. If, however, the McCabe effect is only evident when participants use elaborative strategies, then this would negate our covert retrieval account and instead suggest that time in WM affords the opportunity for elaboration that promotes later EM performance, irrespective of the type of task (slow or complex span).

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, with normal or corrected-to-normal vision, with no history of cognitive impairment, who were using a desktop/laptop, and who had not taken part in a similar experiment (Loaiza et al., 2020) were able to sign up for the study. The experiment lasted about 30-45 min for most participants, and they were compensated with £5.00.

Given the aforementioned issues of power for Experiment 1, we decided to conduct an a-priori power analysis from a Bayesian perspective based on the observed McCabe effect from our

recent prior work (Loaiza et al., 2020). To our knowledge there is no principled way to estimate power for Bayesian logistic mixed effects modeling, and thus we estimated power using Kruschke and Meredith's BEST (2020) R package and a simulated studies method (see OSF for details). The predicted power from these respective methods was 0.82 and 0.91 with 40 participants and 72 items (i.e., 4 memoranda per 18 trials[11]). Thus, we aimed to recruit a minimum sample of at least 24 participants per group (i.e., 72 in total) in line with Experiment 1 and similar prior work (e.g., McCabe, 2008; Souza & Oberauer, 2017), but we assumed that we would need at least 40 participants per group. We checked the results after the first 24 valid datasets per group and continued checking until reaching at least 40 valid datasets per group, allowing up to a maximum of 60 valid datasets per group in case the results were still unclear even after 40 participants per group. Although it was not necessary to continue sampling past our planned sample size or engage in optional stopping, it should be noted that Bayesian inference is considered immune to changes in sampling plan (Rouder, 2014).[12]

In total, 127 participants (M_{age} = 26.79, *SD* = 4.91) were randomly assigned to one of three instruction groups: none (*n* = 41), ineffective (*n* = 41), and elaborative (*n* = 45). The data of an additional 13 participants were excluded for quitting in the middle of the experiment (usually during the practice arithmetic phase). The remaining task factor of task type (simple, complex, or slow span) was manipulated within-subjects. Like Experiment 1, reported strategy use (ineffective or elaborative) was both a measured dependent variable as well as a factor predicting delayed performance. The principal dependent variables were free and serial scoring at both the immediate and delayed tests.

Materials and Procedure. Experiment 2 was programmed with Inquisit (2018). Given the increased number of items for the design, a list of 360 words was developed that was similar to

[11] Although we had planned 18 trials as in the previous experiment, a brief pilot suggested that delayed performance could reach ceiling, and thus we increased the number of trials to 30 total in this experiment.

[12] To briefly explain why this is the case for Bayesian inference but not traditional null-hypothesis testing: If the null hypothesis is true, a researcher will eventually arrive at a significant *p*-value when continuing to collect data past the planned sample size given that *p* values are uniformly distributed under the null. Conversely, under Bayesian methods of hypothesis testing (e.g., Bayes factors), the support for the null will continue to grow as more data are collected (see e.g., Etz et al., 2018; Schönbrodt et al., 2017, for further discussion).

Experiment 1 (letters: *M* = 5.45, *SD* = 0.92, range = 4-8; syllables: *M* = 1.55, *SD* = 0.50, range = 1-2; log HAL frequency: *M* = 9.28, *SD* = 0.91, range = 8.00-12.60).

The study was advertised on Prolific with a general description of the task and advised participants that they should be prepared to do the experiment in one continuous sitting in a quiet, distraction-free environment. Participants were also advised that they must carefully read and follow the instructions and that they could view their general performance at the end of the experiment in order to increase interest and motivation in the study. After signing up, participants installed the Inquisit plugin that forced the experiment to fill the screen, thereby preventing engagement in other tasks on their computers during the experiment.

The remaining procedure of Experiment 2 was very similar to that of Experiment 1. Participants first completed the practice arithmetic phase, followed by instructions regarding the strategy they should implement during the critical task if they had been randomly assigned to the instructed ineffective or elaborative strategy groups. Like Experiment 1, participants in the instructed ineffective or elaborative strategy groups were instructed to use their assigned strategy during the upcoming memory task and were shown a visual example of how they should implement the strategy. Given that the experiment took place online, it was not possible for an experimenter to check their implementation of the strategy, and so we decided to remove the strategy practice session of Experiment 1. However, the instructions strongly and regularly emphasized that it was very important to the study that they used their assigned strategy. Participants assigned to receive no instructions proceeded directly to the critical task instructions, which were identical for all three groups.

During the critical task, there was one block of 30 trials, 10 trials of each task type (simple, complex, and slow span), randomly intermixed. Unlike the six blocks of trials of Experiment 1, the change in retrieval method required us to mass the trials into one block so that performance was not too high during the delayed test. Like Experiment 1, participants were instructed to read each word out loud only one time as it appeared and to try to remember them. They were also instructed to read and respond to the arithmetic problems as quickly and accurately as possible. During immediate

reconstruction, the four presented words were randomly arranged among four never-presented lures in two rows of four frames with the instruction "use the mouse to select the 4 presented words in their original order" above the frames.[13] After selecting four items, participants then completed a concurrent strategy report: a screen appeared asking them how they tried to remember the words in the last trial, with 6 options possible (1 = read each word as it appeared; 2 = repeated the words in my mind as much as possible; 3 = used a sentence to link the words together in my mind; 4 = developed mental images of the words; 5 = grouped the words in my mind in a meaningful way; 6 = did something else). As in Experiment 1, the last option was tabulated (see Table 2) but excluded from analysis. After selecting their response and clicking a "continue" button, the next trial began. Participants were offered the opportunity to take a short break after 10 and 20 trials.

After completing the block, participants completed a 3 min distraction phase, which entailed making symmetry judgments for 8 x 8 grids, with some squares filled in black or white to make a pattern that was symmetrical or not about their vertical axis (taken from Foster et al., 2015; Kane et al., 2004). The change in distraction task from Experiment 1 was merely an issue of convenience given that programming the equivalent in Inquisit was not possible and because the distraction task is not important to the research questions. The task was at least similar to Experiment 1 in the sense that it was visuospatial, although we emphasize that there is no theoretical reason why the nature of the distraction task should be important to the current research questions. After completing the distraction phase, participants received instructions for the delayed reconstruction test. All the trials of the previous task were presented again in a new random order, each comprising the four originally presented words randomly arranged among four never-presented lures in two rows of four frames. As during the immediate reconstruction test, participants were instructed to recall the words in their

[13] Note that, unlike Experiment 1, it was possible to select the same item more than once due to constraints in the program. This only occurred 1.39% and 0.49% of the time during the immediate and delayed tests, respectively. These instances were corrected so that only the first instance of the repeated selection was counted and not marked correct more than once.

original order. Thereafter, participants completed an instruction compliance and demographics survey, followed by the chance to view their overall performance and a debriefing of the experiment.

It is important to note that the task regularly emphasized the importance of following the instructions, and during the WM phase, participants were warned that they would be sent back to the practice arithmetic phase if their responses were not registered. Twenty-nine participants received a first warning during the critical task, and a further 28 participants returned to the practice arithmetic phase once during the block for continuing to not respond to the arithmetic problems after receiving the first warning. Furthermore, at the conclusion of the experiment, the instruction compliance survey enquired whether the participants read and answered the arithmetic problems aloud, read the words aloud only once, and completed the experiment in one sitting in a quiet, distraction-free environment. The instructions made it clear that the participants should answer honestly and that their answers to these questions would not impact their pay. Although most participants reported compliance, 42 participants reported not following at least one of these instructions. The pattern of results was similar when excluding these participants.

Data Analysis. We followed the same analytic approach as in Experiment 1 of fitting Bayesian logistic mixed effects models with the brms package in R.

Results and Discussion

Immediate Performance. We first report on immediate performance in terms of serial scoring and free scoring as a function of task type and instruction group. As in Experiment 1, the results of each model showed credible effects of task type, with no effect of or interaction with instruction group. Overall, immediate reconstruction from complex span was worse than simple span and slow span, with largely no credible differences between simple span and slow span (see Table 1).

Reported Strategy Use. As in Experiment 1, we considered the likelihood of reporting ineffective and elaborative strategies as a function of task type and instruction group (see Table 2 and Figure 3). Different to Experiment 1, the use of concurrent reports in Experiment 2 seemed to reduce the variability in reported strategies in participants who did not receive any instructions as well as

reduced the overall likelihood that the instructed participants reported strategies inconsistent with their group assignment (29% of the time overall; 9% of the time in the ineffective instruction group, 49% of the time for elaborative instruction group).

As is strikingly evident in Figure 3, there was a strong credible effect of instruction group: Compared to those receiving no or ineffective instructions, participants who received elaborative instructions were, overall, more likely to report elaborative strategies (versus none: estimate = 2.85 [1.16, 4.61]; ineffective: estimate = 5.49 [3.59, 7.71]) and less likely to report ineffective strategies (versus none: estimate = -2.73 [-4.43, -1.06]; ineffective: estimate = -4.80 [-6.72, -2.94]). Conversely, there was no credible difference between participants who received no instructions versus ineffective instructions overall in terms of overall reported ineffective strategies (estimate = 1.34 [-0.36, 3.15]), but participants were less likely to report using elaborative strategies overall in the ineffective instruction group compared to the no instruction group (estimate = -2.03 [-4.25, -0.17]).

Most importantly, for both reported strategy types, there was a credible effect of slow span (reported ineffective: estimate = -0.87 [-1.47, -0.25]; reported elaborative: estimate = 0.89 [0.21, 1.48]) and interaction with the elaborative instruction group (reported ineffective: estimate = -1.18 [-2.25, -0.17]; reported elaborative: estimate = 1.20 [0.14, 2.34]). The pairwise comparisons revealed that participants who received no instructions and elaborative instructions were more likely to report elaborative strategies (none: estimate = 0.89 [0.21, 1.48]; elaborative: estimate = 2.09 [1.26, 3.04]) and less likely to report ineffective strategies (none: estimate = -0.87 [-1.47, -0.25]; elaborative: estimate = -2.05 [-2.91, -1.23]) during slow span compared to simple span. As in Experiment 1, there were no credible differences in reported strategies between the task types for the ineffective instruction group. Although there was no such effect of or interaction with complex span in the omnibus analyses, the pairwise comparisons revealed that the elaborative group was less likely to report ineffective strategies (estimate = -1.36 [-1.98, -0.76]) and more likely to report elaborative strategies (estimate = 1.30 [0.73, 1.86]) during complex span compared to simple span. Finally, there

were no credible differences in reported strategies between complex span and slow span for any of the instruction groups.

Overall, the results of reported strategy use demonstrated that participants were overwhelmingly likely to report using ineffective strategies, unless they were specifically instructed to use elaborative strategies. However, like Experiment 1B, participants complied with this elaborative instruction only about half the time overall, and their compliance further depended on the task type, such that participants were much more likely to report using elaborative strategies during slow and complex span compared to simple span. Most importantly, although participants without any instructions reported using elaborative strategies more often during slow span compared to simple span, there were no differences between slow span and complex span in any of the instruction groups, particularly in their reported elaborative strategy use. This conflicts with our hypothesis that participants should use elaborative strategies most often during slow span compared to both simple and complex span, whether spontaneously or when directed to do so.

Delayed Performance. Finally, we report the delayed performance results. The use of the reconstruction test mitigated the influence of baseline differences between task types, yielding immediate performance that was much higher overall in Experiment 2 than Experiment 1 (see Table 1). However, immediate performance from complex span was still characteristically lower overall than simple and slow span, and for the sake of consistency between experiments, we once again report the analyses using corrected delayed performance. The results based on overall delayed performance can be found on the OSF, and the pattern of results was the same regardless of the measure. Furthermore, like immediate performance, the reconstruction method at delay allowed us to consider both free and serial scoring measures of delayed performance, displayed in Figures 4 and 5, respectively.

Like Experiment 1, we tested a series of four models: The first included only an effect of task type, the second and third models further included an effect and interaction with reported strategies and instruction group, respectively, and the fourth model included all three variables (i.e., task type, reported strategies, and instruction group). This allowed us to address our primary research question

regarding whether reported strategies (Model 2) and instructed strategies (Model 3), uniquely or in combination (Model 4), contribute to or moderate the effects of task type (Model 1) on delayed performance. As in Experiment 1, the best-fitting model of the four was Model 2 that included only effects of task type and reported strategies, compared to the next-best Model 4 that included all three predictors (free: ΔLOO = 15.7 [SE = 3.5], ΔWAIC = 15.4 [SE = 3.4]; serial: ΔLOO = 1.8 [SE = 4.2], ΔWAIC = 1.3 [SE = 4.2]). Model 2 also substantially outperformed Model 3 that included only effects of task type and instruction group (free: ΔLOO = 27.7 [SE = 7.8], ΔWAIC = 27.7 [SE = 7.8]; serial: ΔLOO = 73.2 [SE = 13.2], ΔWAIC = 73.5 [SE = 13.2]) and Model 1 that only included an effect of task type (free: ΔLOO = 20.1 [SE = 7.8], ΔWAIC = 20.3 [SE = 7.8]; serial: ΔLOO = 74.0 [SE = 13.1], ΔWAIC = 74.3 [SE = 13.1]).

As evident in Figures 4 and 5, delayed simple span performance was worse than that of complex span (free: estimate = 0.33 [0.22, 0.44]; serial: estimate = 0.37 [0.22, 0.51]) and slow span (free: estimate = 0.49 [0.37, 0.62]; serial: estimate = 0.58 [0.40, 0.74]) in Model 1. This replicates the McCabe and slow span effects of Experiment 1 and from prior research (McCabe, 2008; Souza & Oberauer, 2017), and further demonstrates that these effects are observable online and using a reconstruction retrieval method (Loaiza et al., 2020). Different to Experiment 1, there was also an advantage of slow span over complex span (free: estimate = 0.17 [0.04, 0.30]; serial: estimate = 0.21 [0.07, 0.35]). The improved fit of Model 2 suggests that reported strategies substantially contributed to this effect of task type. As in Experiment 1, there was an overall benefit of reported elaborative over ineffective strategies (free: estimate = 0.43 [0.18, 0.69]; serial: estimate = 0.58 [0.30, 0.87]). Furthermore, the McCabe effect remained (free: estimate = 0.25 [0.14, 0.37]; serial: estimate = 0.35 [0.20, 0.50]) and did not interact with elaborative strategies (free: estimate = 0.10 [-0.17, 0.37]; serial: estimate = -0.14 [-0.43, 0.15]). In free scoring, the effect occurred regardless of whether participants reported using ineffective (estimate = 0.25 [0.14, 0.37]) or elaborative (estimate = 0.36 [0.12, 0.59]) strategies, just as in Experiment 1. In serial scoring, however, although the interaction was not credible, the pairwise comparisons revealed that the McCabe effect was only credible when participants reported using ineffective strategies (estimate = 0.36 [0.20, 0.50]), but not when using

elaborative strategies (estimate = 0.22 [-0.05, 0.47]). The slow span effect also remained in Model 2 (free: estimate = 0.34 [0.21, 0.47]; serial: estimate = 0.43 [0.26, 0.60]). For free scoring, the slow span effect interacted with elaborative strategies (estimate = 0.29 [0.01, 0.59]) but not for serial scoring (estimate = 0.25 [-0.05, 0.54]). However, the pairwise comparisons revealed that the slow span effect was evident in both elaborative (free: estimate = 0.63 [0.37, 0.89]; serial: estimate = 0.68 [0.40, 0.94]) and ineffective strategies (free: estimate = 0.34 [0.21, 0.47]; serial: estimate = 0.43 [0.26, 0.60]). Thus, the interaction was ordinal as in Experiment 1 because the slow span effect was simply stronger with elaborative versus ineffective strategies. Finally, the slow span advantage over complex span from Model 1 was only evident for reported elaborative strategies (free: estimate = 0.27 [0.003, 0.54]; serial: estimate = 0.47 [0.22, 0.71]), whereas delayed performance was similar between complex and slow span when participants reported using ineffective strategies (free: estimate = 0.08 [-0.05, 0.22]; serial: estimate = 0.08 [-0.08, 0.23]).

For the sake of brevity, and given that model fit was substantially worse for Model 3 and very few observations in some cells of Model 4, we will not continue to report on these remaining models. In brief, the pattern of results for Model 3 was similar to that of Model 2, as in Experiment 1 (see Table 3). The interested reader can find the full analyses on the OSF.

Results Summary. The overall results of Experiment 2 reinforced many of the results of Experiment 1. First, participants reported elaborative strategies more often during slow span compared to simple span, both spontaneously (i.e., with no instructions) and when instructed to use elaborative strategies. However, participants were also just as likely to report using elaborative strategies during complex span as slow span, spontaneously and when instructed to do so, thus conflicting with our prediction that slow span affords a disproportionate opportunity to engage in elaboration compared to both of the other tasks. Furthermore, we again replicated the McCabe and slow span effects in both free and serial scoring measures of delayed recall, and, different to Experiment 1, also observed a further advantage of slow span over complex span, replicating Souza and Oberauer (2017). Importantly, reported strategies once again had an overall effect on delayed

performance and substantially contributed to these task effects: An ordinal interaction between the slow span effect and elaborative strategies suggested that the slow span effect was evident regardless of reported strategies, but stronger for elaborative strategies. The McCabe effect was observed regardless of reported strategies in free scoring, but was only evident for reported ineffective strategies in serial scoring. Moreover, different to Experiment 1, the advantage of slow span over complex span was only evident for elaborative strategies, but there was similar recall between the tasks for ineffective strategies. As we go onto discuss, these results overall converge with those of Experiment 1 in suggesting that elaborative strategies do seem to play a role in the McCabe and slow span effects, but the fact that both were still observed during reported ineffective strategies suggests that elaboration does not completely account for their effects. Moreover, the increased power and methodological adjustments of Experiment 2 gives more certainty to this conclusion converging between the two experiments.

General Discussion

The current experiments investigated the long-term advantages of covert retrieval and elaboration in WM by considering the impact of instructed and spontaneously adopted strategies on delayed memory performance of items studied and recalled from WM. We adjudicated between the covert retrieval account (McCabe, 2008) and an elaboration version of the time-in-WM hypothesis (Souza & Oberauer, 2017) for the observed advantages of complex span and slow span over simple span at delay. Participants were instructed to use ineffective or elaborative strategies or received no specific instructions, and they reported on their strategies retrospectively at the end of the entire memory phase (Experiment 1) or concurrently after each trial (Experiment 2). The rationale of the experiments was to determine whether elaborative strategy use similarly accounts for the advantage of complex span and slow span over simple span, thereby suggesting a parsimonious explanation of why time in WM may promote later retrieval from long-term EM. We first overview the most important results concerning our predictions regarding EM performance and strategy use, and then further discuss the theoretical implications of the current work.

Episodic Memory Performance

The results of both experiments showed an overall long-term benefit of spontaneous elaboration in WM, but conflicting with our predictions, participants were not more likely to report using elaborative strategies during slow span compared to complex span. Most importantly, the results of both experiments showed overall long-term advantages over simple span for both complex span (i.e., a McCabe effect; McCabe, 2008) and slow span (i.e., a slow span effect; Souza & Oberauer, 2017), and generally regardless of whether participants reported using or were instructed to use ineffective or elaborative strategies. The slow span effect appeared to be slightly weaker when participants reported using or were instructed to use ineffective versus elaborative strategies. However, as this was an ordinal interaction, it is not possible to say for certain that this reflects an even greater contribution of elaboration to the slow span effect or perhaps simply reflects a mere issue of measurement scale (see Loftus, 1978; Wagenmakers et al., 2012 for further discussion). At least in Experiment 2, an advantage of slow span over complex span was evident only when participants reported using or were instructed to use elaborative strategies, thus suggesting that elaboration may drive the slow span advantage over complex span. Further work will be necessary to investigate this, but the current findings at least suggest that elaboration contributes to both the McCabe and slow span effects, consistent with the elaboration version of the time-in-WM hypothesis. Notwithstanding, the fact that the McCabe and slow span effects were still both observed even with ineffective strategies suggests that there may be other factors besides elaboration that underlie these effects. Still, our primary prediction that elaboration should exclusively account for the slow span effect was contradicted, and instead suggests that spontaneous elaborative strategies promote long-term retention regardless of task type.

It is further noteworthy and encouraging that this general pattern of results was observed with two different retrieval methods (i.e., self-generated recall in Experiment 1 and reconstruction in Experiment 2) and two different methods of reporting strategies (i.e., retrospective in Experiment 1 and concurrent in Experiment 2). This indicates that the McCabe and slow span effects, as well as the

contributions of elaboration to them, are generalizable across different types of retrieval and strategy reports and can be observed regardless of whether participants are tested in the lab or online.

Instructed versus Spontaneous Strategies in Working Memory

Instructing participants to use specific strategies or not and then having participants report their actual strategies, either retrospectively or concurrently, also allowed an interesting investigation into the potential distinction between instructed and spontaneously adopted strategies. Although we had no prior assumptions about any fundamental differences between these methods, the results of both experiments indicated that spontaneous strategies in WM was a more important factor than instructed strategies for understanding retrieval from EM. That is, a model including an effect of reported strategies alongside the effect of task type better fit the data compared to a model including an effect of instructed strategies. This may have occurred at least partly because participants were not always compliant with their strategy instructions, and allowing for spontaneous strategy adoption may have better captured the variability in the data. However, it is important to note that the conclusions from the models were generally consistent (see Table 3). Thus, although instructed strategy compliance was inconsistent, particularly for those instructed to use elaboration, the results generally suggested that participants may spontaneously use elaborative strategies more often when it is possible to do so during slow and complex span, thereby contributing to their long-term effects.

Participants largely reported using ineffective strategies when instructed to do so and when receiving no specific instructions, the latter of which replicates prior work (Bailey et al., 2011; Dunlosky & Kane, 2007). Furthermore, participants' overall reported compliance with their assigned instructions depended on task type, such that participants were less likely to use elaboration during simple span compared to slow span (Experiment 1B and 2) and complex span (Experiment 2). In Experiment 1B, participants were also less likely to report using elaborative strategies during complex span compared to slow span when instructed to do so. However, the use of retrospective reports may have simply biased participants toward reporting more elaborative strategies for items that they remember better over the long-term. That is, participants may have reflected on the fact that they recalled more slow

span items and found the other items less memorable, and thus circularly reasoned that this is because they used effective strategies in a kind of self-fulfilling prophecy. The use of concurrent reports in Experiment 2 allowed us to mitigate the potential influence of this bias as well as make aware participants with no instructions of the different strategies that they could implement. Indeed, reported elaboration was similar between slow and complex span in Experiment 2, and hence the discrepancy in Experiment 1B likely arose due to the use of retrospective reports. These findings suggest that future research concerning strategy use in WM should consider enquiring about participants' spontaneously adopted strategies, even when instructing specific strategies, and preferably using concurrent rather than retrospective reports.

All this said, there are some limitations to considering reported strategies as a variable predicting memory performance. First, if considered in combination with instructed strategies in the same model (e.g., Model 4 in the current experiments), there will likely be situations in which cell size is very small for some combination of conditions. For example, as we observed in the current experiments, participants who are instructed to use ineffective strategies may rarely report using elaborative strategies, making this specific cell very small. Besides the resulting variable performance causing difficulties for ascertaining reliable conclusions, it is also challenging to interpret what it would mean for other factors like task type to vary according to the combined versus mismatched influence of instructed and reported strategies. For example, imagine if the McCabe effect were evident when participants were instructed to use ineffective strategies but reported using elaborative strategies, but not vice versa. It would not be clear what to conclude regarding the effect of strategies, and thus the two methods in combination could add an unnecessary layer of complexity. Thus, the unfortunate byproduct of this sort of design is that it necessarily limits the cell size of some conditions, which could be further problematic for some potential interpretations. At least the results were consistent regardless of whether reported or instructed strategies were included in the model, even if the model including reported strategies provided the better fit. Furthermore, strategy reports may at least serve

as a manipulation check of how well participants complied with their strategy instructions, and relatively low compliance may warrant their use as a predictor as in the current work.

Relatedly, the limited instances of reported refreshing as a strategy in Experiment 1 may not necessarily suggest that participants are not refreshing, but it may be more likely that they cannot distinguish refreshing from other purported maintenance mechanisms such as rehearsal or even covert retrieval. Consequently, drawing inferences from EM performance as a function of refreshing in the different task types would not be advisable given that the results would more likely be attributable to potential lay misunderstandings of what refreshing is or simply noise in the data. Instructing refreshing as a strategy would perhaps be a more fruitful method of investigating the impact of its use (e.g., Bartsch et al., 2018, 2019; Camos et al., 2011; Loaiza & Camos, 2018), or very thorough training to help participants understand what it means so that they can report any spontaneous use appropriately. For example, Loaiza and Camos (2018, Experiment 2) showed that the benefit of semantic cues when participants forgot information from WM was evident when participants were instructed to use refreshing rather than rehearsal as a maintenance mechanism. Such dissociations on the basis of instructed maintenance strategy suggest that refreshing can be successfully instructed, with informative findings for the field. At the same time, this instruction method is limited to show evidence for the existence of spontaneous refreshing and its effects on memory. This remains an outstanding quandary for the field.

Theoretical Implications

There is a long tradition of research that has investigated the impact of different mnemonic strategies on memory performance, with theoretical and applied interests centered on understanding the best methods to improve long-term retention (e.g., Cepeda et al., 2006; Karpicke & Roediger, 2008; Slamecka & Graf, 1978). There has been a recent return to the historical notion that time predicts long-term retention (Atkinson & Shiffrin, 1968), which at one time was considered debunked, particularly by the highly influential levels-of-processing framework (Craik & Lockhart, 1972; Craik & Tulving, 1975). As Craik (2002) reviews, the levels-of-processing framework was developed partly as

an alternative functionalist perspective of memory as a process rather than the structuralist view of transferring memory traces between a series of stores, as in the popular dual store models of the day (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974; Broadbent, 1958; Murdock, 1967; Waugh & Norman, 1965). Rather than there being separate "boxes" representing independent memory systems through which memory traces are passed, Craik and Lockhart (1972) argued that the persistence of those memory traces depends on the depth of their analysis. Shallowly processed information is likely to be forgotten quickly, whereas deeper, semantic processing yields more durable memory traces. Thus, the apparent beneficial effect of time spent processing items in WM may simply reflect that items processed for a short period of time are likely processed more shallowly and ineffectively compared to items that are afforded more time for elaboration.

Perhaps at some level this historical debate has replayed here, but with a twist: Although there is recent evidence that time spent processing items in WM is indeed important to their later retrieval (Hartshorne & Makovski, 2019; Jarjat et al., 2018; Souza & Oberauer, 2017), the current work demonstrates evidence that time in WM at least partly allows greater opportunity for elaboration. If long-term retention was driven by simply the total time items spent in WM, then it would not matter what type of strategy was instructed or spontaneously adopted, as the total time interleaving the processed items was matched between complex span and slow span in the current experiments. However, the fact that McCabe and slow span effects were still observed even when participants reported ineffective strategies suggests that time in WM may allow for other factors besides elaboration. These factors could include a range of other proposed WM processes, such as consolidation (Cotton & Ricker, 2020; Ricker et al., 2018), attentional refreshing (Camos et al., 2018; Johnson, 1992), or covert retrieval (McCabe, 2008). Compared to the strategy instruction and reporting methods used here, it is unfortunately not as straightforward to directly observe the impact of these different processes without relying on indirect inferences from behavioral data that may be just as adequately explained by alternative competing accounts. Furthermore, it may be the case that

these processes differently contribute to the McCabe and slow span effects, as we have tried to investigate in the current work regarding the influence of elaboration for these effects.

Still other possibilities for long-term effects of items processed in WM could include factors that may constrain or even have little to do with the processing of items in WM. For example, in the current work we used a constant set size of four memoranda per trial, as in similar previous work (e.g., Loaiza & McCabe, 2012; McCabe, 2008; Souza & Oberauer, 2017). It may be the case that the propensity to use or the effectiveness of elaboration depends on such parameters of the task, such that increasing the memory load could encourage participants to adopt elaborative strategies. Although further work will be required to investigate this issue, there is at least some relevant prior research speaking to this issue. First, a consistent McCabe effect has been observed even when varying the number of memoranda between two and four items per trial (Loaiza & Halse, 2019; McCabe, 2008). However, this is obviously a limited, sub-span range, which we similarly adopted here according to McCabe's (2008) reasoning that trial lengths of four or less ensure that participants report the simple span items from the typical maximum capacity of the central component of WM (Cowan, 2000). Furthermore, when Loaiza and McCabe (2012) included supra-span trials of simple span that included eight memoranda, there was no long-term advantage similar to the McCabe effect observed compared to sub-span trials of simple span. Of course, this may be because even the extended number of items still does not provide the time necessary to engage in elaboration during simple span. Thus, further work will be needed to explore this issue.

Another potential factor that has little to do with the processing of items in WM per se is that of the nature of overt retrieval from WM, such that it may be warranted to draw a distinction between elaborative encoding versus elaborative retrieval. For example, even when items are shallowly encoded, effortful retrieval after distraction, such as during complex span, may involve generating retrieval cues to delimit a search set that disambiguates current trial items from irrelevant but recently active items (Rose et al., 2014; Unsworth & Engle, 2007). This effortful retrieval may involve a deeper form of retrieval that benefits EM relative to simply reporting items directly from the focus of

attention that have never been distracted, as in for slow span and simple span. This has two main implications: First, participants may validly report elaborative strategies during slow span and complex span, but implement them differently, such that elaborative encoding and maintenance benefits slow span and effortful, elaborative retrieval benefits complex span. Although these may not be fundamentally different processes, there may still be a distinction in how elaboration is implemented during encoding versus retrieval from WM. The fact that the McCabe effect was sometimes smaller or null in Experiment 1 when using overall recall that was not corrected for successful immediate recall may hint at this. However, a McCabe effect was still observed in Experiment 2 when using reconstruction rather than self-generated recall that mitigated the immediate baseline differences between the tasks. This suggests that effortful retrieval is not necessary to observe a McCabe effect.

In fact, we have recently tested the prediction that effortful retrieval from complex span may explain the McCabe effect in a series of experiments that varied the immediate retrieval demands between simple and complex span (Loaiza et al., 2020). That is, requiring participants to serially recall items may be disproportionately more difficult for complex span compared to simple span, and overcoming these demands (perhaps by using elaborative retrieval strategies) may result in a McCabe effect that is specific to difficult, serial-recall conditions compared to having no retrieval demands at all. However, Loaiza and colleagues (2020) showed that a McCabe effect was observed in both serial-recall and no-recall conditions, thus suggesting that overt retrieval demands do not matter for the effect. Coupled with the results of Experiment 2 that used reconstruction rather than self-generated recall, this indicates that baseline differences in initial recall and low delayed performance overall may reduce the opportunity to observe a McCabe effect, but do not explain the effect. Notwithstanding, this notion would need to be further replicated using slow span, but at least both the McCabe and slow span effects appear to be at least partly attributable to activities such as elaboration taking place while encoding and processing items in WM.

Conclusions

In summary, the current work suggests that spontaneous elaborative strategies contribute to the long-term advantages of slow span and complex span over simple span. This suggests that elaboration is a common underlying factor for the advantages of the time spent processing items in WM for long-term EM, regardless of the nature of the intervening task affording this time (i.e., slow span or complex span). This finding conflicts with the covert retrieval account that covert retrieval drives the McCabe effect, whereas elaboration may drive the slow span effect. However, the fact that both effects were observed even under ineffective strategies suggests that there are further factors beyond elaboration that contribute to the beneficial long-term effects of time in WM.

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.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition*, *36*(8), 1383–1390. https://doi.org/10.3758/MC.36.8.1383

Bailey, H. R., Dunlosky, J., & Hertzog, C. (2014). Does Strategy Training Reduce Age-Related Deficits in Working Memory? *Gerontology*, *60*(4), 346–356. https://doi.org/10.1159/000356699

Bailey, Heather, Dunlosky, J., & Hertzog, C. (2009). Does differential strategy use account for age-related deficits in working-memory performance? *Psychology and Aging*, *24*(1), 82–92. https://doi.org/10.1037/a0014078

Bailey, Heather, Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & Cognition*, *39*(3), 447–461. https://doi.org/10.3758/s13421-010-0034-3

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

Bartsch, L. M., Loaiza, V. M., Jäncke, L., Oberauer, K., & Lewis-Peacock, J. A. (2019). Dissociating refreshing and elaboration and their impacts on memory. *NeuroImage*, *199*, 585–597. https://doi.org/10.1016/j.neuroimage.2019.06.028

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

Blumenfeld, R. S., & Ranganath, C. (2006). Dorsolateral Prefrontal Cortex Promotes Long-Term Memory Formation through Its Role in Working Memory Organization. *Journal of Neuroscience*, *26*(3), 916–925. https://doi.org/10.1523/JNEUROSCI.2353-05.2006

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

Broadbent, D. E. (1958). *Perception and communication*. Oxford University Press.

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

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

Camos, V., Mora, G., & Oberauer, K. (2011). Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. *Memory & Cognition*, *39*(2), 231–244. https://doi.org/10.3758/s13421-010-0011-x

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

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.

Cotton, K., & Ricker, T. J. (2020). Working memory consolidation improves long-term memory recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000954

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185. https://doi.org/10.1017/S0140525x01003922

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

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

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.

Craik, Fergus I. M. (2002). Levels of processing: Past, present. and future? *Memory (Hove, England)*, *10*(5–6), 305–318. https://doi.org/10.1080/09658210244000135

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy production? *Psychology and Aging*, *13*(4), 597–607. https://doi.org/10.1037/0882-7974.13.4.597

Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition*, *29*(2), 247–253.

Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *The Quarterly Journal of Experimental Psychology*, *60*(9), 1227–1245. https://doi.org/10.1080/17470210600926075

Ebbinghaus, H. (1885). *Memory: A contribution to experimental Psychology* (H. A. Ruger & C. E. Bussenius, Trans.). Dover.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*(1), 219–234. https://doi.org/10.3758/s13423-017-1317-5

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. https://doi.org/10.3758/s13421-014-0461-7

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

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

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, *82*(3), 472–481. https://doi.org/10.1037/h0028372

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 471–480. https://doi.org/10.1016/S0022-5371(73)80027-1

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

James, W. (1890). *Principles of psychology*. Holt.

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

Johnson, M. K., Reeder, J. A., Raye, C. L., & Mitchell, K. J. (2002). Second thoughts versus second looks: An age-related deficit in reflectively refreshing just-activated information. *Psychological Science*, *13*(1), 64–67. https://doi.org/10.1111/1467-9280.00411

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

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

Kruschke, J. K., & Meredith, M. (2020). *Bayesian estimation supersedes the t-test* (0.5.2) [Computer software]. https://cran.r-project.org/web/packages/BEST/BEST.pdf

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., & Camos, V. (2016). Does Controlling for Temporal Parameters Change the Levels-of-Processing Effect in Working Memory? *Advances in Cognitive Psychology*, *12*(1), 2–9. https://doi.org/10.5709/acp-0182-3

Loaiza, V. M., & Camos, V. (2018). The role of semantic representations in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(6), 863. https://doi.org/10.1037/xlm0000475

Loaiza, V. M., Doherty, C., & Howlett, P. (2020). The long-term consequences of retrieval demands during working memory. *Memory & Cognition*. https://doi.org/10.3758/s13421-020-01079-5

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. (2013). The influence of aging on attentional refreshing and articulatory rehearsal during working memory on later episodic memory performance. *Aging, Neuropsychology, and Cognition*, *20*(4), 471–493. https://doi.org/10.1080/13825585.2012.738289

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

Loftus, G. R. (1978). On the interpretation of interactions. *Memory & Cognition*, *6*, 312–319. https://doi.org/10.3758/BF03197461

Mazuryk, G. F., & Lockhart, R. S. (1974). Negative recency and levels of processing in free recall. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *28*(1), 114–123.

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

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition*, *29*(1), 10–17. https://doi.org/10.3758/BF03195736

Murdock, B. B. (1967). Recent Developments in Short-Term Memory. *British Journal of Psychology*, *58*(3–4), 421–433. https://doi.org/10.1111/j.2044-8295.1967.tb01099.x

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

Nishiyama, R. (2014). Active maintenance of semantic representations. *Psychonomic Bulletin & Review*, *21*(6), 1583–1589. https://doi.org/10.3758/s13423-014-0618-1

Nishiyama, R. (2018). Separability of active semantic and phonological maintenance in verbal working memory. *PLOS ONE*, *13*(3), e0193808. https://doi.org/10.1371/journal.pone.0193808

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin*, *143*(9), 992–1009. https://doi.org/10.1037/bul0000108

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

Richardson, J. T. E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review*, *5*(4), 597–614. https://doi.org/10.3758/BF03208837

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

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*(9), 677–688. https://doi.org/10.1037/0022-3514.35.9.677

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

Rose, N. S., & Craik, F. I. M. (2012). A processing approach to the working memory/long-term memory distinction: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 1019–1029. https://doi.org/10.1037/a0026976

Rose, N. S., Craik, F. I. M., & Buchsbaum, B. R. (2015). Levels of Processing in Working Memory: Differential Involvement of Frontotemporal Networks. *Journal of Cognitive Neuroscience*, *27*(3), 522–532. https://doi.org/10.1162/jocn_a_00738

Rose, N. S., Myerson, J., Roediger, H. L., & Hale, S. (2010). Similarities and differences between working memory and long-term memory: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 471–483. https://doi.org/10.1037/a0018405

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

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

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Shivde, G., & Anderson, M. C. (2011). On the existence of semantic working memory: Evidence for direct semantic maintenance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(6), 1342–1370. https://doi.org/10.1037/a0024832

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

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

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

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–403). Academic Press.

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Turner, M. L., & Engle, R. W. (1989). Is Working Memory Capacity Task Dependent? *Journal of Memory and Language*, *28*(2), 127–154.

Unsworth, N., & Spillers, G. J. (2010). Variation in working memory capacity and episodic recall: The contributions of strategic encoding and contextual retrieval. *Psychonomic Bulletin & Review*, *17*(2), 200–205. https://doi.org/10.3758/PBR.17.2.200

Unsworth, Nash, Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition*, *41*(2), 242–254. https://doi.org/10.3758/s13421-012-0261-x

Unsworth, Nash, & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition*, *40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*(2), 89–104. https://doi.org/10.1037/h0021797

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00433

Table 1. Means (and standard deviations) of proportion of accurate immediate recall as a function of instruction group and task type, as well as the mean posterior estimates [and 95% credibility intervals] comparing the differences between task types across experiments.

Exp.	Instruction Group	Task type	Serial scoring	Free scoring	Pairwise Comparison	Serial Scoring	Free scoring
1A	None	Simple	0.96 (0.19)	0.98 (0.15)	Simple vs. Complex	**3.78 [2.88, 4.99]**	**3.52 [2.41, 5.29]**
		Complex	0.51 (0.50)	0.73 (0.45)	Simple vs. Slow	**1.24 [0.18, 2.52]**	1.28 [-0.14, 3.19]
		Slow	0.90 (0.31)	0.94 (0.23)	Complex vs. Slow	**-2.54 [-3.54, -1.79]**	**-2.23 [-3.26, -1.47]**
1B	Ineffective	Simple	0.93 (0.25)	0.96 (0.19)	Simple vs. Complex	**2.87 [2.16, 3.69]**	**2.35 [1.77, 2.99]**
		Complex	0.53 (0.50)	0.72 (0.45)	Simple vs. Slow	-0.58 [-1.83, 0.49]	-0.66 [-1.90, 0.37]
		Slow	0.92 (0.27)	0.96 (0.19)	Complex vs. Slow	**-3.45 [-4.70, -2.44]**	**-3.01 [-4.21, -2.08]**
	Elaborative	Simple	0.95 (0.23)	0.96 (0.19)	Simple vs. Complex	**3.85 [3.04, 4.84]**	**2.96 [2.21, 3.87]**
		Complex	0.44 (0.50)	0.69 (0.47)	Simple vs. Slow	0.15 [-1.30, 1.44]	-0.06 [-1.23, 1.03]
		Slow	0.92 (0.28)	0.96 (0.19)	Complex vs. Slow	**-3.70 [-5.00, -2.69]**	**-3.01 [-4.05, -2.16]**
2	None	Simple	0.93 (0.26)	0.98 (0.14)	Simple vs. Complex	**0.96 [0.40, 1.53]**	**0.73 [0.23, 1.25]**
		Complex	0.85 (0.36)	0.96 (0.19)	Simple vs. Slow	-0.23 [-0.83, 0.34]	-0.01 [-0.63, 0.58]
		Slow	0.94 (0.24)	0.98 (0.14)	Complex vs. Slow	**-1.19 [-1.80, -0.61]**	**-0.74 [-1.35, -0.18]**
	Ineffective	Simple	0.87 (0.34)	0.97 (0.19)	Simple vs. Complex	**0.78 [0.16, 1.43]**	**0.83 [0.33, 1.37]**
		Complex	0.82 (0.38)	0.93 (0.25)	Simple vs. Slow	**-0.86 [-1.89, -0.01]**	-0.25 [-1.01, 0.40]
		Slow	0.89 (0.31)	0.97 (0.18)	Complex vs. Slow	**-1.64 [-2.62, -0.86]**	**-1.08 [-1.82, -0.46]**
	Elaborative	Simple	0.91 (0.28)	0.97 (0.17)	Simple vs. Complex	**1.61 [1.07, 2.20]**	**0.78 [0.25, 1.38]**
		Complex	0.80 (0.40)	0.95 (0.22)	Simple vs. Slow	-0.62 [-1.50, 0.14]	-0.73 [-1.85, 0.09]
		Slow	0.92 (0.27)	0.97 (0.17)	Complex vs. Slow	**-2.23 [-3.05, -1.58]**	**-1.51 [-2.56, -0.72]**

Note. Exp. = experiment. Effects in boldface are credible.

Table 2. Means (and standard deviations) of proportion of reported strategies as a function of instruction group and task type.

			Ineffective Strategies			Elaborative Strategies						
Exp.	Instruction Group	Task type	Passive reading	Rote rehearsal	Overall	Sentence generation	Imagery	Grouping	Overall	Refreshing	Other	Don't remember
1A	None	Simple	0.22 (0.28)	0.37 (0.23)	0.59 (0.28)	0.03 (0.08)	0.07 (0.14)	0.10 (0.15)	0.21 (0.22)	0.06 (0.11)	0.01 (0.05)	0.11 (0.15)
		Complex	0.21 (0.23)	0.30 (0.23)	0.51 (0.25)	0.06 (0.09)	0.06 (0.10)	0.13 (0.18)	0.25 (0.21)	0.12 (0.18)	0.01 (0.05)	0.08 (0.16)
		Slow	0.23 (0.26)	0.25 (0.23)	0.48 (0.28)	0.08 (0.11)	0.12 (0.15)	0.13 (0.20)	0.33 (0.28)	0.13 (0.18)	0.01 (0.05)	0.03 (0.06)
1B	Ineffective	Simple	0.26 (0.25)	0.38 (0.33)	0.63 (0.27)	0.01 (0.05)	0.04 (0.09)	0.05 (0.09)	0.10 (0.13)	0.06 (0.13)	0.01 (0.05)	0.15 (0.20)
		Complex	0.20 (0.25)	0.40 (0.31)	0.60 (0.32)	0.02 (0.06)	0.05 (0.09)	0.09 (0.15)	0.16 (0.18)	0.09 (0.15)	0.03 (0.08)	0.08 (0.16)
		Slow	0.23 (0.21)	0.35 (0.28)	0.58 (0.29)	0.05 (0.12)	0.09 (0.16)	0.08 (0.16)	0.22 (0.25)	0.05 (0.08)	0.05 (0.14)	0.06 (0.13)
	Elaborative	Simple	0.21 (0.22)	0.19 (0.20)	0.40 (0.29)	0.06 (0.14)	0.18 (0.22)	0.08 (0.12)	0.32 (0.28)	0.07 (0.18)	0.00 (0.00)	0.18 (0.22)
		Complex	0.17 (0.19)	0.22 (0.25)	0.40 (0.25)	0.10 (0.18)	0.18 (0.24)	0.12 (0.23)	0.41 (0.30)	0.10 (0.16)	0.00 (0.00)	0.06 (0.08)
		Slow	0.08 (0.17)	0.12 (0.14)	0.21 (0.26)	0.22 (0.28)	0.28 (0.29)	0.16 (0.24)	0.66 (0.32)	0.06 (0.14)	0.01 (0.03)	0.02 (0.06)
2	None	Simple	0.69 (0.35)	0.20 (0.31)	0.89 (0.20)	0.01 (0.05)	0.04 (0.09)	0.05 (0.15)	0.11 (0.20)	-	0.00 (0.00)	-
		Complex	0.34 (0.35)	0.50 (0.38)	0.84 (0.27)	0.04 (0.12)	0.06 (0.15)	0.06 (0.14)	0.16 (0.27)	-	0.00 (0.02)	-
		Slow	0.35 (0.34)	0.46 (0.38)	0.82 (0.26)	0.02 (0.06)	0.11 (0.19)	0.05 (0.15)	0.18 (0.26)	-	0.00 (0.00)	-
	Ineffective	Simple	0.45 (0.43)	0.47 (0.44)	0.92 (0.19)	0.01 (0.03)	0.01 (0.03)	0.05 (0.17)	0.06 (0.18)	-	0.01 (0.08)	-
		Complex	0.25 (0.35)	0.63 (0.39)	0.88 (0.22)	0.03 (0.10)	0.03 (0.10)	0.04 (0.14)	0.10 (0.22)	-	0.01 (0.07)	-
		Slow	0.26 (0.36)	0.63 (0.39)	0.89 (0.24)	0.03 (0.08)	0.03 (0.09)	0.04 (0.17)	0.10 (0.23)	-	0.01 (0.08)	-
	Elaborative	Simple	0.44 (0.39)	0.16 (0.29)	0.60 (0.39)	0.08 (0.17)	0.19 (0.28)	0.13 (0.25)	0.40 (0.38)	-	0.01 (0.03)	-
		Complex	0.28 (0.33)	0.18 (0.31)	0.45 (0.40)	0.12 (0.20)	0.29 (0.32)	0.14 (0.21)	0.54 (0.39)	-	0.01 (0.04)	-
		Slow	0.25 (0.36)	0.16 (0.30)	0.40 (0.42)	0.14 (0.25)	0.32 (0.34)	0.14 (0.21)	0.60 (0.42)	-	0.00 (0.01)	-

Note. Exp. = experiment.

Table 3. Mean posterior estimates [and 95% credibility intervals] comparing the delayed performance differences between task types for Models 2 and 3 (see text for details).

				Experiment 1	Experiment 2	
Model	Strategies		Pairwise Comparison	Free Scoring	Free Scoring	Serial Scoring
2	Reported	Ineffective	Complex vs. Simple	**0.69 [0.41, 0.98]**	**0.25 [0.14, 0.37]**	**0.36 [0.20, 0.50]**
			Slow vs. Simple	**0.35 [0.06, 0.64]**	**0.34 [0.21, 0.47]**	**0.43 [0.26, 0.60]**
			Slow vs. Complex	**-0.34 [-0.66, -0.03]**	0.08 [-0.05, 0.22]	0.08 [-0.08, 0.23]
		Elaborative	Complex vs. Simple	**1.01 [0.62, 1.39]**	**0.36 [0.12, 0.59]**	0.22 [-0.05, 0.47]
			Slow vs. Simple	**1.05 [0.68, 1.41]**	**0.63 [0.37, 0.89]**	**0.68 [0.40, 0.94]**
			Slow vs. Complex	0.04 [-0.34, 0.43]	**0.27 [0.003, 0.54]**	**0.47 [0.22, 0.71]**
3	Instructed	None	Complex vs. Simple	**1.16 [0.79, 1.53]**	**0.35 [0.18, 0.52]**	**0.39 [0.15, 0.62]**
			Slow vs. Simple	**0.82 [0.42, 1.22]**	**0.44 [0.25, 0.64]**	**0.56 [0.27, 0.85]**
			Slow vs. Complex	-0.34 [-0.77, 0.09]	0.09 [-0.11, 0.30]	0.17 [-0.07, 0.41]
		Ineffective	Complex vs. Simple	**0.46 [0.02, 0.92]**	**0.24 [0.06, 0.42]**	**0.39 [0.15, 0.63]**
			Slow vs. Simple	0.32 [-0.10, 0.72]	**0.32 [0.12, 0.53]**	**0.43 [0.13, 0.73]**
			Slow vs. Complex	-0.14 [-0.64, 0.38]	0.09 [-0.12, 0.31]	0.04 [-0.23, 0.30]
		Elaborative	Complex vs. Simple	**0.99 [0.57, 1.43]**	**0.37 [0.17, 0.58]**	**0.32 [0.05, 0.58]**
			Slow vs. Simple	**1.35 [0.93, 1.80]**	**0.66 [0.43, 0.91]**	**0.72 [0.44, 1.00]**
			Slow vs. Complex	0.36 [-0.12, 0.82]	**0.29 [0.04, 0.54]**	**0.40 [0.14, 0.66]**

Note. Effects in boldface are credible.

Figure 1. Mean posterior estimates of the likelihood of reporting elaborative (top) and ineffective (bottom) strategies as a function of instruction group and task type in Experiment 1. Error bars reflect 95% credibility intervals and individual points reflect posterior predicted responses based on the models.

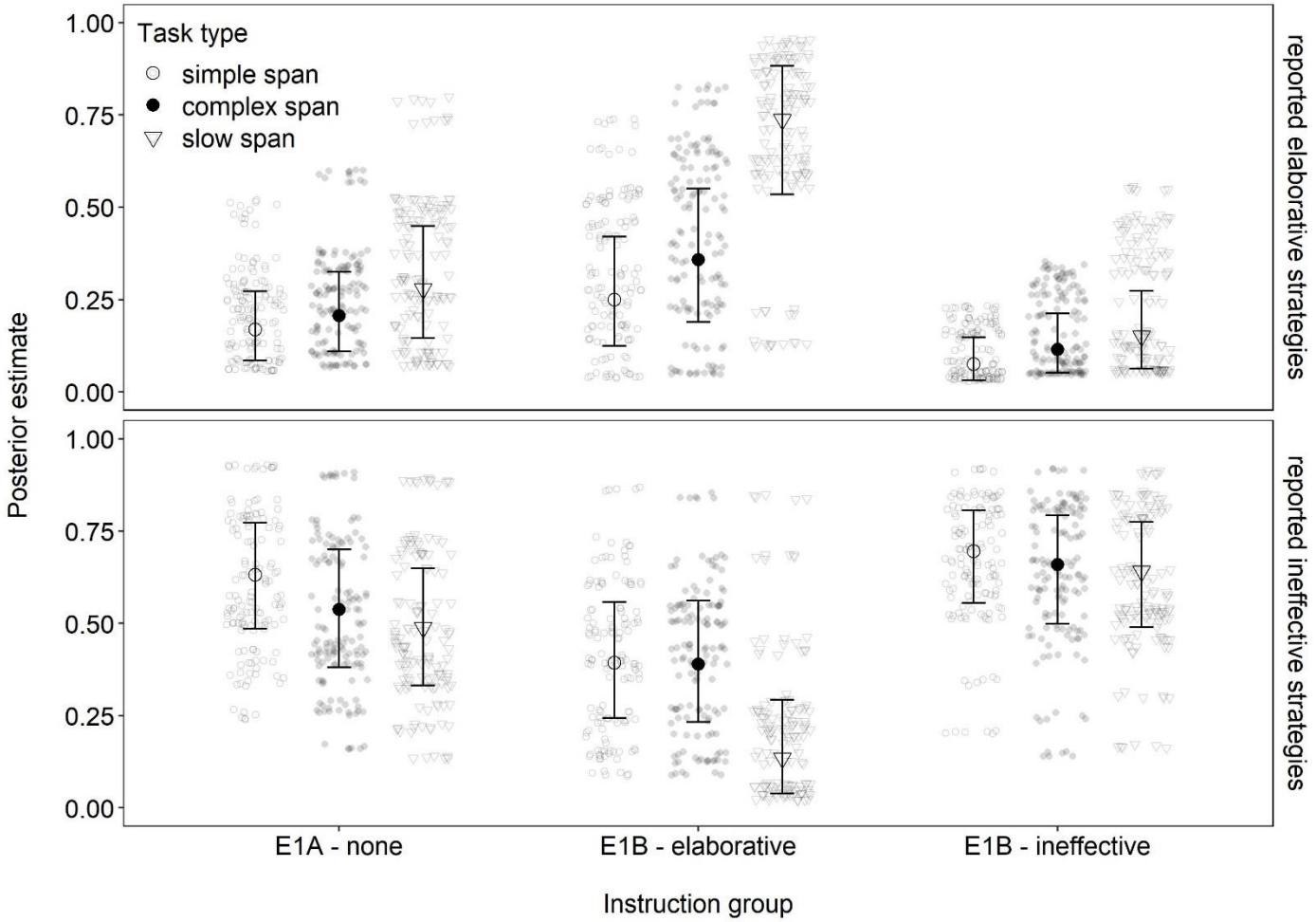

Figure 2. Mean posterior estimates of delayed recall in Experiment 1 as a function of task type, reported strategy, and instruction group, depending on the tested model (see text for details). Error bars reflect 95% credibility intervals and individual points reflect posterior predicted responses based on the models.

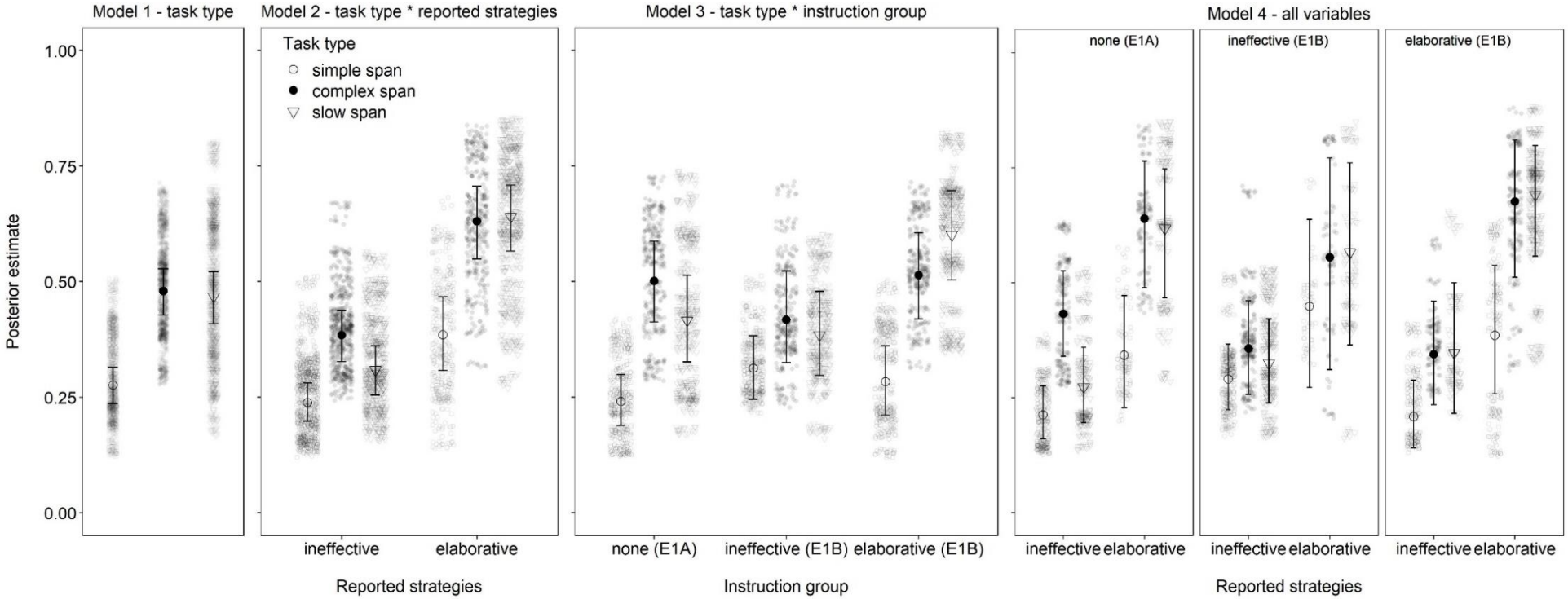

Figure 3. Mean posterior estimates of the likelihood of reporting elaborative (top) and ineffective (bottom) strategies as a function of instruction group and task type in Experiment 2. Error bars reflect 95% credibility intervals and individual points reflect posterior predicted responses based on the models.

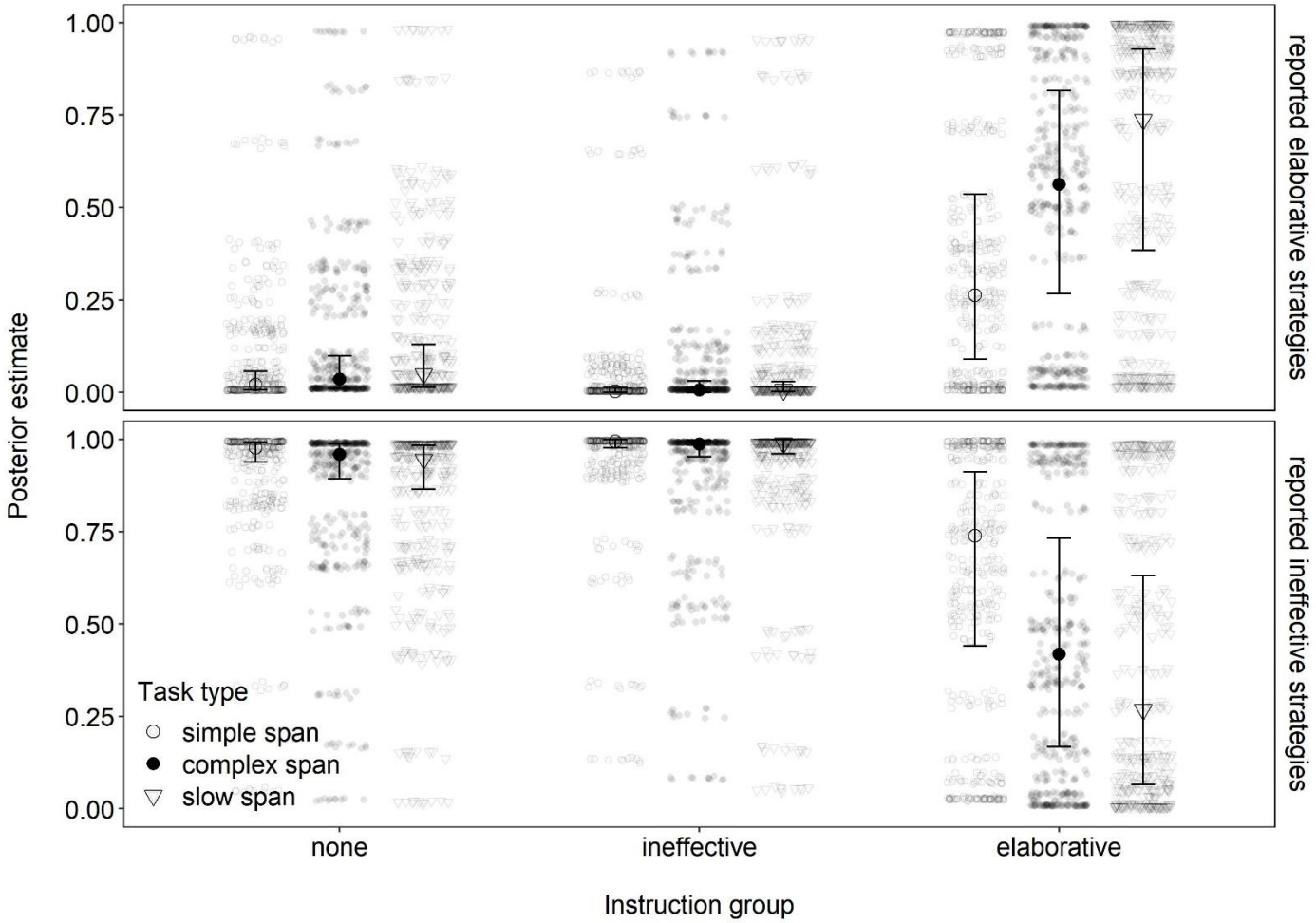

Figure 4. Mean posterior estimates of delayed performance (free scoring) in Experiment 2 as a function of task type, reported strategy, and instruction group, depending on the tested model (see text for details). Error bars reflect 95% credibility intervals and individual points reflect posterior predicted responses based on the models.

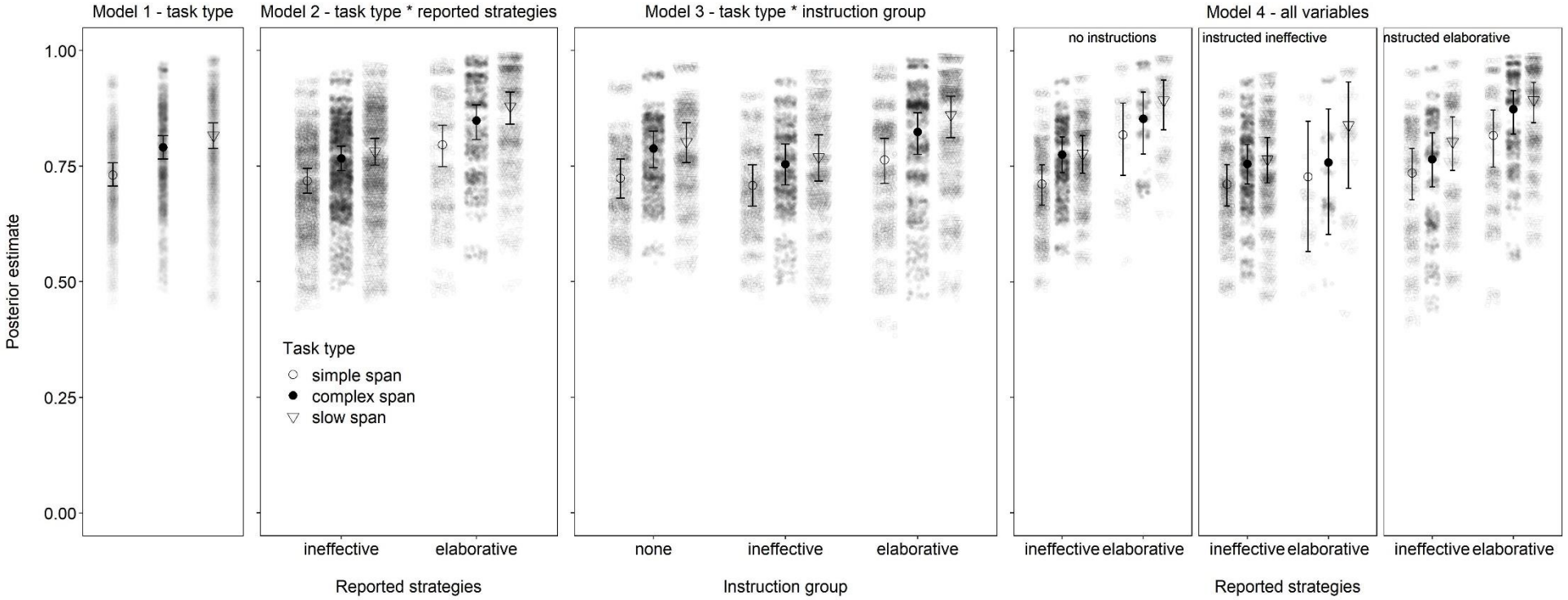

Figure 5. Mean posterior estimates of delayed performance (serial scoring) in Experiment 2 as a function of task type, reported strategy, and instruction group, depending on the tested model (see text for details). Error bars reflect 95% credibility intervals and individual points reflect posterior predicted responses based on the models.