



Published in final edited form as:

Psychon Bull Rev. 2021 April ; 28(2): 537–547. doi:10.3758/s13423-020-01847-z.

Working Memory Limits Severely Constrain Long-term Retention

Alicia Forsberg¹, Dominic Guitard², Nelson Cowan¹

¹University of Missouri

²Université de Moncton

Abstract

There has been considerable controversy in recent years as to whether information held in WM is rapidly forgotten or automatically transferred to LTM. Although visual working memory (WM) capacity is very limited, we appear able to store a virtually infinite amount of information in visual long-term memory (LTM). Still, LTM retrieval often fails. Some view visual WM as a mental sketchpad that is wiped clean when new information enters, but not a consistent precursor of LTM. Others view the WM and LTM systems as inherently linked. Distinguishing between these possibilities has been difficult, as attempts to directly manipulate the active holding of information in visual WM has typically introduced various confounds. Here, we capitalized on the WM system's capacity limitation to control the likelihood that visual information was actively held in WM. Our young-adult participants ($N = 103$) performed a WM task with unique everyday items, presented in groups of either two, four, six, or eight items. Presentation time was adjusted according to the number of items. Subsequently, we tested participants' LTM for items from the WM task. LTM was better for items presented originally within smaller WM set sizes, indicating that WM limitations contribute to subsequent LTM failures, and that holding items in WM enhances LTM encoding. Our results suggest that a limit in WM capacity contributes to an LTM encoding bottleneck for trial-unique familiar objects, with a relatively large effect size.

Working Memory (WM) is a system for holding mental representations temporarily for use in thought and action (Cowan, 2017). A crucial feature of this system is its limit to 3–4 objects concurrently (Adam et al., 2017; Cowan, 2001). In contrast, we seem able to store unlimited information in Long-Term Memory (LTM) (e.g., Brady et al., 2008). Still, a large proportion of the information we encounter is quickly forgotten. For instance, you might forget your colleague's outfit despite recently seeing it in the hallway. This illustrates what we term the LTM encoding bottleneck. Even if all of the information one encounters is somehow encoded into LTM, much of it is at least not encoded in a manner sufficient for later recognition. Here, we asked whether this bottleneck can be explained by WM limitations.

Corresponding author: Alicia Forsberg, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri-Columbia, Columbia, Missouri 65211-2500 USA, 573 397 8113, aliciaforsberg@missouri.edu.

Publisher's Disclaimer: This Author Accepted Manuscript is a PDF file of a an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

How WM and LTM systems interact has implications for theories of memory, learning, and cognition. WM capacity is linked to fluid intelligence (Conway et al., 2005; Daneman & Carpenter, 1980) and educational attainment (Gathercole et al., 2004). High-WM-capacity individuals' advantages may stem partly from their ability to retain comparatively more information in WM at once, which promotes better LTM transfer.

The relationship between WM¹ and LTM remains controversial (Cowan, 2019). Some theorists view the two systems as closely related, regarding WM as a temporarily activated, capacity-limited subset of LTM information, which can include rapidly-learned and still-active, new associations (e.g., Cowan, 1988, 2019; Morey et al., 2013). Others propose that WM and LTM are separate streams, based largely on neurological findings of selective deficits of WM with relatively intact LTM (e.g., Shallice & Warrington, 1970). In the visual modality, on which we focus, some propose that information is held in WM via a 'visual cache', likened to a mental sketchpad from which information is not retained long; it is seen as not a consistent precursor of episodic long-term memory (Logie et al., 2009; Shimi & Logie, 2018). Here, we re-examine whether holding visual information in WM improves LTM encoding.

The idea that WM acts as a bottleneck for LTM storage was already present in Atkinson and Shiffrin's (1968) seminal model of memory. The transfer from verbal WM to LTM has been controversial, and the time available to encode items is often conflated with the time available for maintenance. Hartshorne and Makovski (2019) meta-analytically reviewed and supplemented this literature, which includes only a few studies using visual materials and suggests a small effect size ($d \approx 0.2$) for the benefit of WM encoding on LTM. A WM encoding bottleneck may contribute to discrepant results, and our study pursues the hypothesis that the likelihood of evidence of LTM encoding depends on the likelihood of an item from an array entering WM.

Recent attempts to explore the relationship between visual WM and LTM have produced intriguing results. Some evidence indicates that WM is cleared out on a moment-by-moment basis, as participants failed to notice the repetition of identical arrays of color–shape–location bindings during a WM task, suggesting that arrays were not well-encoded into LTM (Logie et al., 2009; Shimi & Logie, 2018). In contrast, Fukuda and Vogel (2019) found that participants remembered items from a WM task in a subsequent LTM test, and that high-WM-capacity individuals recalled more items from the WM task. Others found that while repetition of WM arrays does not necessarily boost WM performance, repeated arrays were recognized above chance levels in a subsequent LTM test (Olson & Jiang, 2004; Olson et al., 2005; see also Chun & Jiang 1998; Chun & Jiang 2003). Hartshorne and Makovski (2019) compared LTM for single items that were passively viewed, retained in WM for recognition, or attended and used as probes, and found mixed evidence for an LTM benefit for WM items.

¹We use the term visual WM for our procedure but some find the term 'visual short-term memory' more appropriate, and some consider the two terms interchangeable (Cowan, 2017).

Here, we test LTM formation in a simple fashion, presenting arrays of different numbers of known objects for immediate recognition of a probe that could come from the array, and later testing LTM recognition of items not probed before. We allowed an equal processing time per item regardless of the array set size. In previous work, Logie et al. (2009) and Shimi and Logie (2018) instead examined slow effects of array repetitions in situations in which participants may not have managed to encode the complete WM array on a given trial because the number of objects exceeded typical WM capacity. Comparing partial array representations might explain the failure to notice array repetition. Fukuda and Vogel (2019) tested LTM for the WM items, but did not allow an equal amount of processing time per item across all array sizes. Time is important not only to allow initial entry of items into a capacity-limited WM (Woodman & Vogel, 2005), but also to allow further consolidation of WM information (Ricker et al., 2018). By using known objects rather than abstract objects such as colored bars, and presenting each object in only one array, we attempt to maximize the likelihood that pre-existing knowledge can play a role to allow the creation of accessible new LTM representations (Endress & Potter, 2014; Shoval et al., 2020).

We relied on a core distinguishing feature of the WM system: its capacity limit (3 – 4 items, Adam et al., 2017; Cowan, 2001). Thus, by manipulating an item's WM set size, we manipulated the probability that it was held in WM. We restricted LTM testing to items that were not probed in WM or used as WM probes to avoid repeated exposure or, worse, a repeated testing confound (Roediger III & Karpicke, 2006). If items being held in WM matters for their LTM encoding, then items from lower array sizes in the WM procedure should be better remembered in a later LTM test, because an item in a smaller array is more likely to be held in WM.

We examined our basic hypothesis by assessing the mean LTM contents as a function of the mean WM contents and by examining the correlation between an individual's success at the WM task and the LTM task.

Method

The methods, and all analyses (except those labeled 'as exploratory') were pre-registered on the Open Science Framework at [https://osf.io/jrzwg?view_only=fca6a8ce3aa54c05b18675f73291d9be].

Participants

Pilot data and sample size rationale—We tested 30 pilot participants to detect any issues in our online data collection procedure and to ensure that participants could complete the study in a reasonable timeframe (see the Supplement, Section 1, for mean performance values and sample size determination procedures). Based on simulations, 100 participants appeared necessary and sufficient for results to be of interest to others in the field, regardless of the outcome.

Final Sample—For online data collection, the software PsyToolKit was used (Stoet, 2010, 2017). We recruited participants via prolific.co, for safe and convenient remote data collection, producing results seemingly comparable to those obtained in laboratory studies

(Germine et al., 2012; Peer et al., 2017). Participants were paid £5 for completing the study. See the Supplement, Section 2, for detailed inclusion criteria. Our pre-registered target sample size was 100, but 104 participants were accidentally recruited online. One participant was excluded per our set exclusion criteria (see Online Supplement, Section 3), so the final N was 103 participants. The age of the final sample was 24.2 years ($SD = 3.50$, range 18–30 years), with the gender distribution female (62.1%), male (37.9%). Detailed demographic information is reported in the Supplement, Section 4. On average, participants completed the experiment in 32.2 minutes ($SD = 8.4$; range = 21–78 minutes). This study was approved by the local IRB committee.

General Study Design

Participants completed three different tasks: First, a WM probe-recognition task, followed by a brief mathematical distraction task and, finally, a second probe-recognition memory test, assessing LTM for items from the WM task. Participants read written instructions prior to each task. Figure 1 shows a trial in the WM phase and two trials within the LTM phase. The crucial manipulation was the WM set size (i.e., the number of items presented simultaneously; 2, 4, 6, or 8 items). The WM array presentation time was adjusted according to the number of items (250 ms per item). Participants were not informed that they would be tested on the WM items again but, at the very end of the session, we asked them whether they had expected a retest on these items.

Working Memory (WM) Task

Memory items were selected from the Microsoft Office ‘Icons’, consisting of various easily recognizable images, including animals, symbols, furniture, food items, etc. All items were presented in black on a light grey background. Participants studied a total of 384 unique memory items in the WM task, at varying set sizes (2, 4, 6, or 8 items). Items in an array were presented in an imaginary circle around a central fixation cross (+), such that one item could be placed at every 45-degree increment (Figure 1). For set size 8, each space was occupied, and for lower set sizes, locations were selected at random.

Each trial started with a 250 ms central fixation cross, followed by the memory array, and a 2000 ms delay, before the probe item and response options were presented. On each trial, an array was presented for 250 ms times the number of items in the array, and was followed by a 2000-ms delay and then a probe item, which was drawn randomly from the array on half of the trials and was a new item not seen in any array on the other half. They responded by clicking on one of the following options presented on the screen along with the probe: ‘*sure same*’, ‘*believe same*’, ‘*guess same*’, ‘*guess different*’, ‘*believe different*’, or ‘*sure different*’. As a way to ensure that participants could not let the experiment finish without responding, after a 10-minute period of no response the program timed out and the participant would not be counted in the analysis, though no one did this. The order of trials and selection of items for each trial were randomized for each participant. The number of trials at the four array sizes was 48, 24, 16, and 12, respectively, resulting in 96 unique memory items at each set size.

Distraction Task

Next, participants completed a distraction task lasting 60 s. During this task, participants verified mathematical equations of the form $a \times b + c = d$, where a , b , and c were integers from 1 to 9, and d was equal to $a \times b + c$ or differed from that expression by ± 1 . The integers for a , b , and c were drawn randomly with replacement from the integers 1 to 9, inclusive. The number d that was shown was correct on half the trials and incorrect on the other half. Participants responded by clicking ‘Correct’ or ‘Incorrect’ on the screen. Participants completed as many trials as possible during the 60 s interval.

Long-Term Memory (LTM) Task

On each trial, participants saw a probe item and had to say whether that item was studied in the WM task, or was a new item (see Figure 1). Participants responded to each probe item by clicking on one of the following options on the screen: ‘*sure studied*’, ‘*believe studied*’, ‘*guess studied*’, ‘*guess new*’, ‘*believe new*’, or ‘*sure new*’. Items which were probed in the WM task or served as lures in that task were not probed in the LTM task, to avoid repeated exposure. Each participant responded to a total of 213 items in the LTM task (46 new items, 36 items from set size 2, 42 items from set size 4, 44 items from set size 6, and 45 items from set size 8). As in the WM task, a 10-minute pause was grounds for exclusion of the participant.

Results

Memory Tasks

Accuracy—In statistical analyses, we use a nomenclature in which BF_{10} refers to the Bayes Factor for the presence of an effect and BF_{01} refers to absence of an effect, where $BF_{01} = 1 / BF_{10}$. Figure 2 shows the use of the response scale for the WM task (left panel) and the LTM task (right panel). Lower ratings indicate more confident same/studied responses, and higher ratings indicate more confident ‘different/new’ responses. The general, important pattern can be observed from that figure. Clearly, in WM, participants distinguished fairly well between trials with probes that were old (present in the array, or same) versus new (not in any array, or different). As set size increased, however, ratings began to come closer together, indicating poorer average performance. Similarly, the average LTM rating for items originally presented in set size 2 ($M=2.80$, $SD=1.70$) was lower than for those presented at set size 8 ($M=3.63$, $SD=1.63$; $d=0.50$), reflecting a stronger tendency to correctly – and confidently – identify items from lower WM set sizes as studied.

We used two types accuracy scoring. First, ‘strict’ scoring, in which responses were only considered correct when participants reported some confidence in their response (using ‘Sure’ or ‘Believe’) but guessing (e.g., guess new for a new item) was considered as incorrect in this scoring. Next, ‘lenient’ scoring in which all correct responses (including guessing in the correct direction) were scored as correct. WM accuracy across set sizes is presented in Figure 3. Using Bayesian Logistic Regression (*brms*; R, Bürkner; 2017, 2018), and considering responses marked as guesses as incorrect (i.e., ‘strict’ scoring), we found credible evidence that memory performance decreased as set size increased ($\eta = -0.47$; $SE = 0.01$, 95% CI $[-0.50, -0.45]$). This trial-level analytical approach was appropriate to

account for the increased uncertainty at set sizes with lower trial numbers. For details see the supplement Section 5. The BF in favor of the model including set size was 3.38×10^{306} over a model not including this factor, indicating that the set size manipulation influenced WM performance as expected.

Two separate pre-registered analyses addressed the key question of whether successful encoding of items in WM influences subsequent LTM representations. First, we tested whether performance in the LTM task varied as a function of WM set size (coded as a continuous numeric variable) using ‘*generalTestBF*’ (*R* package BayesFactor, see the Supplement, Section 6 for details). We found ‘decisive’ evidence that LTM memory performance was better for items presented for lower-set size items, both when coding ‘guesses’ as incorrect ($BF_{10}=7.06 \times 10^{97}$) and correct ($BF_{10}=1.99 \times 10^{72}$). See Figure 3 for accuracy rates across set sizes. The LTM accuracy for items originally presented in set size 2 ($M=0.65$, $SD=0.48$) was higher than that for items presented at set size 8 ($M=0.45$, $SD=0.50$; $d=0.41$). The ratio of novel to old items was uneven (46 new items vs. 167 old). Although the prevalence of old items may alter the bias so as to affect the levels of both hits and false alarms, any such effect would presumably be across all set sizes and therefore could not in itself produce the set size effect.

The 29.1% of participants who reported expecting a LTM test of the WM items did not remember items better than naïve participants (exploratory Bayesian ANOVA; $BF_{01}=51.68$ and 7.88 for strict and lenient accuracy scoring).

Analysis of the Number of Items in WM and LTM for each WM Set Size—A

second kind of measure was used to address the question of how many of the items encoded into WM in fact made it into LTM. For each individual, the proportion of items from a given size of WM set was in memory at the time of (1) WM testing; $p(\text{WM})$, and (2) LTM testing; $p(\text{LTM})$. The goal was to form a ratio of LTM to WM item presence. For this measure, we ignored whether a response was *sure*, *believe*, or *guess* and just considered whether the correct half of the response scale was used. The $p(\text{WM})$ estimates were obtained using correct detection of an old item (i.e., a hit, h) and calling a new item old (i.e., a false alarm, f). The model to estimate $p(\text{WM})$ is derived from work by Pashler (1988) as applied to the present test situation by Cowan et al. (2013, “reverse-Pashler” formula).² Participants should respond correctly when the probed item is in WM and otherwise guess that the item is new with a certain rate (g). The rate of correct detection of old items, h , equals the probability that the probe item is in WM plus the probability that it is not in WM but that a correct “old” guess g is given:

$$h = p(\text{WM}) + [1 - p(\text{WM})](g)$$

When the item is new there is no match, so performance depends on the guessing rate and an incorrect response (f) is made at the rate, $f=g$.

²Note that we have found it more natural to redefine hits as correct detection of an ‘old’ or studied item, and false alarms as incorrect indications that a novel item was ‘old’ or studied, differing from Pashler and Cowan et al.

Combining these formulas, it can be shown that

$$p(\text{WM}) = \left(\frac{h-f}{1-f} \right)$$

A comparable formula can be used with LTM data (hits, hl for correctly detected old items and false alarms, fl for incorrect responses that a novel item was old) to yield the proportion of items in LTM:

$$p(\text{LTM}) = \left(\frac{hl-fl}{1-fl} \right)$$

Multiplying $p(\text{WM})$ and $p(\text{LTM})$ by the set size yields an estimate of the number of items from an array available at those test stages. In accordance with preregistered exclusion criteria, 45 observations of negative $p(\text{WM})$ values (6) or $p(\text{LTM})$ values (39) resulted in the exclusion of 34 participants, some of whom had more than one negative value. Figure 4 shows the number of items in WM (left-hand panel) and LTM (right-hand panel) for each WM array size.

As Table 1 shows, the likelihood that items that were encoded in WM were subsequently remembered in the LTM task seems higher for Set Size 2 (54%) than for larger set sizes, which seem similar to one another (37–39%), and the LTM/WM. The ratio differed across set sizes ($\text{BF}_{10}=40.93$). An exploratory analysis excluding Set Size 2 showed evidence against a difference between the remaining set sizes ($\text{BF}_{01}=6.59$). These findings suggest that when WM reached capacity, items encoded into WM had a certain likelihood of being recalled in LTM. When the stimuli were presumably below capacity at Set Size 2, it appears that more intensive transfer into LTM could take place. Thirty-four participants with floor-level scores in at least one set size had been excluded but, to include them, in an exploratory analysis we adjusted $k < 1$ to 1, and $p(\text{LTM}) = 0$ to 0 and again obtained the difference in LTM/WM ratio between set sizes, $\text{BF}_{10}=9.85$, and evidence for the null when Set Size 2 was excluded, $\text{BF}_{01}=7.70$ (see Figure 5).

The Effect of WM Trial Inaccuracy on LTM—We conducted an exploratory analysis to test the effect of WM trial-accuracy on LTM retention. We used only data from *same* trials, since errors guessing ‘different’ when the item was the same is the clearest indication that at least one item (i.e., the probed item) in that array was not held in WM. The beneficial effect of WM trial accuracy on LTM appeared greater at lower set sizes ($\text{BF}=29.36$; see Table S3, and for detailed parameter estimates, see Supplement, Section 10, in which we also report a similar analysis for Different trial data and a more holistic registered analysis). The effect was by far the clearest for 2-item arrays with a “same” probe, which resulted in higher LTM performance on array items from trials in which the WM probe was recognized ($M=.53$, $SD=.50$) compared to when it was not recognized ($M=.37$, $SD=.49$).

Correlations between Items in WM versus LTM—We carried out an exploratory analysis in which we averaged the number of items across supra-capacity set sizes (4, 6, and 8), at which most participants were below ceiling, for WM (k) and for LTM (multiplying

$p(LTM)$ by the relevant set size). To allow inclusion of all participants at each set size, we replaced instances in which $WM < 1$ with 1 and values of $p(LTM) = 0$ with zero. The result (Figure 6) indicated a positive correlation, $r = .24$, $BF_{10} = 6.01$. There was imperfect transfer from WM to LTM in every participant (i.e., all points below the diagonal line). See online supplement (Section 11) for an alternative, pre-registered analysis.

Mathematical Distraction Task

On average, participants attempted 14.8 ($SD = 5.0$, range 4 – 32) problems during this one-minute distraction task, and the average accuracy rate was 86.6% ($SD = 12.3$, range 40 – 100% accurate), indicating that participants were generally engaged with this task.

Discussion

Although the visual WM and LTM systems are often considered to be linked, specifics of their relationship are contentious. Our results provide strong evidence that WM encoding enhances subsequent LTM representations, contradicting suggestions that items held in visual WM are quickly erased (e.g., Logie et al., 2009; Shimi & Logie, 2018). To summarize our key findings: (1) both WM and LTM performance levels were higher for items presented as part of a smaller set in WM (Figures 3 & 4. For LTM, this difference appeared especially large between items sub-capacity (set size 2), and supra-capacity set sizes (4, 6, and 8); (2) the ratio of items in LTM to WM was constant across arrays of four, six, and eight items (Figure 5); and (3) performance levels on WM and LTM were correlated on an individual-participant level. When one attempts to hold an overwhelming amount of information in mind, one is likely to forget some of it, in both immediate and delayed testing. Our results suggest that WM encoding acts as a bottleneck for visual LTM retention (Atkinson & Shiffrin, 1968; Fukuda & Vogel, 2019), and verify that WM and LTM encoding both are constrained by a WM capacity limit (Brady et al., 2008; Endress & Potter, 2014; Shoval et al., 2020). Our effect size for unique, familiar objects, comparing the smallest and largest WM array set sizes ($d = 0.41$), is about double what Hartshorne and Makovski (2019) obtained in the research literature using less-familiar objects, reinforcing the importance of linking into distinct information already in LTM (Brady et al., 2008; Endress & Potter, 2014; Shoval et al., 2020).

The pattern of ratios between items in LTM and WM indicated stronger LTM encoding at a sub-capacity set size of 2 items (Figure 5). That pattern warrants follow-up research as it would be consistent with studies indicating that precision declines when set size increases from one to two to three items, though not much beyond three. Embedded processes models of WM suggest that some items may be held in a limited focus of attention (Cowan, 1995), which may hold 1–2 items (Öztekin et al., 2010; Sutterer et al., 2019) or 3–5 items (Cowan, 2001). Being in the focus of attention might enhance LTM encoding (see Cowan, 1988; 2019). Perhaps the greater difference between LTM retention at set size two, as compared to all larger set sizes, reflects that items from set size 2 are more likely to enter the focus of attention, compared to items presented as part of larger set sizes.

At lower WM set sizes, and especially in trials with a 2-item array and a probe drawn from the array, WM trial-failure resulted in poor LTM memory for items in arrays drawn from

those trials. This result for 2-item arrays cannot be attributed to a capacity limit, but it corresponds to the expected effect of trial inattention (looking away or mind-wandering), as in the model of Rouder et al. (2008). Indeed, ongoing fluctuations of attention, WM maintenance, and LTM performance are known to be linked (e.g., Adam et al., 2015; Aly & Turk-Browne, 2016; deBettencourt et al., 2019; Murray et al., 2011; Unsworth & Robison, 2016).

We observed a correlation among individuals for information in WM and LTM, in an exploratory analysis of the number of items held in WM and LTM averaged across Set Sizes 4, 6, and 8. This result provides further support for the notion that holding information in WM is beneficial for subsequent LTM retrieval, and is aligned with the results of Fukuda and Vogel (2019). Given that the association was not very strong, future work could investigate individual differences in encoding style that might distinguish between those with good versus poor LTM (e.g., along the lines of Craik & Lockhart, 1972), or on individual differences in the functioning of brain areas highly relevant to LTM, including the hippocampus and nearby areas (e.g., Wixted et al., 2018).

Various WM processes may produce the LTM bottleneck effect we have observed. It may be driven by a limit in the number of items that can be actively represented in the brain simultaneously, if such active representations underlie LTM encoding (Cowan, 2019). Interestingly, our results differed from those of Bartsch et al. (2019), who did not find evidence for a set size effect on LTM, using sequential presentation of word pairs. Perhaps sequential presentation resulted in equal entry to focus of attention for all items, regardless of set size. In contrast, in our procedure, perhaps only certain items were held in the focus of attention during the 2000 ms retention interval. This discrepancy may indicate that an item's presence in the focus of attention during WM maintenance may be crucial for subsequent LTM recognition (Cowan, 1998; 2019).

Maintenance of items in WM may rely on attentional refreshing (Barrouillet et al., 2004; Camos et al., 2018) or verbal rehearsal (Baddeley et al., 1975; Forsberg et al., 2019) but their effects on LTM have been disputed (Bartsch et al., 2018; Hartshorne & Makovski, 2019). Elaboration strategies, such as mental imagery or chunk formation, do boost LTM (Dunlosky & Hertzog, 2001) and should only be possible for items concurrently in WM. Undoubtedly, participants approach the task with various strategies (Logie, 2018). Inasmuch as the strategic approaches efficient for WM differ from those efficient for LTM (Bartsch et al., 2018), voluntary elaborative strategies seem unlikely to drive our finding that WM array size affected LTM retention, since expecting the LTM test did not seem to improve performance (cf. Fukuda & Vogel, 2019). However, it is still possible that participants opted for different strategies for different set sizes. Future research, perhaps with sequential presentation, should explore which specific WM processes contribute to the effect.

To conclude, items successfully held in WM were more likely retained in LTM. The ratio of the number of items held in LTM to items held WM from the array during which the item was first presented appeared fairly constant for arrays above capacity. When the WM array did not fill up WM capacity, additional resources seemed to boost LTM encoding further.

Overall, our results suggest that WM processes are indeed part of the LTM encoding bottleneck.

Open Practices Statement

The methods, and all analyses (except those labeled ‘as exploratory’) were pre-registered on the Open Science Framework at [https://osf.io/jrzwg?view_only=fca6a8ce3aa54c05b18675f73291d9be]. Due to space limitations, some preregistered analyses are reported in the online Supplement. Data, analysis code, and study materials are available at [https://osf.io/qfzn3/?view_only=490083f12cb1461694f8afc6b1109b70].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Bret Glass for assisting in data collection and acknowledge NIH Grant R01-HD021338 to Cowan.

References

- Adam KC, Mance I, Fukuda K, & Vogel EK (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience*, 27(8), 1601–1616. [PubMed: 25811710]
- Adam KC, Vogel EK, & Awh E (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, 97, 79–97. [PubMed: 28734172]
- Aly M, & Turk-Browne NB (2016). Attention promotes episodic encoding by stabilizing hippocampal representations. *Proceedings of the National Academy of Sciences*, 113(4), E420–E429.
- Atkinson RC, & Shiffrin RM (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation*, 2(4), 89–195.
- Baddeley AD, Thomson N, & Buchanan M (1975). Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior*, 14, 575–589.
- Barr DJ, Levy R, Scheepers C, & Tily HJ (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- 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. [PubMed: 14979753]
- Bartsch LM, 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. [PubMed: 29557069]
- Brady TF, Konkle T, Alvarez GA, & Oliva A (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38), 14325–14329.
- Bürkner PC (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1–28.
- Bürkner PC (2018). Advanced Bayesian Multilevel Modeling with the R Package brms.” *The R Journal*, 10(1), 395–411.
- Camos V, Johnson MR, Loaiza VM, Portrat S, Souza AS, & Vergauwe E (2018). What is attentional refreshing in working memory? *Annals of the New York Academy of Sciences*, 1424(1), 19–32. [PubMed: 29542133]

- Chun MM, & Jiang Y (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36(1), 28–71. [PubMed: 9679076]
- Chun MM, & Jiang Y (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(2), 224.
- Conway AR, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, & Engle RW (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12(5), 769–786. [PubMed: 16523997]
- Cowan N (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, 104(2), 163. [PubMed: 3054993]
- Cowan N (1992). Verbal memory span and the timing of spoken recall. *Journal of Memory and Language*, 31, 668–684.
- Cowan N (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1), 87–114.
- Cowan N (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, 24, 1158–1170. [PubMed: 27896630]
- Cowan N (2019). Short-term memory based on activated long-term memory: A review in response to Norris (2017). *Psychological Bulletin*, 145, 822–847. [PubMed: 31328941]
- Cowan N, Blume CL, & Saults JS (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 731.
- Craik FI, & Lockhart RS (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 671–684.
- Daneman M, & Carpenter PA (1980). Individual differences in working memory and reading. *Journal of Memory and Language*, 19(4), 450.
- DeBettencourt MT, Keene PA, Awh E, & Vogel EK (2019). Real-time triggering reveals concurrent lapses of attention and working memory. *Nature Human Behaviour*, 3(8), 808–816.
- 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. [PubMed: 11352207]
- Endress AD, & Potter MC (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General*, 143(2), 548. [PubMed: 23937181]
- Fukuda K, & Vogel EK (2019). Visual short-term memory capacity predicts the “bandwidth” of visual long-term memory encoding. *Memory & Cognition*, 47(8), 1481–1497. [PubMed: 31236821]
- Forsberg A, Johnson W, & Logie RH (2019). Aging and feature-binding in visual working memory: The role of verbal rehearsal. *Psychology and Aging*, 34(7), 933. [PubMed: 31535870]
- Gathercole SE, Pickering SJ, Knight C, & Stegmann Z (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 18(1), 1–16.
- Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, & Wilmer JB (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5), 847–857. [PubMed: 22829343]
- Hartshorne JK, & Makovski T (2019). The effect of working memory maintenance on long-term memory. *Memory & Cognition*, 47(4), 749–763. [PubMed: 31073790]
- Logie R (2018). Human cognition: Common principles and individual variation. *Journal of Applied Research in Memory and Cognition*, 7(4), 471–486.
- Logie RH, Brockmole JR, & Vandembroucke AR (2009). Bound feature combinations in visual short-term memory are fragile but influence long-term learning. *Visual Cognition*, 17(1–2), 160–179.
- Ma WJ, Husain M, & Bays PM (2014). Changing concepts of working memory. *Nature Neuroscience*, 17, 347–56. [PubMed: 24569831]
- Morey CC, Morey RD, van der Reijden M, & Holweg M (2013). Asymmetric cross-domain interference between two working memory tasks: Implications for models of working memory. *Journal of Memory and Language*, 69(3), 324–348.

- Morey RD, Rouder JN, Jamil T, & Morey MRD (2015). Package 'bayesfactor'. URLh <http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf> (accessed 1006 15).
- Murray AM, Nobre AC, & Stokes MG (2011). Markers of preparatory attention predict visual short-term memory performance. *Neuropsychologia*, 49(6), 1458–1465. [PubMed: 21335015]
- Olson IR, & Jiang Y (2004). Visual short-term memory is not improved by training. *Memory & Cognition*, 32(8), 1326–1332. [PubMed: 15900926]
- Olson IR, Jiang Y, & Moore KS (2005). Associative learning improves visual working memory performance. *Journal of Experimental Psychology: Human Perception and Performance*, 31(5), 889. [PubMed: 16262486]
- Öztekın I, Davachi L, & McElree B (2010). Are representations in working memory distinct from representations in long-term memory? Neural evidence in support of a single store. *Psychological Science*, 21(8), 1123–1133. [PubMed: 20624934]
- Pashler H (1988). Familiarity and visual change detection. *Perception & Psychophysics*, 44, 369–378. [PubMed: 3226885]
- Peer E, Brandimarte L, Samat S, & Acquisti A (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163.
- Ricker TJ, Nieuwenstein MR, Bayliss DM, & Barrouillet P (2018). Working memory consolidation: insights from studies on attention and working memory. *Annals of the New York Academy of Sciences*, 1424, 8–18. [PubMed: 29635694]
- Roediger III HL, & Karpicke JD (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210. [PubMed: 26151629]
- Rouder JN, Morey RD, Cowan N, Zwilling CE, Morey CC, & Pratte MS (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences USA (PNAS)*, 105, 5975–5979.
- Schönbrodt FD & Stefan AM (2018). BFDA: An R package for Bayes factor design analysis (version 0.3). Retrieved from <https://github.com/nicebread/BFDA>
- Shallice T, & Warrington EK (1970). Independent functioning of verbal memory stores: A neuropsychological study. *Quarterly Journal of Experimental Psychology*, 22, 261–273.
- Shimi A, & Logie RH (2019). Feature binding in short-term memory and long-term learning. *Quarterly Journal of Experimental Psychology*, 72(6), 1387–1400.
- Shoval R, Luria R, & Makovski T (2020). Bridging the gap between visual temporary memory and working memory: The role of stimuli distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46, 1258–1269.
- Stoet G (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096–1104. [PubMed: 21139177]
- Stoet G (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24–31.
- Sutterer DW, Foster JJ, Adam KC, Vogel EK, & Awh E (2019). Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. *PLoS biology*, 17(4), e3000239. [PubMed: 31026274]
- Unsworth N, & Robison MK (2016). The influence of lapses of attention on working memory capacity. *Memory & Cognition*, 44(2), 188–196. [PubMed: 26450588]
- Wixted JT, Goldinger SD, Squire LR, Kuhn JR, Papesh MH, Smith KA, Treiman DM, & Steinmetz PN (2018). Coding of episodic memory in the human hippocampus. *PNAS*, 115, 1093–1098. [PubMed: 29339476]
- Woodman GF, & Vogel EK (2005). Fractionating working memory: Consolidation and maintenance are independent processes. *Psychological Science*, 16, 106–113. [PubMed: 15686576]
- Xu Z, Adam KCS, Fang X, & Vogel EK (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods*, 50(2), 576–588. [PubMed: 28389852]
- Zhang W, & Luck SJ (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, 453, 23–35.

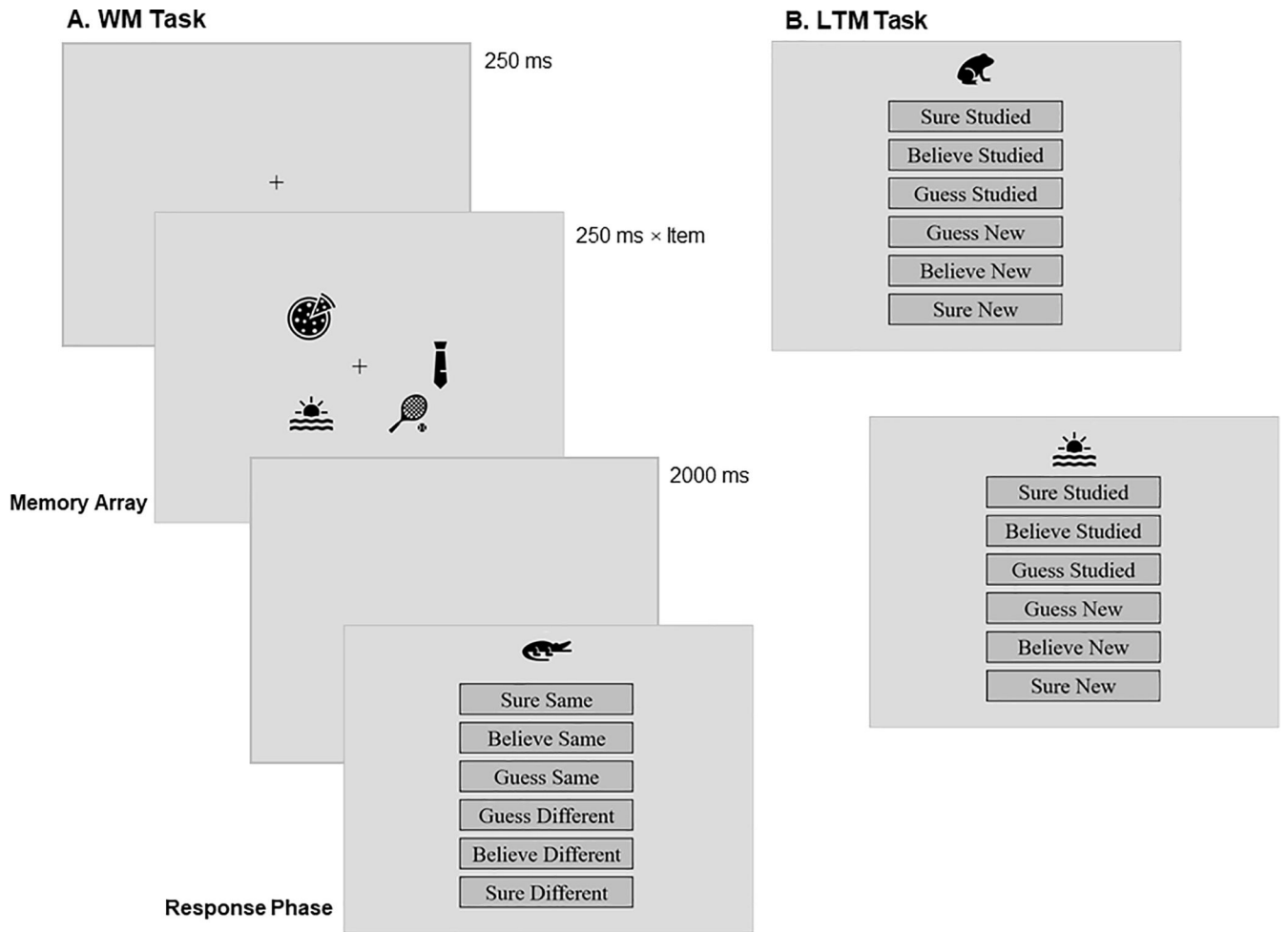


Figure 1.

Outline of some typical trials. **Panel A**, Working Memory (WM) Task trial. **Panel B**, two trials in the Long-Term Memory (LTM) Task. The memory array set size in the WM task varied between 2, 4, 6, or 8 items, and the presentation time was adjusted to be 250 ms per item. During the WM response phase, participants indicated whether the probe item was the same as an item in the array, or different, and also selected their level of confidence, by a mouse-click on the relevant option. In the LTM task, participants indicated whether items had been studied in the WM task.

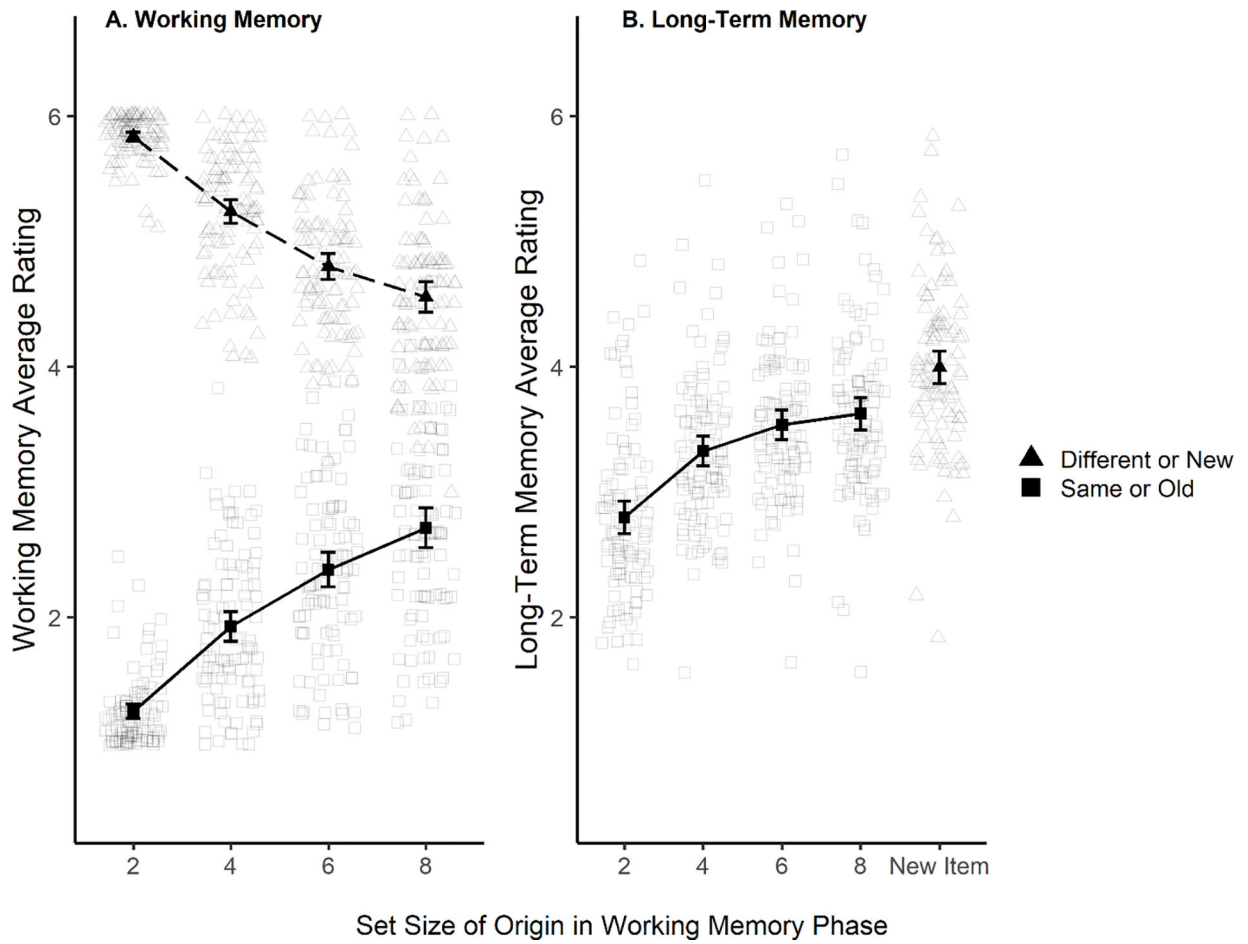


Figure 2.

Average ratings (1 – 6; 1 = *sure same/studied*, 2 = *believe same/studied*, 3 = *guess same/studied*, 4 = *guess different/new*, 5 = *believe different/new*, 6 = *sure different/new*) by Set Size in the WM task. **Panel A**, WM task ratings; **Panel B**, LTM task ratings. Circles show ratings on trials when the probe item was different or new, and diamonds show performance when the probe item was the same as one in the studied set, or old. The black circles and diamonds represent the overall mean ratings. The transparent, smaller circles and diamonds represent individual participants' mean ratings. These are jittered to avoid overlap. Higher ratings indicate higher confidence that the item was different, and lower ratings indicate higher confidence that the item was the same. New items in the LTM phase were not studied within any set size in the WM phase. Error bars on the group means represent 95% confidence intervals.

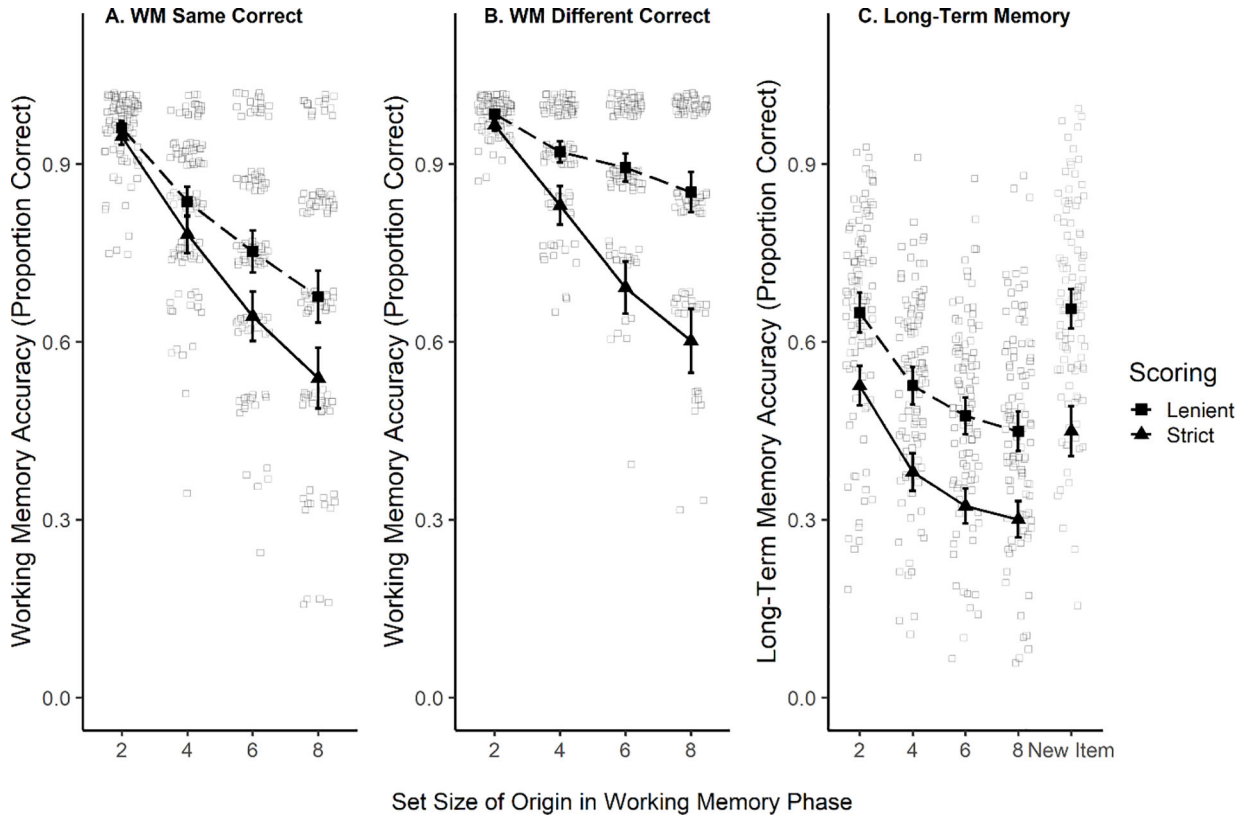


Figure 3. Memory accuracy by WM set size. **Panel A**, WM hits (i.e., correctly identified ‘Same’ trials); **Panel B**, WM correct rejections (i.e., correctly identified ‘Different’ trials); **Panel C**, LTM accuracy. Black triangles and the solid line show the average WM accuracy across trials in which responses in the guessing range were always scored as incorrect (strict scoring). The dashed line and squares show accuracy in selecting same or different regardless of participants’ exact confidence rating (lenient scoring). Light squares show individual subjects’ accuracy by set size for the lenient scoring (the points are jittered slightly in the figure to avoid overlap). Error bars represent 95% confidence intervals.

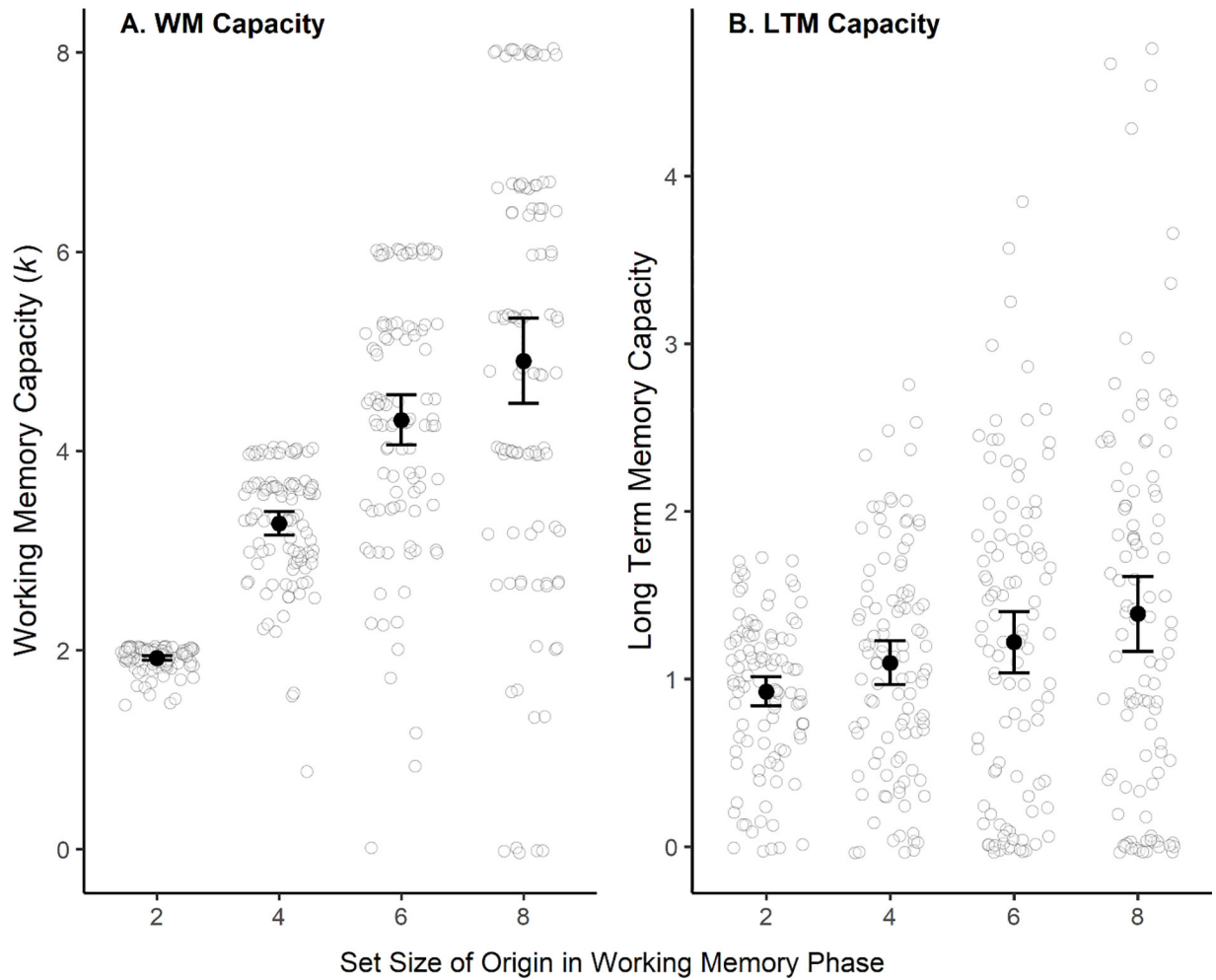


Figure 4.

Estimated items in WM and LTM. **Panel A**, items in WM (k). **Panel B**, items in LTM. Items in LTM were calculated for each participant by multiplying $\rho(\text{LTM})$ by the number of items in the relevant WM set size. $\rho(\text{LTM}) < 0$ values were recoded as 0. Black circles represent the mean number of estimated items. Grey circle outlines represent individual subject estimates and are jittered slightly to avoid overlap. Error bars on the mean are 95% confidence intervals.

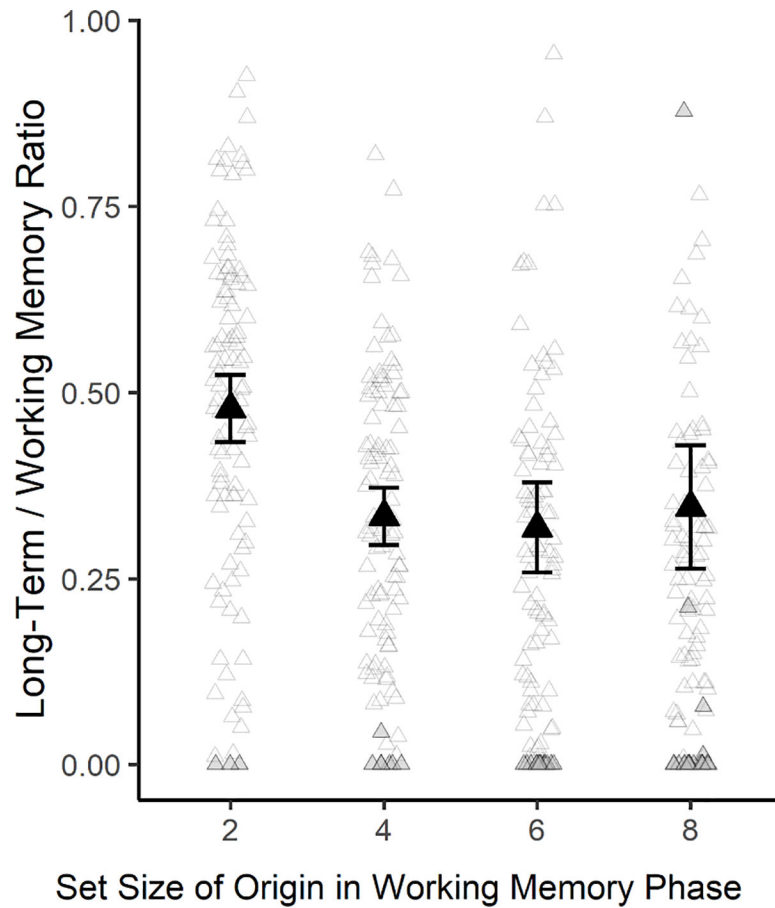


Figure 5.

LTM / WM Ratios, by WM set size. The large black triangles show the average for each set size (and error bars are 95% confidence intervals). Smaller triangle outlines show individual subject points. Grey triangles show adjusted values (either WM $k < 1$, and was adjusted to 1, or $p(\text{LTM})$ was negative and was adjusted to 0). To avoid excessive whitespace, 11 LTM / WM Ratio values > 1 were removed from this figure (4 from Set Size 6, 7 from Set Size 8); see Supplement, Figure S2 for the complete figure including these values.

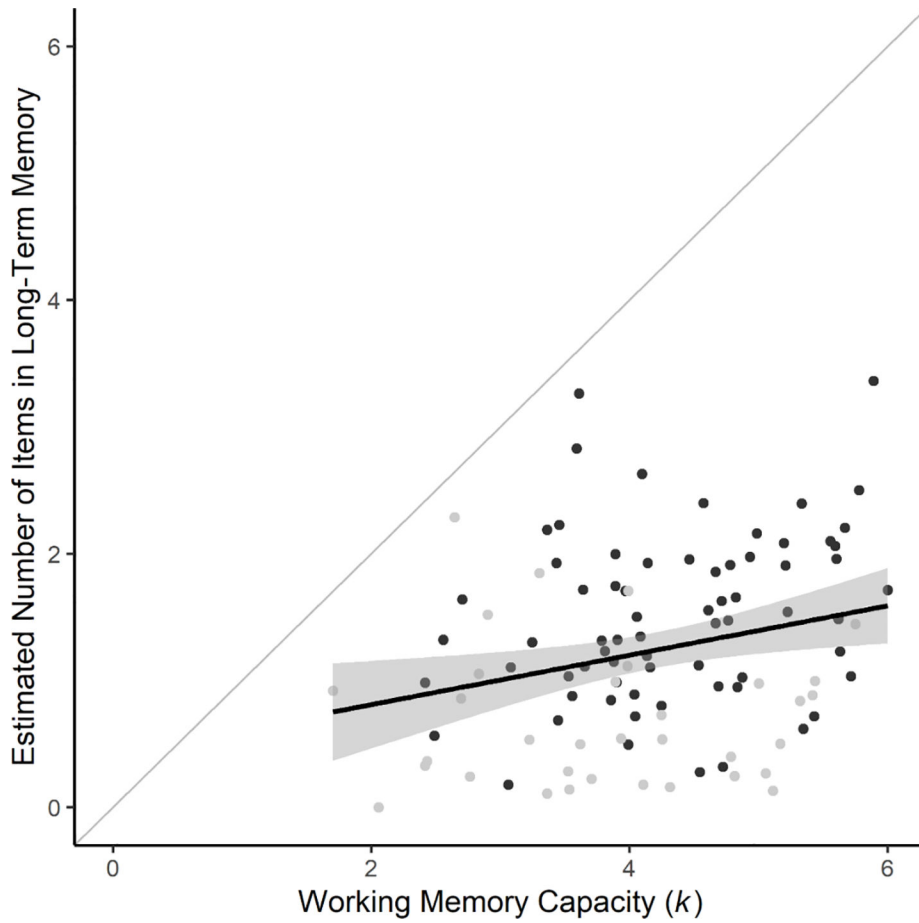


Figure 6. Estimated Number of Items in Long-term Memory for each subject as a function of their Working Memory Capacity (k). The Long-Term Memory estimate was calculated for each participant by multiplying $\rho(LTM)$ by the number of items in the WM set size, and averaging this across all set sizes, not including values from set size 2. The Working Memory capacity (k) value was obtained by averaging k from all set sizes except set size 2. Black points represent individual participants, grey points participants for which at least one $\rho(WM)$ or $\rho(LTM)$ value was adjusted. The black line represents a frequentist linear regression line, and the shaded area includes the 95% confidence region. The grey diagonal line represents hypothetical perfect transfer in which the number of items in LTM=WM k .

Table 1.

Averages for transformed data by Set Size.

Set Size	$p(\text{WM})$ ($N=69$)	$p(\text{LTM})$ ($N=69$)	LTM/WM ratio ($N=69$)	LTM/WM ratio (adjusted) ($N=103$)	Items in WM (k) ($N=103$)
2 items	.96 (.06)	.52 (.20)	0.54 (0.20)	0.48 (0.23)	1.92 (0.12)
4 items	.83 (.14)	.32 (.16)	0.38 (0.19)	0.33 (0.20)	3.27 (0.61)
6 items	.74 (.20)	.25 (.14)	0.37 (0.26)	0.32 (0.31)	4.31 (1.28)
8 items	.66 (.24)	.21 (.13)	0.39 (0.33)	0.35 (0.42)	4.90 (2.18)

Note. Values in Parenthesis represent Standard Deviations. The first LTM/WM ratio includes data only from participants who had no negative values of either $p(\text{WM})$ or $p(\text{LTM})$ at any set size. For the second LTM/WM ratio, we multiplied $p(\text{WM})$ and $p(\text{LTM})$ with the relevant set size. Values of WM $k < 1$ were replaced with 1, and values of $p(\text{LTM}) \times \text{Set Size} = 0$ were replaced with zero.